Liber Amicorum Pascal Engel

Liber Amicorum Pascal Engel

Edited by

Julien Dutant, Davide Fassio and Anne Meylan

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES LETTRES
Département de philosophie

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES LETTRES
Département de philosophie

rue de Candolle 2, 1205 Genève

ISBN 978-2-8399-1562-5
http://www.unige.ch/lettres/philo/publications/engel/liberamicorum/

Contents

Part Two: Knowledge and Belief

Part Three: Mind and Language

Part Four: Norms and Values

Part Five: The Dispute

Part Six: Further papers

Dedication

Dear Pascal,

This volume is presented to you by your friends, colleagues past and present, and former students, on the occasion of your sixtieth birthday. It is the expression of our admiration for your curiosity, your knowledge and your talent, but also our appreciation for your humour and our gratitude for your intellectual generosity. *Tu es une galette qu'aucun renard ne croquera.*[1]

Anne Meylan, Davide Fassio and Julien Dutant.

[1] Cf. Pierre Belvès & Natha Caputo, *Les albums du Père Castor: roule galette.*

Contributors

CAROLA BARBERO University of Turin, Italy.

CLAUDIA BIANCHI University Vita-Salute San Raffaele, Italy.

YVES BOUCHARD Université de Sherbrooke, Canada.

INGAR BRINK Lund University, Sweden.

J. ADAM CARTER University of Edinburgh, United Kingdom.

STÉPHANE CHAUVIER University of Paris-IV Sorbonne, France.

JEAN-MARIE CHEVALIER Collège de France, France.

FRANÇOIS CLEMENTZ Aix-Marseille University, France.

ANNALISA COLIVA University of Modena, Italy.

VASCO CORREIA New University of Lisbon, Portugal.

PAOLO CRIVELLI University of Geneva, Switzerland.

ALAIN DE LIBERA Collège de France, France.

JÉRÔME DOKIC EHESS, Institut Jean Nicod, France.

IGOR DOUVEN University of Groningen, Netherlands.

JULIEN DUTANT University of Geneva, Switzerland.

SANTIAGO ECHEVERRI University of Geneva, Switzerland.

PAUL EGRE CNRS, Institut Jean Nicod, France.

JON ELSTER Columbia University, United States.

MITSUHIRO OKADA Keio University, Japan.

ERIK J. OLSSON Lund University, Sweden.

CLAUDE PANACCIO UQAM, Canada.

CARLO PENCO University of Genoa, Italy.

JOHANNES PERSSON Lund University, Sweden.

ROGER POUIVET LHSP-Archives Poincaré, Université de Lorraine, France.

DUNCAN PRITCHARD University of Edinburgh, United Kingdom.

FRANÇOIS RÉCANATI Institut Jean Nicod, France.

NILS-ERIC SAHLIN Lund University, Sweden.

DANIEL SCHULTHESS University of Neuchâtel, Switzerland.

KATUSYA TAKAHASHI Saitama University, Japan.

FABRICE TERONI University of Bern, Switzerland.

ÉMILE THALABARD University of Paris-IV Sorbonne, France.

GIOVANNI TUZET Università Bocconi, Italy.

STELIOS VIRVIDAKIS University of Athens, Greece.

JACQUES-HENRI VOLLET University of Geneva, Switzerland.

MARCEL WEBER University of Geneva, Switzerland.

Tabula Gratulatoria

The following were unable to present a paper but would like to associate themselves with the present *hommage*:

THOMAS BALDWIN
RENÉE BILODEAU
SIMON BLACKBURN
ALBAN BOUVIER
ARIEL CECCHI
ZOÉ CHRISTOFF
J. MATHIAS FLEURY
RICHARD GLAUSER
GHISLAIN GUIGON
LEILA HAAPARANTA
GERHARD HEINZMANN
MAX KISTLER
SIMO KNUUTTILA
CHARLES LARMORE
DIEGO MARCONI
OLIVIER MASSIN
JACQUES MORIZOT
ROBERT NADEAU
MARTINE NIDA-RÜMELIN
ELISABETH PACHERIE
CHRISTOPHER PEACOCKE
PHILIP PETTIT
JOËLLE PROUST
WLODEK RABINOWICZ
ERNEST SOSA
OLIVIER SOUAN
BENJAMIN SYLVAND
JESÚS VEGA ENCABO

PART ONE

Truth and Fiction

1

What ever happened to the correspondence theory of truth?

MICHAEL P. LYNCH

Introductory remark Pascal Engel is an inspiration. For those of us who have toiled in the rough fields of the truth literature, his work is celebrated for, among other things, having established the question of the value of truth as central to the question of the nature of truth. Engel's work here as elsewhere combines thoroughness and technical sophistication with an extraordinary grasp of how and why the details matter for not only philosophy in general but for our intellectual lives. As such, he remains the very model of what a good philosopher should be. I am honored to contribute to this symposium.

*

1. Overview

Once upon a time, the story goes, the correspondence theory was everyone's theory of truth. Everyone believed it, or at least everyone believed that everyone else believed it. It was – so were' told— the "everyman" of truth theories.

This is no longer the case – if it ever was. While many philosophers who work on truth for a living continue to pay some lip service to the theory, sustained defenses are increasingly rare. If you were to judge just by the airtime it gets among the truthies, so to speak, one might well wonder: What ever happened to the correspondence theory? And where is it now?

Here, in brief, are my answers to these questions. The correspondence theory of truth is missing in action because, at least as it is traditionally conceived, it is a bad theory of truth. But that doesn't mean it should be rejected. For it is a good theory of something else.

2. What counts as a theory of truth?

I want to explain why the correspondence theory, as traditionally conceived is implausible. So I first have to explain how the correspondence theory is traditionally conceived.

Well, as it has been traditionally conceived, the correspondence theory of truth is... a theory of truth. Now for the less obvious: when would a theory count as a theory of truth? I suggest that a theory counts a theory of truth when it incorporates many of the key truisms about truth – when it addresses truth's nominal essence, as Locke might have put it. The nominal essence of F, in the sense I intend, is our folk concept of F. It embodies our preconceptions, the way we tacitly think about it in ordinary life – even if, normally, we don't even recognize ourselves as doing so. A natural way of identifying something's nominal essence, therefore, is to appeal to the set of largely implicit beliefs we folk have about it. By appealing to those folk beliefs, or truisms, we won't typically learn *everything* about the object or property we are interested in. And our later discoveries may force us to revise our preconceptions of it. But however these questions play out, keeping one eye on our folk beliefs about the thing about which we are curious will hopefully tell us whether our subsequent theories of its nature address the topic we were concerned with when our theorizing began.

So I suggest that a theory is a theory about truth as opposed to something else if it incorporates most of core truisms about truth– the nominal essence of truth. So what are these? Here I am interested in just one: the idea that truth is objective. To speak truly is to "say of what is, that it is", as Aristotle said[1] And since what we say, at least when we are sincere, is an expression of what we believe or judge, a parallel truism holds about belief. That is,

> *Objectivity*: The belief that p is true if, and only if, with respect to the belief that p, things are as they are believed to be.

Together with some further and reasonably obvious assumptions, Objectivity underwrites further derivative principles which are typically highlighted by philosophers. One related principle is that when, for example, I believe that roses are red, things are as I believe them to be just when roses are red. That is,

> With respect to the belief that p, things are as they are believed to be if, and only if, p.

With this point in hand, we can derive, together with Objectivity, instances of:

> BS: The belief that p is true if and only if p.

So I we'll count a theory as a theory of *truth* (as opposed to something else) just when, arguably, it incorporates truisms like the above. But we'll count it as a THEORY of truth (as opposed to just a chat about it say) just when it *explains* those truisms.

To explain the truisms, in the sense of "explain" relevant here, is to show why they are true by pointing to some property or properties that all true propositions have that results in those propositions satisfying the truisms.

The correspondence theory, as traditionally conceived, attempts to give an explanation of truth in this sense, and moreover to give a reductive explanation. That is, it attempts to explain the nominal essence of truth in terms of an underlying real essence. *Correspondence* is the name for that underlying real essence—the property that all true beliefs have in common to which, we might say the property truth is "reduced". And it is the having of this property that – according to the theory – explains why true beliefs satisfy the central truisms I just discussed.

[1] Aristotle, *Metaphysics: Books Gamma, Delta, and Epsilo*n (Clarendon Press, 1993)..

3. Vacuity

The version of the correspondence theory that is most well known – what we might call, probably misleadingly, the "original theory" – is also the one which is easiest to be skeptical about. Since its problems are generally well-known, I won't spend much time discussing it. But a few words are in order. The starting point for the original correspondence theory is the idea that

C: A belief is true just when it corresponds to reality.

Taken by itself, however, C is not much of a "theory". For unless we are told what "correspondence" and "reality" mean, this seems to amount to little more than a restatement of the Objectivity truism, or at the most, another simple truism any theory could accept. Indeed, most advocates of the coherence and pragmatist accounts for example, did accept C.

As a result, advocates of the correspondence theory often suggested that C really comes to

CT The belief B is true =def B corresponds to some fact.

And a belief is false, the typical thought runs, just when it fails to correspond to any fact[2] (. This is, perhaps, no longer *completely* vacuous. But it isn't especially promising either, and for by now well-known reasons. Since I have other fish to fry, I'll just briefly mention two.

First, CT is committed to a ontology of facts. And one might wonder what they are. If, on the one hand, we have a rather thin account of "facts" – that is, by "fact " we mean something like "true proposition" then CT might amount to nothing more than the thought that beliefs are true when beliefs correspond to truths. If on the other hand, we say that facts are distinct entities – so distinct that they are over and above the objects and properties that populate the world, we must say something about what they consist in. And that can prove difficult

However, the big problem with CT is not typically thought to be with "facts". It is with "corresponds". Here again the most obvious answers threaten to be vacuous. Thus Millikan writes:

[2] For discussion of recent formulations, see: Pascal Engel, 'Truth', (McGill Queens Univ, 2002), 177, Marian Alexander David, 'Correspondence and Disquotation', *An Essay on the Nature of Truth* (Oxford University Press, 1994), 206, Richard A Fumerton, 'Realism and the Correspondence Theory of Truth', (Rowman & Littlefield, 2002), 149.

> If any certainty has emerged from the last thirty years of philosophy it is that a pure correspondence theory is vacuous. By a *pure* correspondence theory I mean a theory that signs or representations when true or correct, are true or correct merely by virtue of their being a, some, mapping function that maps these representations onto part of the world or reality. [Such a theory will not work] because mathematical mapping relations are infinitely numerous and ubiquitous ... If any correspondence theory of truth is to avoid vacuousness, it must be a theory that tells what is *different* or *special* about the mapping relations that map representations onto representeds...[3]

In order words, what the original theory needs – to really count as a theory – is an understanding of correspondence that allows there to be a real, substantive relationship between beliefs – items in the head—and fact-sized bits of reality. Moreover, this relation must be such as to pick out some *particular* fact-sized bit of reality. And it is difficult to see what that relation could be.

4. Truthmaking to the rescue?

So far, one might think that all I've really said is that the original correspondence theory is not much of a theory. But surely there are more plausible theories that could count as successors of the original correspondence theory?

For example, Far from thinking the correspondence theory is dead, some think the correspondence theory is alive and well but living under an assumed name in Australia. This is the idea that what is right about the correspondence theory is captured by

> Truthmaker: For every truth, there is something that exists which makes it true.

The thought is that Truthmaker captures two core thoughts behind the traditional correspondence theory. First, truth "supervenes on being". Second, truths are made true by something. Correspondence is truth-making[4]

[3] Ruth Garrett Millikan, 'Language, Thought, and Other Biological Categories', *New Foundations for Realis*m (MIT Press, 1984), 355 at 86-88.

[4] David Lewis, 'Forget About the 'Correspondence Theory of Truth'', *Analysis*, 61/272 (2001), 275–80, Marian David, 'Don't Forget About the Correspondence Theory of Truth', *Australasian Journal of Philosophy*, 82/1 (2004), 42 – 47.

Moreover, Truthmaker, as it is typically understood, is sometimes thought to answer the question that the original theory had difficulty with: what is the special relation that exists between a truth and some particular bit of reality. This is because Truthmaker theorists typically say that truth-making is plain old necessitation. A proposition is true when made true. It is made true when necessitated by some existing entity. An existing entity necessitates a proposition just when its existing is metaphysically sufficient for that proposition's being true. And so what was earlier a vice is now a virtue: one entitiy can necessitate more than one truth. There are, so to speak, "many correspondences."[5]

This is all well and good. But it shows that whatever else we might say about Truthmaker – whether it is true for example—it is not a theory of truth, or at least not a correspondence theory traditionally conceived. Here's why. If Truth-maker were a theory of truth, it would be a theory that according to which being true is being made true. But as we just saw, truth making is defined in terms of necessitation, and necessitation in terms of truth. That's a small circle. Thus truthmaker theory may be many things, but it is not an account of truth – or at the very least, it is not a traditional, reductive account.

5. Heir to the throne? Representational theories

A more promising successor to the original correspondence theory is what I call the representationalist theory of truth.

Many of the core elements of the representational theory of truth were initially developed to understand how *sentences* and their component words represent, or *refer* to the world. But the basic elements can, and have been adapted to mental representations, to beliefs and their component concepts. And whether it is applied to sentences or beliefs, contemporary naturalistic representationalism can be understood as offering a two-part theory of truth.[6] First, the truth of a belief, say, is defined in terms the representational features of its component concepts (what I will here call "denotation"). Thus in the

[5] Trenton Merricks, *Truth and Ontology* (Oxford University Press, 2007).)

[6] For early statements of the view, see Hartry H Field, 'Tarski's Theory of Truth', *The Journal of philosophy* (69: Journal of Philosophy, Inc., 1972), 347-75, Michael Devitt, 'Realism and Truth', (Princeton University Press, 1997), 371. An important recent formulation of this sort of approach can be found in Terence Horgan, 'Contextual Semantics and Metaphysical Realism: Truth as Indirect Correspondence', *The Nature of Truth: Classic and Contemporary Perspectives*, 2001.See also R Barnard and T Horgan, 'Truth as Mediated Correspondence', *The Monist* (2006).

case of a belief whose content has the simple predicational structure *a is F*, we get:

> REPRESENT: The belief that a is F is true if and only if the object denoted by <a> has the property denoted by <F>. [7]

The basic thought is that beliefs are true because their components stand in certain representational relations to reality and that reality is a certain way. Adopting machinery made familiar with Tarski, the representationalist then applies this insight to beliefs with more complicated structures.[8] The result is a view according to which the truth of complex beliefs is recursively defined in terms of the truth of simpler beliefs and the rules for logical connectives, while less complex beliefs "correspond to reality" in the sense that their component parts – concepts – themselves represent objects and properties.

The second part of any representational view of truth is a theory of how concepts denote objects and properties. Toy versions of two familiar views are these.

> *CAUSAL:* <cat> denotes cats = cats, cause, under appropriate conditions, mental tokenings of <cat>.[9]
>
> *TELEOLOGICAL: <cat>* denotes cats = the function of <cat> is to be mentally tokened in the presence of cats.

I am not interested in defending either of these familiar proposals here. Rather, I want to stress simply that both are best thought of as a *framing hypothesis* for *naturalistically* investigating mental representation. And for our purposes, the real promise of a naturalistic theory of representation is that theories like CAUSAL and TELEOLOGICAL can be combined with REPRESENT to give a representational theory of truth. According to this theory, truth is defined in terms of representation, representation is defined in terms of denotation, and denotation is defined as a property that either is, or supervenes on natural relations like those specified in CAUSAL or TELEOLOGICAL. Thus,

---

[7] Throughout, I use brackets in the usual way: <dog> means the concept of a dog; <snow is white> means the proposition that snow is white.

[8] Alfred Tarski, 'The Concept of Truth in Formalized Languages', in A. Tarski (ed.), *Logic, Semantics, Metamathematic*s (Oxford University Press, 1936), 152–278.

[9] Obviously I am simplifying in the text the complexities of these theories, and passing over numerous differences in formulation.

to give a toy example, a representational view might be constructed as follows. Let's say that an object or property, which, under appropriate conditions, causes (or its instances cause) mental tokenings of some concept to be "causally mapped" by that concept. If so, we can construct:

> CC (Causal-correspondence): The belief that a is F is true if and only if the object causally mapped by <*a*> has the property causally mapped by <*F*>.

Likewise with a teleological theory of representation: Let us say that a concept that has as its biological function to be mentally tokened in the presence of a particular object or property *functionally maps* that object or property. If so, then one might construct:

> TC (Teleological correspondence): The belief that a is F is true if and only if the object functionally mapped by <a> has the property functionally mapped by <*F*>.

6. Two problems

There are various objections and challenges one might raise against any particular representational theory of truth, including of course, against the views just presented.[10] But there are two more general points to make. As I see it, representational theories – even when more sophisticated than the toy versions just presented – are implausible theories of *truth.* But I think that very are plausible theories of something else.

Why do I think representational theories are implausible when taken as theories of truth? I'll give two reasons. The first is that it is open to doubt whether they really count as theories of *truth* in the first place, plausible or implausible. According to the standard introduced above, a theory counts as a theory of truth just when it not only incorporates the truisms as part of the theory, and offers an explanation of at least most of those truisms. Consider, for example:

[10] For objections and discussion, see for example, Jerry A Fodor, 'Psychosemantics: The Problem of Meaning in the Philosophy of Mind.', (MIT Press, 1987). Peacocke *A Study of Concepts*; (Cambridge: MIT Press 1992); K. Neander "Malfunctioning and Misrepresenting" in *Philosophical Studies*, 79 (1995): 109-141; Millikan, 1996, "On Swampkinds" *Mind and Language*, 11 (1996): 70-130.

> *Objectivity*: My belief that p is true if only if, with respect to the belief that p, things are as I believe them to be.

How does (CC) or (TC) explain Objectivity? Presumably their advocates will say they explain Objectivity by giving an account of what it is for things to be *as* I believe them to be. Suppose, for example, that I believe that Oliver is a cat. According to (CC), things are as I believe them to be if and only if the object causally mapped by <Oliver> has the property causally mapped by <cat>. A similar explanation is available for advocates of (TC).

But is this really an explanation in terms of truth? What makes it different from the following conjunctive explanation, roughly:

> Where I believe that Oliver is a cat, things are as I believe them to be if and only <Oliver> maps Oliver; <cat> maps cats and Oliver is a cat.

More broadly, one can complain that the representationalist theory is something of this form:

> B is true if and only if B = the belief that p; and p.

The second half of this biconditional is a conjunction. And representationalism, one might argue, really only enters into the account of the first conjunct. If we stipulate that the belief has content, the first conjunct falls out, and we are left with

> The belief that p is true if and only if p.

In short, the complaint is that representationalism is really a theory of content married to a simpler theory of truth.

In and of itself, this objection is not, I think, all that devastating. For the representationalist can say that their view just is, in effect, a theory of both content and truth at once. So there.

Yet I think a worry behind this first objection –that representationalism isn't really about truth – remains. That worry takes on more force when combined with a second problem. Representational theories face what I've elsewhere called a problem of scope. This is because such theories require that the objects and properties mapped by our beliefs be capable of entering into at least indirect causal interaction with our minds. This is plausible when we are concentrating on beliefs about cats and cars. But it implausible when we are talking about beliefs like *two and two are four;* or *torture is wrong*. Whatever

these beliefs are about, they aren't plausibly about objects and properties that are in causal contact with our thoughts. Yet both beliefs seem true.

Of course, representationalists are well aware of these examples. Indeed, the history of twentieth century philosophy is replete with isms that have been posed to deal with them – from expressivism to fictionalism. Such theories try to explain away the appearance of truth for the troublesome sorts of beliefs. And maybe these strategies – tired as they are—will work out. I have a different suggestion. It seems to me that *the very fact that we feel the need to construct such theories – to explain away the appearance of truth—points to our having a grip on truth independently of having a grip on representation.* This was, in effect, the point of the first objection as well, but we now have another reason to accept it. It doesn't seem that in order to get a grip on what truth is in general, we need to get a grip on the nature of representation, if only because beliefs that don't represent can be true.

Representationalists favor examples involving cats on mats and the like. There is a reason for this. Theories like (CC) or (TC) are plausible wherever we can make the case that our thoughts about G's are responsive to the antics of the G's themselves. And when it comes to cats on mats, this case seems easy to make. If my belief that there is a cat on the mat is true, it is a *response* to – what else? – there being a cat on the mat. When things are working as they should, when our cognitive machinery is firing on all cylinders so to speak, human beings are good detectors of cats on mats. This suggests a constraint on representationalist theories. A correspondence theory like (CC) will seem likely as theory of what makes mental states with X-ish content true only when we can establish that mental states with X-ish content are causally responsive to an external environment that contains X's. In a bumper sticker, *if we are to correspond, we must respond*.

Moreover, where responsiveness *does* seem plausible, and we have independent reasons for thinking that the content in question is assessable for truth or falsity, it becomes more *likely* that our mental states with x-ish content have that content in virtue of *representing* X's. Accordingly, it will seem more likely that when I believe such content *correctly*– when cognitively speaking, success has been achieved –what *makes* my belief that, e.g. the ubiquitous cat is on some mat correct is that it accurately represents an actual cat on an actual mat.

But where responsiveness is not plausible – either because the states in question aren't appropriately causally responsive or because the external environment contains no x's that can be so causally responsive, then it is less likely that mental-states with x-ish content have that content because they

represent x's. Some other explanation of their content becomes more likely. And thus if we nonetheless wish to maintain that the relevant mental states are *true*, then – to anticipate my closing point – some other account of what makes them true must be pushed onto the field.

Wrapping up, the scope problem suggests that representationalist theories of truth are implausible because they face counterexamples. Certain beliefs are true which couldn't be so if the representationalists theories are correct. We might well conclude then, that as a theory of truth, representationalism fails. But this would be to throw the baby out with the bathwater. For what I think we should conclude is not that representationalism is a false theory of truth, but that it is a true theory of a property the possession of which *makes* some beliefs true. Representational theories are about what makes some kinds of belief true, not truth itself.

The possibility I am suggesting might be put this way: when we talk about the nature of truth, we are speaking ambiguously. First, we might be interested in giving an account of what is in common between all and only those beliefs that are true – be they about mathematics or macramé. That is, we might be talking about truth itself. Second, we might be talking about what makes some particular kind of belief true. That is, we might be talking about a property which, when had by a particular kind of belief, entails that it is true.

And notice: Once we distinguish the projects in this way, the following possibility arises. There may be only one property *being true*, but there may be more than property the possession of which makes or entails that a belief is true. Tempting analogy: there may be only one property *being in pain*, but there may well be more than one neural property the possession of which makes an organism possess that property.

In conclusion: The correspondence theory of truth, taken as a theory of the property of truth, is implausible. But taken as a theory of what makes some kinds of belief have that property, it is very plausible. The correspondence theory of truth is dead. Long live the correspondence theory!

7. References

Aristotle (1993), *Metaphysics: Books gamma, delta, and epsilon* (Clarendon Press).

Barnard, R and Horgan, T (2006), 'Truth as mediated correspondence', *The Monist*.

David, Marian (2004), 'Don't forget about the correspondence theory of truth', *Australasian Journal of Philosophy,* 82 (1), 42 – 47.

David, Marian Alexander (1994), 'Correspondence and Disquotation', *An Essay on the Nature of Truth* (Oxford University Press), 206.

Devitt, Michael (1997), 'Realism and Truth', (Princeton University Press), 371.

Engel, Pascal (2002), 'Truth', (McGill Queens Univ), 177.

Field, Hartry H (1972), 'Tarski's Theory of Truth', *The Journal of philosophy* (69: Journal of Philosophy, Inc.), 347-75.

Fodor, Jerry A (1987), 'Psychosemantics: The problem of meaning in the philosophy of mind.', (MIT Press).

Fumerton, Richard A (2002), 'Realism and the Correspondence Theory of Truth', (Rowman & Littlefield), 149.

Horgan, Terence (2001), 'Contextual semantics and metaphysical realism: truth as indirect correspondence', *The Nature of Truth: Classic and Contemporary Perspectives*.

Lewis, David (2001), 'Forget about the 'correspondence theory of truth'', *Analysis,* 61 (272), 275–80.

Merricks, Trenton (2007), *Truth and Ontology* (Oxford University Press).

Millikan, Ruth Garrett (1984), 'Language, Thought, and Other Biological Categories', *New Foundations for Realism* (MIT Press), 355.

Tarski, Alfred (1936), 'The concept of truth in formalized languages', in A. Tarski (ed.), *Logic, Semantics, Metamathematics* (Oxford University Press), 152–278.

2

Engel vs. Rorty on Truth

ERIK J. OLSSON

Abstract My concern in this paper is with a debate between Pascal Engel and Richard Rorty on truth, as documented in *What's the Use of Truth?* There Engel defends, against his opponent, the view that truth plays a crucial role in our intellectual and daily lives. In the present paper, I attempt an evaluation of the debate, which can give the superficial impression of ending in a stand-off, from the point of view of a general theory of rational goal-setting. This move has the notable effect that Rorty's central argument against truth being a goal of inquiry is undermined, and that Engel's truth-friendly position is correspondingly vindicated.

1. Introduction

I would like to start by paying tribute to Pascal Engel's great contribution to analytic philosophy in France, where he has been for a long time a leading analytical philosopher, as well as in Europe at large. Trained in the continental school, he at some point converted to analytic philosophy and has remained an enthusiastic devotee ever since. No doubt his change in view was in part caused by an insider's realization that intellectual anarchy looms if central distinctions such as that between true and false, rational and irrational, objective and subjective are neglected – a failure for which some French thinkers like Derrida and Foucault have demonstrated a particularly tragic disposition.

On a personal note, my first encounter with Pascal must have at one of the many workshops he arranged in the 1990s at the Sorbonne, at the time a center for the continental school thought, as part of his persistent efforts to introduce analytical philosophy in his home country. I am greatly indebted to him for repeatedly inviting me to give talks in this stimulating setting although at the time I was a doctoral student and not an established researcher. Since then I have had the privilege to meet and discuss with Pascal on many occasions as well as to enjoy his friendship and kindness.

I turn now to the actual topic of this article. In a stimulating little book entitled *What's the Use of Truth?*, edited by Patrick Savidan, Pascal Engel and Richard Rorty engage in a debate which, according to an observer quoted on the back cover, "starts off in university tweed and ends up in a street fight". It is not difficult to foresee that there should be a certain philosophical tension between the two because their intellectual trajectories could hardly have been more divergent. Where Engel converted from continental to analytic philosophy, Rorty made the opposite intellectual journey. From Engel's perspective Rorty must be something of an intellectual conundrum: how could he, having had the great fortune of being schooled in the analytical tradition in the company of some of its most distinguished American practitioners, even contemplate taking seriously thinkers such as Derrida and Foucault? Similarly, Engel's decision to distance himself from the continental tradition must appear, from where Rorty stands, as equally incomprehensible and erratic.

After reviewing the Engel-Rorty debate in section 2, I will make, in section 3, the move to invoke, as a vehicle of conceptual clarification and reconstruction, the theory of goal-setting as it has been developed and applied in management science and technology. The benefit of this framework, which may strike the reader as an unlikely source of philosophical enlightenment, is that it in fact provides a standard vocabulary for discussing goal rationality that

is richer and more precise than the apparatus typically used by philosophical authors writing on the subject. As I argue, this richness and precision can inform the evaluation of the slippery claim that truth is the goal of inquiry, upon which so much of the Engel-Rorty controversy depends. A concrete illustration is given in in section 4 where I consider Peirce's view on the aim of inquiry as a preliminary to identifying, in section 5, what I take to be Rorty's central thesis on the subject.[1]

2. The Engel-Rorty debate on truth

The dialogue takes off with a main statement by Engel, the author *The Norm of Truth* and *Truth*, reflecting on the curious conflict between our general longing for truth, on the one hand, and the deep skepticism regarding that very concept expressed by some intellectuals, on the other. Engel recalls having observed firsthand the personification of this tension in Foucault (Rorty and Engel 2007, p. 2):[2]

> "It always used to astonish me, when I was attending Michel Foucault's courses at the Collège de France in the 1970s, to hear him explaining to us that the notion of truth was no more than an instrument of power, and that, since all power was bad, truth could only be the expression of some malign intent, and then see him marching in demonstrations under banners bearing the slogan Truth and Justice."

Engel proposes, tentatively, that intellectual skepticism regarding truth does not concern its role in our daily affairs but rather Truth as a metaphysical concept: "We dislike preachers who speak in the name of Truth, but we pay attention to everyday truths, like the ones in the periodic statement of our bank balance" (p. 3). But, he asks, what is the concept we are meant to reject and what is the concept that can supposedly still cling to? And is it really coherent to reject the one while retaining the other?

One could ask the further question: why engage with Rorty on these matters? As Engel notices, Rorty has defended ideas similar to those expressed by Derrida, Foucault and others but without succumbing to their abstruse prose and literary ambitions, writing in the more accessible and systematic style of the analytical philosopher he used to be. Where Derrida and Foucault simply

[1] Sections 3-5 draw on Olsson (in press).

[2] Page references are to *What's the Use of Truth?* unless otherwise indicated.

state their views, one finds in Rorty's work explicit arguments much to the same effect, drawing on American pragmatists like William James and John Dewey but also on broader thinkers like Quine, Davidson and Sellars. As Engel observes, "Rorty claims a place in the American pragmatist tradition", adding that "his pragmatism is very different from that of the founder of this current, C. S. Peirce". In a later section, I will problematize the last statement. Although it is indeed true that Rorty's pragmatism is in many ways different from Peirce's, both denounce the idea that truth figures essentially in the goal of inquiry, and they do so for very similar reasons, or so I will argue. Finally, because of his background "Rorty knows exactly what he is talking about when he discusses the thesis of analytic philosophy" (p. 5) increasing the prospects of a fruitful and informed debate.

Engel proceeds (pp. 6-8) to describe what he takes to be Rorty's view on truth, as summarized in the following catalogue:

1) The notion of truth has no explanatory use and does not cover any essence or substance or designate any profound substantial or metaphysical property or any object (the True).

2) The traditional correspondence or realist theory of truth is "devoid of meaning".

3) The debates between realism and antirealism are "hollow".

4) There is no distinction to be made between truth and justification, and the latter "is nothing other than agreement among the members of a group or a community, and there is no ultimate, final agreement or ideal convergence of statements".

5) The concept of truth being empty, truth cannot be a norm of scientific or philosophical inquiry or an ultimate goal of our search. A fortiori, neither can it be a value.

6) We cannot hope for a naturalist, reductionist theory of representation and reality.

7) Rather than objectivity and truth, the values that are to be pursued are those of solidarity, tolerance, liberty, and a sense of community.

As Engel notices, Rorty relies in his argumentation on a deflationist or minimalist theory of truth, according to which the legitimate uses of the word true are exhausted by the following list:

a. an endorsing or performative use as in "your belief is true"

b. a cautionary use as when one says "your belief is justified but it is not true"

c. a disquotational use in the sense of the Tarski equivalences ("p is true" is true if and only if p")

Engel gives a detailed account of where he thinks he and Rorty advocate different views. This is not the place to give a complete coverage of Engel's intricate argumentation. Rather, I will be mainly concerned with claims 4) an 5) above, i.e. on Engel's reasons for rejecting Rorty's theses that truth and justification are in a sense indistinguishable and that truth cannot be the goal of inquiry.

The rejection of what he calls the "argument from indistinguishability" plays a central role in Engel's critical reflections on Rorty. As an "initial response", Engel suggest the following indirect approach. Suppose it were true that the words true and justified (or warrantedly assertible) mean the same thing.

> "If that were the case, the negation of a statement would be the same thing as the affirmation that it is not warrantedly assertible. But to say that the Loch Ness monster does not exist is not the same thing as saying that it is not warrantedly assertible that the Loch Ness monster does not exist." (p. 19)

Engel proceeds, in his second line of criticism, to concede that there is a close link between justification and truth, but this link, he claims, is not one of identity (ibid.).

> "When one has reasons, guarantees, or justifications for believing that P, these are justifications for believing that P *is true*. But this does not entail that saying 'I am justified in believing that P' and saying 'P is true' signify the same thing. On the contrary, this shows that, when one has reasons to assert or believe a proposition, one has reasons to believe that it is *true*. One cannot therefore maintain that *true* and *justified* convey the same thing, since justified *presupposes* the very notion of truth."

Yet, Engel may have misinterpreted Rorty on this particular point. As I understand him, Rorty is not claiming that justification and truth can be strictly speaking identified. The point is rather that once we are in possession of the

one, we cannot tell the difference between that situation and one in which we are in possession of the other. At the end of his response to Engel's first statement, Rorty clarifies his position thusly: "I am perfectly ready to admit that one cannot identify the concept of truth with the concept of justification or with any other", adding that "this is not a sufficient reason to conclude that the nature of truth is an important or interesting question" (p. 45). In Engel's favor, it should be noted that Rorty was not always this explicit about the content of his actual thesis.

Finally, Engel considers the possibility of a collective brainwashing, asking "would we say in that case that our beliefs were justified in relation to one audience but not in relation to another?" His answer is no because "we would say that our beliefs are justified but false". In response, Rorty could probably agree that, in the case of collective brainwashing, we would say that our beliefs are justified but false. It is only that he would reinterpret these words in a way that does not refer to truth or falsity. This is also what I take to be the gist of actual Rorty's response to this particular point, in which he does little more than restate his position.

I am inclined to think of the subsequent discussion of Rorty's rejection of truth as being a goal of inquiry as a central part of Engel's critique (pp. pp. 22-). In this connection, Engel ascribes to Rorty the following argumentative chain:

A. If there is a truth as norm or goal of inquiry, then there must be a real property in it such as "the truth of our assertions".

B. There is no real property of this kind.

C. Thus there is no truth as norm or goal of inquiry.

But premise A is false, Engel thinks, "because the fact that there does not exist a property such as the correspondence between our utterance and reality does not entail, from the point of view of inquiry, that we are not seeking to attain a certain objective" (p. 22). In other words, "[t]he notion of a norm does not presuppose the existence of the property in question or its reality" (p. 23). This seems to me to be on the right track. However, it should be bore in mind that Engel concurs with Rorty, perhaps merely for the sake of the argument, that truth is not a goal of inquiry in a "profound" sense of being a Supreme Value (p. 23). Rather, what Engel has in mind is "the relatively innocent sense in which we say that our beliefs aim at truth because it forms part of the concept

of belief that if we discover that one of our beliefs is false we try to change it" (ibid.).

This is a point where I believe that Engel may be conceding too much to Rorty. While I acknowledge that our practice of belief is governed by the norm mentioned by Engel, I will argue below that Rorty's, and before him Peirce's, argument to the effect that there is no profound sense in which truth is the goal of inquiry is unconvincing, or even refutable. However, because the matter is delicate and philosophical pitfalls abound we need to approach it more systematically than is typically done.

3. A general theory of goal-setting rationality

Goal rationality has been studied extensively in management theory, where it is central in so-called MBO, an acronym standing for Management By Objectives (e.g. Mali, 1972). This has led to the development of a common approach, codified in the acronym SMART, according to which goals should be Specific, Measurable, Achievable, Realistic and Time-bound. This theory has been refined and systematized by Sven Ove Hansson and his research group at the Royal Institute of Technology (KTH) in Stockholm (e.g. Edvardsson and Hansson, 2005). In the following, I will refer to the framework developed by Hansson et al as SMART+, signaling that it represents an updated, philosophically more sophisticated, version of the original SMART conditions. (This is my terminology, not theirs.) The KTH group has used the theory in its study of environmental (Edvardsson, 2004) and transport objectives (Rosencrantz et al, 2007).[3]

The thesis that a theory originating in management science could have any bearing whatsoever on a philosophical issue as sublime and profound as that of truth may seem chocking to some. I intend to prove this argument from guilt by association wrong. It is only from an implausible "first philosophy" standpoint that one could object to philosophy being informed by other parts of science. For an epistemological naturalist like myself, there is no problem in principle with borrowing ideas and concepts from other fields if there is some concrete hope that greater clarity can thereby be achieved. From this perspective, management science seems to be as good a field as any other.

A goal is typically set for the purpose of achieving it. We will say that a goal is *achievement-inducing* if setting it furthers the desired end-state to which

[3] The account of SMART+ in this section draws mainly on Edvardsson and Hansson (2005). The reader is advised to consult that paper for additional references.

the goal refers. Thus the goal of becoming rich is achievement-inducing (for me) if my setting that goal makes it more likely that I will in fact become rich, e.g. by inspiring me to focus on accumulating wealth, which may eventually lead to my actually becoming wealthy. As a first approximation, a goal G is achievement-inducing for a subject S just in case the probability that S attains the goal G is increased by S setting herself the goal G, i.e., in semi-formal terms, just in case P(S attains the goal G | S sets herself the goal G) >P(S attains the goal G).

Edvardsson and Hansson proceed to use the notion of achievement-inducing to define the concept of goal *rationality*: in their view, a goal is rational if it performs its achievement-inducing function (sufficiently) well. This is a satisficing rather than an optimizing notion of rationality (Simon, 1956). Evidently, in order to be achievement-inducing and therefore, on this proposal, rational a goal should guide as well as motivate action. One could also argue that rational goals serve to coordinate actions among several agents, but that aspect will not play any major role in the following.

There is certainly more to be said about this proposed concept of goal rationality. First, as it stands it begs the question against visionary goals such as "world peace" or, in general, goals that cannot be fully attained. An example from Swedish transport policy is the so-called "vision zero" goal stating that, in the longer run, no one should be killed or seriously injured as the effect of a traffic accident (Rosencrantz, Edvardsson and Hansson, 2007). A goal that cannot be attained is not achievement-inducing and is therefore irrational according to the proposed definition. However, there is an obvious way to avoid this untoward result by a suitable redefinition of the concept of achievement inducement. A goal G is achievement-inducing for a subject G, on the revised proposal, just in case the probability that S attains the goal at least partially or, alternatively, at least approaches the attainment of G, is increased by S setting herself the goal G.

Second, achievement-inducement, even in the less demanding sense, cannot be all there is to goal rationality. If it were, the rational thing to do would be to set oneself trivial goals that can be easily attained: poking one's nose, lifting one's hand, and so on. The likelihood that I manage to raise my hand if I set myself the goal to do so is very close to one. Goals which are more difficult to achieve, such as getting oneself a solid education, would be dismissed as irrational. However, the proposal does make good sense as a tie-breaking condition in a setting where there are already a number of candidate goals that have been singled out on the basis of other considerations. Faced with a set of goals that are equally attractive in other respects, it is reasonable to

select one that is achievement-inducing.

With these clarificatory remarks in mind, what does it mean, more specifically, to say that a goal can guide and motivate action? It is useful at this point to distinguish between three types of criteria of goal-rationality: those related to what the agents *know*, what they *can do* and what they *want to do*. From the first, epistemic perspective, goals should be *precise* and *evaluable*. A goal such as "achieving a better society" fails on the first account, that of precision. That goal is not very useful for guiding action unless supplemented with more precise instructions. There are at least two different aspects of precision: directional and temporal. A goal is *directionally complete* if it specifies in what direction one should go in order to reach the goal. Take for example the goal to substantially decrease the number of unemployed in Sweden. That goal is directionally complete because it suggests in what direction progress towards the goal is to be made. If employment has decreased, then the goal has been approached or achieved, otherwise not. A goal is *temporally complete* if it specifies the timeframe within which it should be attained.

A goal is *end-state evaluable*, moreover, if it is possible to know whether it has been achieved. The goal to reduce a pollutant in the atmosphere to a certain level that is far below what can be measured would fail to satisfy the criterion of end-state evaluability. A goal is *progressively evaluable* if it can be determined how far we are from satisfying it. This property of goals is crucial in determining whether a certain course of action should be maintained, changed or given up. It has also been argued that such feedback enhances the agent's motivation so that she will make an intensified effort to act in ways that further the goal.

For an illustration, suppose my goal is to reach Geneva by the end of the day. In order for that goal to be rational, I must be able to determine whether or not this is the city I am actually in by the end of the day. However, in many situations it is not enough to be able to determine whether or not the goal state has been fully achieved. In the example, I must also be able to tell whether I am travelling in the right direction, and how far I have left to go. In particular, if a goal is distant, or difficult fully to achieve we need to be able to judge the degree of success in approaching the goal. In other words, degrees of *partial attainment* must be distinguishable.

The second aspect of goal rationality concerns what the agent *can do*. It is reflected by the requirement that a goal should be *attainable*, or at least *approachable* (i.e. attainable at least to some degree). The goal to become a wizard (in the sense of a person with true magical powers) would not be classified as attainable or even approachable. There are at least three dimensions of ap-

proachability: *closeness*, *certainty* and *cost*. The dimension of closeness is the most obvious one. It concerns how close to the goal it is possible to come. The goal to achieve a perfectly just society is probably not fully achievable, and would therefore qualify as utopian, but it can be approached by acting in ways that increase social justice.

The third aspect of goal rationality is the volitional one. It concerns what we *want to do*. Goals, in order to be rational, should be motivating. Setting ourselves the goal should motivate us to act in a way which furthers the realization of the goal state. The motivation that a goal may give rise to in the agent can be characterized according to degree of intensity or durability. Studies indicate that goals are more action-generating when they are explicit and specific, and that such goals are more likely than do-your-best goals to intensify effort. There is also evidence suggesting that specific and challenging goals lead people to work longer at a task. We have already mentioned a connection between evaluation and motivation: when people can check how they stand in relation to a goal, their motivation to carry out the task often increases.

An insight into the nature of goal-setting emerging from SMART+ is that the criteria of rational goal-setting may conflict in the sense that the satisfaction of one criterion to a high degree may lead to a failure to satisfy substantially some other criterion. The probably most common type of such conflicts are occasioned by the fact that some of the properties that make a goal action-guiding may at the same time make it less capable to motivate action. Consider, for example, the following two goals (Edvardsson and Hansson, 2005):

(1) The team shall win 12 out of 20 games with a least a two goal advantage, 3 out of 20 games with at least a one goal advantage, and never lose a game with more than one goal.

(2) The team shall beat all opponents hands down.

Here, the second goal, though less action-guiding than the first, is plausibly more achievement inducing, and therefore more rational, because of its greater action-motivating capacity.

In general, visionary and utopian goals are more likely to motivate action than less visionary goals, which on the other hand may be more action-guiding. This point is elaborated in Edvardsson and Hansson (2005). The task of goal-setting therefore may very well involve a trade-off between goals that are action-motivating and goals that are action-guiding. This may lead to the formulation of one single goal reflecting this compromise. However, it is often

a better idea to adopt not a single goal but a whole system of goals at different levels. As Edvardsson and Hansson point out: "One way of balancing the criteria so as to optimize goal realization is to adopt goal systems in which goals are set on different levels in order to supplement each other. In this way visionary and highly motivating goals can be operationalized through more precise and evaluable subgoals, or interim targets." (ibid., p. 359)

On first sight, goal rationality in inquiry seems attractively simple: the goal of inquiry is simply to find the truth. If this were correct, there wouldn't be any need for a theory of goal-setting in this domain. On closer scrutiny, however, considerable complexity emerges. For one, the goal to find the truth does not by itself suggest any very definite course of action; it does not specify in what direction one should go in order to reach the goal, except possibly that one should use a method that is reliable – one that is likely to lead to true beliefs. Still, the goal itself does not indicate what those methods are. In fact, not only directional completeness but also the other aspects of goal rationality identified in the SMART+ model make good sense as principles governing goal rationality in inquiry, and it is not clear that they are maximized by the goal of truth.

For another example, it would clearly be desirable in science to have a goal that is end-state evaluable in the sense that it is possible to know whether it has been achieved. Once more, the goal of truth is not an obvious candidate. Similarly, we would like scientific goals to be temporally complete, progressively evaluable, attainable, and we would be happy to have goals that exert the proper motivational force on the inquirer. Finally, there seems to be no reason to think that science is devoid of goal conflicts. For instance, the goal of truth could be satisfied by simply adopting a trivial theory, one which is logically true. To avoid this, we need the further goal of informativity. But as many epistemologists have observed (e.g. Levi, 1967), if we decide to adopt both goals as ultimate ends this is likely to lead to a goal conflict since a more informative theory is often less likely to be true. A theory that is very specific regarding the causes of a particular kind of cancer may thereby be less likely to be true than a less committed theory.[4] These remarks will, I hope, suffice as a background and dialectical bridge to the following reconstruction of Peirce's argument to the effect that truth cannot be a goal of inquiry.[5]

[4] For a related issue and some complications, see Bovens and Olsson (2002).

[5] For the purposes of simplicity and definiteness, I will in the following take "truth" in its objectivist or realist sense as referring to correspondence with an external reality, although I conjecture that much of the reasoning that follows would survive a weakening to "empirical adequacy", or the like.

4. Peirce on belief as the goal of inquiry

In a famous essay, Peirce argues that, contrary to the received view, the goal of inquiry is not truth, or true belief, but merely belief or "opinion" (Peirce, 1955, pp. 10-11):

> [T]he sole object of inquiry is the settlement of opinion. We may fancy that this is not enough for us, and that we seek, not merely an opinion, but a true opinion. But put this fancy to the test and it proves groundless; for as soon as a firm belief is reached we are entirely satisfied, whether the belief be true or false. And it is clear that nothing out of the sphere of our knowledge can be our object, for nothing which does not affect the mind can be the motive for mental effort. The most that can be maintained is, that we seek for a belief that we shall *think* to be true. But we think each one of our beliefs to be true, and, indeed, it is mere tautology to say so.

We recall that, for Peirce, belief or opinion is, by definition, that upon which an inquirer is prepared to act. Hence, Peirce is proposing to reduce the goal of scientific inquiry to the goal of attaining that upon which we are prepared to act.

In the latter part of the quote, Peirce seems to be maintaining that the true state of things does not affect the mind and therefore cannot be the motive of mental effort. But the claim that the facts of the matter do not affect the mind is a counterintuitive one. When I look out the window, I come to believe that there is a tree just 10 meters away. Normally, this belief is caused by the tree, or the fact that there is a tree, which is thus affecting my mind.[6]

On another interpretation, Peirce is thinking of objective truth as essentially "mind-independent". If so, one could be led to think that it follows trivially that objective truth cannot affect the mind, for nothing that is mind-independent can if that is what "mind-independent" means. But this is an irrelevant sense of mind-independence. In a less trivial sense, something is mind-independent and objective if it does not depend entirely on our will. Truth is mind-independent in the latter sense but not in the former. What is true – for example that there is a tree outside the window – does not depend

[6] It could be objected that Peirce is here using "truth" in a technical sense, signifying what is collectively accepted by all researchers once scientific inquiry has come to an end. Truth in that sense presumably does not exert any direct influence on a particular mind now. Still, this is an implausible interpretation of Peirce in the present context, as there is no concrete sign that truth should be given any special technical meaning.

entirely on our will but it is still something that can affect us in various ways, and typically does so through our observations.

Peirce is right, though, in stating that once we believe something, e.g. that there is a tree out there, we cannot, pending further inquiry, distinguish the state we are in from a state of *true* belief. If S believes that p, or believes truly that p, she cannot tell whether she has attained the first goal or the second. She will, from the position of the goal end state, judge that she believes that p just in case she will judge that she believes truly that p. Peirce can be understood as maintaining that this fact alone makes it more rational, or appropriate, to view the goal of inquiry in terms of fixing belief rather than in terms of fixing *true* belief. Is that correct?

Let us look at the matter from a more abstract perspective. We will say that two goals G_1 and G_2 are *end-state evaluation equivalent* for a subject S if, upon attaining one of G_1 or G_2, S cannot tell whether she attained G_1 or G_2. Peirce, in the argument under scrutiny, is relying on the following principle:

> (Peirce's Principle) If (i) G_1 and G_2 are end-state evaluation equivalent for a subject S, and (ii) G_1 is logically stronger than G_2, then G_2 is more rational, or appropriate, than G_1 for S.

Is this principle valid as a general principle of goal rationality? I will argue that it is not. Here is a counterexample:

Suppose that P is a pollutant that is dangerous to humans and that M is a device which indicates whether or not the amount of P in the air exceeds the limits that have been set by an international body. Moreover, there is no other device that can be used for this purpose. However, M is not fully reliable and it sometimes misfires. Let G_1 be the goal of using the device M *and* successfully determining whether the air is free of P-pollution; and let G_2 be merely the goal of using the device M. G_1 and G_2 are end-state evaluation equivalent for the measuring person S: upon attaining G_1 or G_2 she cannot distinguish one from the other. Moreover, G_2 is logically entailed by G_1. It would follow from Peirce's Principle that G_2 is more rational than G_1. But this conclusion can be questioned. It is true that G_2 is more easily attained than G_1. But G_1 is surely more inspiring than G_2; it is, to use Peirce's own expression, a stronger "motive for mental effort". It cannot, therefore, be concluded that G_2 is more rational, or achievement-inducing, than G_1. Hence, the principle presupposed by Peirce is plausibly not generally valid. This observation is sufficient to undermine Peirce's argument that the goal of belief is more rational, or appropriate, than the goal of true belief.

Indeed, the goal of true belief, or the goal of truth for short, does sound more inspirational than the goal of settling belief. Many people, not least those equipped with a scientific mind, will go to almost any length to find the truth of the matter, sometimes even in practically insignificant affairs. Disregarding the special case of religious faith, comparatively few would be willing to incur similar personal and other costs for the sole gain of settling a corresponding opinion.

Apart from the general invalidity of Peirce's Principle, there may be other differences between the goal of belief and that of true belief that are worth attending to. One such factor is a difference in precision. We recall that a goal is said to be directionally complete if it specifies in what direction one should go in order to reach the goal. We have noted that the goal of true does not do terribly well on this score. But it might still do better than the goal of belief. For the goal of true belief suggests, albeit imperfectly, that the belief be fixed, not by any old method, but by one that is likely to establish the truth of the matter. This would suggest to the inquisitive mind such things as evidence-gathering, hypothesis-testing, the use of scientific instruments, and so on. The goal of belief does not suggest as vividly any particular course of action. It is compatible with using a wider range of methods, including methods that are not truth-oriented but focus, say, on the systematic disregard of contravening evidence.

Finally, there is a difference between the two goals on the ability dimension, concerning what we can do to approach the respective goals. This is related to the presumed difference in directional completeness. The goal of belief can be approach and evaluated along one dimension only: degree of belief. The stronger our belief is, the closer we are to achieving the goal of (full) belief. The goal of truth, by contrast, can in addition be approach, at least in principle, along the dimension of truth-likeness: the closer we are to the truth, the closer we are to achieving the goal of true belief *ceteris paribus*.

5. Rorty on justification as the goal of inquiry

After this warm-up on Peirce, I turn now to Rorty himself and an argument presented in a paper from 1995, drawing partly on earlier work (e.g. Rorty, 1986), to the conclusion that truth is not legitimately viewed as the goal of inquiry. This is a conclusion also drawn by Peirce, as we saw, but where Peirce thought that the goal of truth should be replaced by the goal of belief, Rorty proposes that the proper replacement is rather *justified* belief. Apart from this

notable difference, their respective arguments are strikingly similar.

The starting point of Rorty's 1995 paper is the following declaration (p. 281):

> Pragmatists think that if something makes no difference to practice, it should make no difference to philosophy. This conviction makes them suspicious of the philosopher's emphasis on the difference between justification and truth. For that difference makes no difference to my decisions about what to do. If I have concrete, specific doubts about whether one of my beliefs is true, I can resolve those doubts only by asking whether it is adequately justified – by finding and assessing additional reasons pro and con. I cannot bypass justification and confine my attention to truth: assessment of truth and assessment of justification are, when the question is about what I should believe now (rather than about why I, or someone else, acted as we did) the same activity.

He adds, a few pages later on (p. 286):

> The need to justify our beliefs and desires to ourselves and our fellow agents subjects us to norms, and obedience to these norms produces a behavioral pattern that we must detect in others before confidently attributing beliefs to them. But there seems no occasion to look for obedience to an additional norm – the commandment to seek the truth. For ... obedience to that norm will produce no behavior not produced by the need to offer justification.

Thus, in Rorty's view the goal of scientific inquiry is not truth but being in a position to justify one's belief. Rorty, moreover, views justification as essentially unrelated to truth, which in the end is a notion he favors dropping altogether (p. 299). One of the conclusions of his essay is that, on the Dewey-inspired theory which he advocates, "the difference between the carpenter and the scientist is simply the difference between a workman who justifies his action mainly by reference to the movements of matter and one who justifies his mainly by reference to the behavior of his colleagues" (ibid.).

Let us properly dissect this central line of reasoning, drawing on the theory of goal-setting previously introduced. Rorty, as quoted above, is contrasting two goals: the goal of attaining a true belief and the goal of attaining a justified belief. On the reading I would like to highlight, he is offering an argument that is similar to Peirce's argument for the propriety of the goal of belief, but –

again – for a slightly different conclusion. Rorty is pointing out that the goal of attaining a true belief and the goal of attaining a (sufficiently) justified belief are end-state evaluation equivalent from the point of view of the inquirer: once the inquirer has attained either of these goals, she cannot tell which one she attained. This much seems true. Yet Peirce's Principle is not directly applicable as it demands that, among the goals under consideration, one goal be logically stronger than the other. The two goals of true belief and justified belief are not at all logically related, at least not as justification is standardly conceived.[7]

Still, the goal of justified belief is plausibly more directionally complete than the goal of true belief, it specifies more clearly in what direction to proceed in order to satisfy the goal, and in the quote above this is a feature that Rorty highlights. On a plausible reconstruction, the general principle underlying Rorty's reasoning, then, is this:

> (Rorty's Principle) If (i) G_1 and G_2 are end-state evaluation equivalent for a subject S, and (ii) G_2 is more directionally complete than G_1, then G_2 is more rational, or legitimate, than G_1 for S.

But this principle shares the fate of Peirce's Principle of being plausibly generally invalid. Since the problem is similar in both cases, I shall not this time give an explicit counterexample. Suffice it to note that beside directional completeness, there are – as we have seen – several other aspects of a goal that play a part in determining its relative rationality or suitability. One such aspect is, to repeat, the motivational one. This aspect is interesting in this context because, as we noted, it often offsets the directional aspect. Goals that are strongly motivational are in practice rarely directionally complete, and vice versa. Thus many are motivated by goals such as achieving "world peace" or "a completely just society" and yet these goals do not *per se* suggest any particular cause of action. Conversely, goals that give detailed advice for how to act tend, as a matter of psychological fact, to be less inspirational.

As we have already noted, the goal of truth, though directionally less complete than the goal of justification, may still be more rational in virtue of its inspirational qualities. Hence, *pace* Rorty we cannot conclude, from the presumed fact that the goal of true belief and the goal of justified belief are end-state evaluation equivalent and the latter more directionally compete than the former, that the latter is also the more suitable aim.

[7] I am here assuming a standard fallibilist account of justification according to which a belief can be justified without being true.

Leaving Rorty's discussion aside, a natural view to adopt concerning the relation between the two goals of true belief and justified belief, from a SMART+ perspective, is that they could very well live side by side, supplementing each other: the goal of truth providing the visionary, motivating factor and the goal of justification playing the more action-guiding part. Drawing on the upshots of section 3, there are *prima facie* two ways of implementing this recommendation. One would be to adopt a system of goals wherein both goals figure, the goal of truth as a high-level goal and the goal of justification as lower-level goal, the latter operationalizing the former. The other way would be to compress the two goals into one goal, the goal, namely, to attain a justified true belief. The latter goal amounts, incidentally, to the goal of attaining *knowledge*, as that concept is traditionally conceived.

6. Conclusion

Reconnecting to the Engel-Rorty dispute, the first point I wish to make, based on the above considerations, is the marginal one that while Engel is right to point out that Peirce and Rorty advocate very different versions of pragmatism in general, they reject the proposal that truth is a goal of inquiry for reasons that are, in fact, striking similar. They both rely on an "indistinguishability" principle according to which a subject is unable to distinguish the goal of truth from some other, more mundane goal, an observation which is then taken to speak in favor of the latter as the more legitimate aim to pursue.

More important, I have tried to make likely that Rorty's argument against truth as a goal of inquiry, the force of which Engel seems to concede, possibly only for the sake of argument, is unconvincing from the perspective of the general theory of goal-setting. The main problem is that the argument relies on an empirical thesis which, in its general form, is plainly false or at least highly controversial: the thesis, namely, that visionary and utopian goals – to which we must count the goal of truth as a special case – do not affect practice. The general theory of goal-setting, and the empirical work upon which it partly relies, suggests that exactly the opposite holds: such goals do affect practice through the increased "mental effort", to borrow Peirce's phrase, which they induce in the subjects entertaining them. They do so to the extent that their disadvantages from a goal-setting perspective are in many cases offset.

To make the point more vivid, consider the following closing statement by Rorty in the the debate with Engel (pp. 44-45):

> "Trying to do the right thing will lead us to do just the same things we would do when we try to justify our actions to ourselves and others. We do not have any way to establish the truth of a belief or the rightness of an action except by reference to the justifications we offer for thinking what we think or doing what we do. The philosophical distinction between justification and truth seems not to have practical consequences. This is why pragmatists think it is not worth pondering."

We can now see that, far from being a priori certain, Rorty is here relying on a substantial empirical hypothesis about human psychology, a hypothesis which we have – once more – considerable reasons to doubt.

Very probably, then, the goal of truth should be cherished rather than shunned by pragmatists as a goal which, due to its inspirational qualities, is as practice-affecting as one could ever wish. It is indeed curious that this point has not yet, as it appears, received widespread recognition and acceptance. In his patient and insightful engagement with Rorty on these and other issues, Engel makes much in the direction of setting the record straight.

7. References

Bovens, L., and Olsson, E. J. (2002), "Believing More, Risking Less: On Coherence, Truth and Non-trivial Extensions", *Erkenntnis* 57: 137-150.

Edvardsson, K. (2004), "Using Goals in Environmental Management: The Swedish System of Environmental Objectives", Environmental Management 34 (2): 170-180.

Edvardsson, K., and Hansson, S. O. (2005), "When is a Goal Rational?", Social Choice and Welfare 24: 343-361.

Engel, P. (1991), *The Norm of Truth: An Introduction to the Philosophy of Logic*, University of Toronto Press.

Engel, P. (2002), *Truth*, Acumen Press.

Engel, P. (2007), "Main Statement by Pascal Engel", pp. 1-30 in *What's the Use of Truth?*, Savidan, P. (ed.), Columbia University Press.

Levi, I. (1967), *Gambling with Truth: An Essay on Induction and the Aims of Science*, Routledge and Kegan Paul, Ltd: London.

Mali, P. (1972), *Managing by Objectives: An Operating Guide to Faster and More Profitable Results*, John Wiley and Sons: New York.

Olsson, E. J. (in press), "Goal Rationality in Science and Technology: An Epistemological Perspective", in Hansson, S. O. et al (Eds.) *How Technology Shapes Science: Philosophical Perspectives on the Role of Technology in Science*, Springer.

Peirce, C. S. (1955) "The Fixation of Belief", pp. 5-22 in *Philosophical Writings of Peirce*, Buchler, J. (ed.), Dover Publications: New York. The manuscript was first published in *Popular Science Monthly* in 1877.

Rorty, R. (1986), "Pragmatism, Davidson and Truth", pp. 333-368 in *Truth and Interpretation: Perspective on the Philosophy of Donald Davidson*, LePore, E. (ed.), Oxford: Basil Blackwell.

Rorty, R. (1995), "Is Truth a Goal of Inquiry? Davidson vs. Wright", *The Philosophical Quarterly* 45, no. 180: 281-300.

Rorty, R. (2007), "Main Statement by Richard Rorty", pp. 31-46 in *What's the Use of Truth?*, Savidan, P. (ed.), Columbia University Press.

Rosencrantz, H. K., Edvardsson, K., and Hansson, S. O. (2007), "Vision Zero – is it Irrational?", *Transportation Research Part A: Policy and Practice* 41(6): 559-567.

3

La vérité regonflée? Réflexions sur le réalisme minimal de Pascal Engel *

CLAUDE PANACCIO

1. Vers une théorie de la vérité

La notion de vérité a fasciné Pascal Engel depuis les tout débuts de sa carrière philosophique. Son premier livre, en 1989, s'intitulait *La norme du vrai* et la deuxième partie (sur quatre) en était consacrée au thème « Vérité et signification ».[1] L'ouvrage sur Donald Davidson ensuite, *Davidson et la philosophie du langage* en 1994, accordait comme de raison une place centrale à la théorie de la vérité qui est l'un des accomplissements les plus connus du philosophe américain.[2] Et sans parler de nombreux articles ici et là[3], il y a eu encore le petit livre de 1998, *La vérité. Réflexions sur quelques truismes*[4], le débat de 2002 à Paris avec Richard Rorty, publié sous le titre *À quoi bon la vérité ?* et traduit depuis en plusieurs langues[5], et surtout, en 2002, aussi, son livre en anglais, *Truth*, sur lequel je me concentrerai ici principalement.[6] L'ouvrage se présente

*Une première version de ce texte a été présentée lors des Conférences Hugues Leblanc 2008 de l'Université du Québec à Montréal, dont Pascal Engel était l'invité d'honneur.

[1] P. Engel, *La norme du vrai. Philosophie de la logique*, Paris, Gallimard, 1989.

[2] P. Engel, *Davidson et la philosophie du langage*, Paris, PUF, 1994.

[3] Par exemple : P. Engel, « Truth and the aim of belief », dans G. Gillies (éd.), *Laws and Models in Science*, Londres, King's College, 2005, pp. 77-97 ; « Is truth effable », dans R. E. Auxier et L. E. Hahn (éds.), *The Philosophy of Jaakko Hintikka*, La Salle, Ill., Open Court, 2005, pp. 625-641 ; « Vérité, croyance et justification : propos d'un béotien dogmatique », dans A. Wald Lasowski (éd.), *Pensées pour le siècle*, Paris, Fayard, 2008, pp. 212-234.

[4] P. Engel, *La vérité. Réflexion sur quelques truismes*, Paris, Hatier, 1998.

[5] P. Engel et R. Rorty, *À quoi bon la vérité ?* Paris, Grasset, 2005.

[6] P. Engel, *Truth*, Montréal/Kingston, McGill-Queen's University Press, 2002.

en quatrième de couverture comme un « introduction critique aux questions philosophiques contemporaines en théorie de la vérité », mais Pascal Engel, en réalité, y met en place sa propre théorie, qu'il appelle le *réalisme minimal*, une position qu'il avait commencé à développer dès *La norme du vrai*, mais qui trouve là sa formulation la plus achevée.

Comme dans ses autres travaux, Engel développe son approche à travers une discussion riche et précise de ce qu'il y a de plus pertinent sur la question dans la philosophie analytique depuis les Frege, Russell, Wittgenstein, Quine, Davidson et Putnam jusqu'aux écrits les plus récents, dont il est, comme d'habitude, exceptionnellement bien informé, qu'il s'agisse de Crispin Wright, Hartry Field, Paul Horwich, John McDowell, David Wiggins ou d'autres. Cette érudition, cependant, ne cache jamais qu'il y a là une démarche unifiée, une perspective fermement assumée et surtout une *théorie*, qui s'articule en un certain nombre de thèses précises.

Une des composantes de la ligne qu'adopte Engel ? je le dis d'emblée parce que c'est un élément majeur de sa démarche, mais sur lequel je ne reviendrai pas dans ce qui va suivre ?, c'est bien évidemment la polémique, la « dispute » oserais-je dire, qu'il poursuit depuis des années avec le courant postmoderniste français, pour lequel la notion même de vérité n'est qu'un leurre assez suspect. Au cœur de son ouvrage de 1997, *La dispute*[7], où la notion de vérité occupe une place centrale, ce débat est aussi présent dans *Truth*, du moins en filigrane. Foucault notamment entre en scène dès l'introduction, avec Heidegger et Nietzsche, pour reparaître ensuite dans les dernières pages du livre, où il est décrit comme l'« un des plus flamboyants représentants » de cet historicisme postmoderniste que Pascal Engel entend récuser.[8] Mais entre ces deux apparitions spectrales du trio de sorcières, le scénario de *Truth* se développe à l'intérieur du paradigme de la philosophie analytique.

Le premier chapitre est consacré aux théories « classiques » de la vérité de Wittgenstein, Russell et d'autres, Engel les qualifie de théories *substantielles* parce qu'elles tiennent la vérité pour une propriété véritable, objective, indépendante de l'esprit et que l'on peut caractériser ou définir de façon éclairante dans une théorisation qui a une portée réelle, métaphysique même, que ce soit dans les théories de la vérité correspondance, dans les théories cohérentistes ou même dans certaines théories pragmatistes comme celle de Peirce. Il soulève à propos de chacune un certain nombre de difficultés importantes, de

[7] P. Engel, *La dispute. Une introduction à la philosophie analytique*, Paris, Éditions de Minuit, 1997.

[8] P. Engel, *Truth*, *op. cit.*, p.149. Toutes les traductions françaises de citations tirées de cet ouvrage dans la suite du présent article sont de moi.

sorte qu'à la fin du chapitre, si l'on suit l'auteur, on ne met plus trop d'espoir de ce côté-là.

Le chapitre 2 se tourne vers les approches dites « déflationnistes », qui soutiennent que la vérité n'est pas une propriété réelle de quoi que ce soit. Les expressions comme « il est vrai que » et le prédicat de vérité ? le terme « vrai » ? ne servent à leurs yeux qu'à marquer l'assentiment du locuteur à l'endroit de certains contenus. Dans les versions les plus radicales du déflationnisme, une expression comme « il est vrai que » n'est rien d'autre qu'un marqueur de force illocutoire. Il y a certes des positions déflationnistes plus modérées que d'autres, mais l'idée de base est toujours que du point de vue du *contenu* d'information, « il est vrai que p » ou « p est vrai » n'ajoutent rien à « p » tout court et n'introduisent, en particulier, aucune nouvelle référence à une propriété ou une relation objective quelconque. Mais de nouveau Engel soulève toutes sortes de difficultés et l'on semble dans l'impasse à la fin de ce chapitre puisque ni les théories qui gonflent la vérité (les théories substantielles) ni celles qui la dégonflent (le déflationnisme) ne paraissent avoir quelque chance de succès.

Que faire donc ? Les trois derniers chapitres, et surtout le 3 et le 4, montrent la voie et mettent en place la solution qu'Engel favorise, le *réalisme minimal*. Cela consiste, pour le dire caricaturalement à stade-ci, à dégonfler la notion de vérité dans un premier temps, mais pas complètement (c'est la composante minimaliste de la théorie) pour ensuite la regonfler (c'est la composante réaliste). Regardons-y de plus près.

2. Le réalisme minimal

Engel lui-même caractérise cette approche en termes un peu semblables dans la conclusion de *Truth* :

> … l'approche réaliste minimale […] permet à notre concept de vérité de rester mince sans nous empêcher d'accepter le réalisme quant à la véri-aptitude [*truth-aptness*]. Cela nous a conduit à un regonflement, ou une resubstantialisation du concept de vérité et de la propriété qu'il dénote. Mais la « substance » ainsi réintroduite n'est pas celle que visaient les tentatives définitionnelles des théories traditionnelles.[9]

[9] *Ibid.*, p. 147.

Le problème que je soulèverai, autant le dire tout de suite, est qu'il est bien difficile à la notion de vérité de rester mince une fois qu'elle est regonflée ! Mais il faut d'abord dire plus précisément en quoi consiste le réalisme minimal d'Engel. Je le ferai en commentant les sept thèses par lesquelles l'auteur le définit, quatre pour la composante minimaliste et trois pour la composante réaliste.[10]

> *Thèse 1 : La notion de vérité est une notion « mince » qui se caractérise par un certain nombre de truismes sur l'assertion, la correspondance, etc.*

C'est la thèse minimaliste de base. Les truismes dont il est question sont inspirés de Crispin Wright[11] ; Engel déjà en avait donné une liste dans *La vérité* en 1998[12] :

> (a) Asserter un énoncé, c'est le présenter comme vrai.

On ne peut pas, en d'autres mots, asserter sérieusement un énoncé et refuser en même temps de le tenir pour vrai. C'est là d'entrée de jeu un problème majeur pour ceux qui prétendent récuser la notion de vérité, mais qui ne se font pas faute pour autant d'affirmer toutes sortes de choses.

> (b) « p » est vrai si et seulement si p.

C'est la célèbre condition d'adéquation matérielle à laquelle devait satisfaire selon Tarski toute théorie acceptable de la vérité.[13]

> (c) Les énoncés susceptibles d'être vrais ont des négations susceptibles d'être vraies.

À quoi l'on peut ajouter que ces énoncés peuvent également être insérés dans des conjonctions, des disjonctions et des conditionnelles susceptibles, elles aussi, d'être vraies.

> (d) Être vrai n'est pas la même chose qu'être justifié.

---

[10] *Ibid.*, p. 89.

[11] Voir C. Wright, *Truth and Objectivity*, Harvard, Mass., Harvard University Press, 1992.

[12] P. Engel, *La vérité*, *op. cit.*, p. 57.

[13] *Cf.* A. Tarski, « The semantic conception of truth and the foundations of semantics », *Philosophy and Phenomenological Research*, 4, 1944, pp. 341-376 ; trad. fr. : « La conception sémantique de la vérité et les fondements de la sémantique », dans G. G. Granger *et al.* (éds.), *Alfred Tarski. Logique, sémantique, métamathématique 1923-1944*, Paris, Armand Colin, 1974, vol. II, pp. 265-305.

L'idée de vérité n'est pas essentiellement épistémique. Dire que quelque chose est vrai, ce n'est encore rien dire quant à la façon dont nous le savons ni quant à ce qui justifie épistémiquement de le dire. C'est là un point crucial même si c'est un truisme, et à propos duquel on se heurte à de fréquentes confusions dans les débats philosophiques.

(e) Être vrai, c'est correspondre aux faits.

Encore faut-il prendre cet énoncé lui-même de façon relaxe et purement tautologique : dire « c'est un fait qu'il pleut » revient exactement au même que « c'est vrai qu'il pleut ». Tel quel, ce principe n'est pas censé nous engager sur la voie d'une métaphysique de la correspondance et d'une ontologie des faits et des états de choses comme celle de David Armstrong par exemple.[14]

Tel est le genre de truismes dont parle la thèse 1. Engel en propose une liste un peu différente dans *Truth*[15], mais elle demanderait des explications supplémentaires qui sont moins directement pertinentes pour mon propos et je m'en tiendrai ici à ces cinq-là. Deux remarques s'imposent tout de suite cependant. D'abord, la thèse minimaliste n'est pas que la notion de vérité doive satisfaire au moins à ces truismes, c'est qu'elle n'a pas besoin de satisfaire à plus : l'ensemble des truismes devrait nous fournir à toutes fins pratiques des conditions suffisantes pour la notion de vérité. Deuxièmement, truismes ou pas, la série proposée ci-dessus est déjà trop exigeante, parce que (b), la chose est bien connue, conduit à des paradoxes, celui du Menteur notamment : toute notion qui satisfait à la condition d'adéquation matérielle de Tarski (« p » est vrai si et seulement si p) est ultimement incohérente. La notion minimaliste de vérité telle qu'elle est présentée par Engel paraît donc *prima facie* condamnée. C'est un indice, déjà, de la nécessité d'une reconstruction théorique plus robuste que ce que laisse entendre au départ le minimalisme. La chose est majeure et Engel lui-même reproche à Paul Horwich de laisser de côté le problème des paradoxes[16], mais il n'y insiste guère non plus dans *Truth* et je laisserai moi aussi ce point de côté dans le présent contexte.

Thèse 2 : Le terme « vrai » ne constitue pas un simple marqueur d'assertion ou de décitation.

C'est ici que le minimalisme d'Engel, avec celui de Crispin Wright, s'écarte du déflationnisme radical : le terme « vrai » et les autres expressions apparentées

[14] Voir notamment D. Armstrong, *A World of States of Affairs*, Cambridge, Cambridge University Press, 1997 ; ou *Sketch for a Systematic Metaphysics*, Oxford, Clarendon Press, 2010.

[15] Voir P. Engel, *Truth*, *op. cit.*, p. 67.

[16] *Ibid.*, p. 51.

ne sont pas pour lui de purs marqueurs de force illocutoire. C'est pourquoi je disais plus haut que même à l'étape minimaliste de la théorie, la notion de vérité n'y est pas entièrement dégonflée. Engel là-dessus offre divers arguments assez classiques, et qui me paraissent décisifs. Le déflationnisme radical, notamment, s'accommode mal du fait qu'un énoncé de forme « *p* est vrai » peut être inséré dans une clause conditionnelle où « *p* » n'est pas asserté.

Thèse 3 : Les porteurs de vérité sont des propositions.

Engel prend le terme de « proposition » au sens qui est courant en philosophie analytique pour désigner des entités abstraites et non linguistiques qui peuvent être *exprimées* par des phrases et sont des *contenus* possibles de croyances. La thèse donc est litigieuse puisque l'existence même de telles entités est problématique. Je trouve préférable, pour ma part, d'élire des unités linguistiques comme porteurs de vérité, des phrases notamment, ou mieux encore des occurrences de phrases. Mais il ne sera pas nécessaire d'entrer ici très avant dans cette question controversée. Je me contenterai de signaler que cette thèse pose au réalisme minimal un problème de cohésion interne. L'approche en effet ? on y reviendra ci-dessous ? accorde une grande importance à l'idée de « véri-aptitude » (*truth-aptness*), la véri-aptitude, dans ce vocabulaire, étant la capacité d'une unité quelconque d'être vraie ou fausse. Mais les propositions, si elles existent, sont toutes vraies ou fausses et la notion de véri-aptitude donc n'a pas grande pertinence dans leur cas. L'intérêt de cette notion dans la démarche d'Engel tient à ce que la question se pose pour certains types de discours ? ceux de l'éthique, par exemple, ou de l'esthétique ou même des mathématiques ? de savoir si les *énoncés* produits dans ces domaines ont ou non des valeurs de vérité. Cette question, bien évidemment, ne concerne pas les propositions, puisque la réponse alors serait triviale. Dans la mesure où la question de la véri-aptitude est philosophiquement intéressante, comme Engel le pense, il faut que les porteurs de la véri-aptitude soient en général des énoncés ou des phrases, c'est-à-dire des unités linguistiques. Et c'est bien ainsi qu'Engel formule lui-même les problèmes de véri-aptitude au chapitre 4 de *Truth*. Un certain antiréalisme en éthique, par exemple, y est présenté comme soutenant que les *énoncés* moraux (*ethical statements*) ne sont pas véri-aptes, alors que le réalisme au contraire affirme qu'ils le sont.[17] Mais cette querelle n'aurait pas de sens si les « énoncés » n'étaient pas, dans ce vocabulaire, quelque chose de linguistique ? ou du moins qui dépende du langage de façon essentielle.

[17] *Ibid.*, pp. 105-112.

Le problème que je veux signaler au sujet de la thèse 3 est donc le suivant. Pour que la véri-aptitude joue le rôle qu'Engel veut lui confier dans le cadre de son réalisme minimal, les porteurs de la véri-aptitude doivent être des unités linguistiques. Or la véri-aptitude étant la capacité d'être vrai ou faux, les porteurs de la vérité ou de la fausseté doivent en principe être les mêmes que ceux de la véri-aptitude. Il en résulte que les porteurs de la vérité, dans le cadre du réalisme minimal, devraient être des unités linguistiques et non des propositions. On peut certes éviter cette conséquence en disant que la véri-aptitude n'est pas en fin de compte la capacité d'*être* vrai ou faux, mais la capacité d'exprimer quelque chose de vrai ou de faux, c'est-à-dire la capacité d'exprimer une proposition. Les porteurs de la véri-aptitude pourraient différer dès lors des porteurs de vérité. Dans ce cas, cependant, la façon même d'introduire la notion de véri-aptitude devrait être revue de même que son rapport précis avec la notion de vérité. La chose est faisable sans doute, mais on se simplifierait la vie, de ce point de vue du moins, en admettant d'emblée que les valeurs de vérité sont attribuées à des unités d'ordre linguistique. C'est en tout cas ce que je ferai dans la suite de ce texte, mais à vrai dire cela n'aura pas grand impact quant au problème principal que je veux soulever.

Thèse 4 : La notion de vérité est univoque d'un domaine à l'autre.

Engel s'oppose ici au pluralisme de Crispin Wright. On n'a pas, selon lui, une notion de vérité en sciences, une autre en éthique et une autre encore en esthétique etc., mais une seule et même notion partout, une thèse que je concéderai bien volontiers.

Ces thèses 1 à 4 constituent la composante minimaliste du réalisme minimal. L'important parmi elles pour ce qui nous concerne, ce sont les deux premières : minimalisme, oui (thèse 1), mais non pas déflationnisme radical (thèse 2). La composante réaliste ensuite s'exprime dans les thèses 5 à 7.

Thèse 5 : Malgré l'univocité de la notion de vérité, la question du réalisme et de l'antiréalisme ne se pose pas de la même façon dans tous les domaines.

L'idée ici est que l'on peut être réaliste dans un domaine sans l'être dans un autre, ce qu'Engel exprime - un peu énigmatiquement, me semble-t-il - en disant que le minimalisme eu égard à la vérité n'implique pas le minimalisme eu égard à la véri-aptitude. Le réalisme minimal se révèle alors ne pas être minimaliste eu égard à la véri-aptitude. C'est là quelque chose d'étonnant à ce stade-ci et d'assez problématique. J'y reviendrai.

Thèse 6 : La véri-aptitude dans chaque domaine doit être évaluée à l'aune d'un critère réaliste.

La question qu'il faut se poser est en gros la suivante : jusqu'à quel point le domaine en question est-il indépendant des jugements que nous portons ou pourrions porter à son sujet ? On doit alors recourir pour évaluer la véri-aptitude des énoncés d'un domaine en particulier à des considérations complexes et beaucoup moins truistiques que celles de la thèse 1.

Thèse 7 : Dans chaque domaine, la vérité réaliste, au sens de la thèse 6, est la norme de nos enquêtes.

Ainsi se retrouve-t-on en bout de piste, petit coup de théâtre, avec une *vérité regonflée*, qu'Engel appelle ici la « vérité réaliste au sens de la thèse 6 », c'est-à-dire au sens d'une indépendance totale par rapport à nos jugements et même à nos capacités de jugement.

Je voudrais soulever à propos de cette approche un problème de fond : pourquoi ne s'oriente-t-elle pas sans réserve vers une conception substantielle de la vérité ?

3. Le réalisme minimal et la vérité substantielle

Arrêtons-nous d'abord, pour amorcer cette réflexion, sur les rapports entre vérité et véri-aptitude. Si la véri-aptitude est la capacité d'être vrai ou faux, tout ce qui est vrai est véri-apte. Ou du moins, pour faire une concession à l'idiome des propositions : un énoncé ne peut pas exprimer une proposition vraie s'il n'est pas véri-apte. Cela paraît tautologique. Mais s'il en est ainsi, le théoricien ne peut pas en toute cohérence se montrer plus restrictif pour la véri-aptitude que pour la vérité. *Prima facie* donc, il paraît difficile de s'en tenir avec Engel à des conditions minimales pour ce qui est de la notion de vérité, mais d'imposer à celle de véri-aptitude des critères réalistes plus contraignants et je trouve difficile à comprendre à ce stade-ci la thèse 5 ci-dessus, selon laquelle le minimalisme eu égard à la vérité n'implique pas le minimalisme eu égard à la véri-aptitude.

Deux interprétations semblent possibles. La première est que l'on se retrouve en fait avec deux notions différentes de vérité. Cela est suggéré par la formulation même de la thèse 7 quand elle évoque « la vérité réaliste au sens de la thèse 6 », comme si l'on avait une notion minimaliste de vérité pour les thèses 1 à 4 et une autre, plus exigeante, pour les thèses 6 et 7. La première,

plus relaxe, satisferait aux truismes énumérés plus haut *et à rien d'autre*, alors que la seconde, plus robuste, satisferait *en plus* à la contrainte d'indépendance eu égard à nos capacités épistémiques. Cette lecture dualiste, cependant, ne saurait convenir à Engel. D'abord, elle compromettrait l'unité même du réalisme minimal. Au lieu d'une seule, on aurait là deux théories, une approche minimaliste pour la notion relaxe et une théorie réaliste pour l'autre.

Il en découlerait en outre que d'un point de vue philosophique, la notion relaxe n'aurait que bien peu d'intérêt, puisque les questions philosophiques les plus pertinentes, au dire d'Engel lui-même, concernent les débats régionaux entre réalisme et antiréalisme, par exemple en éthique, en mathématiques ou même en physique.[18] Un antiréaliste qui soutiendrait, disons, que les énoncés éthiques ne sont ni vrais ni faux « au sens réaliste » serait assez mal venu d'asserter quand même certains de ces énoncés. Il est incohérent de dire qu'il n'est ni vrai ni faux que le mensonge soit moralement condamnable, et d'affirmer néanmoins : « le mensonge est moralement condamnable ». Un tel antiréaliste ne peut pas de façon cohérente asserter sérieusement des énoncés moraux et n'aura que faire donc, dans le cadre de l'éthique, d'une notion de vérité qui satisferait aux truismes de la thèse 1. Il ne pourrait pas de façon cohérente soutenir que les énoncés moraux ne sont ni vrais ni faux au sens réaliste (c'est-à-dire qu'ils ne sont pas véri-aptes), mais que pour autant certains sont vrais au sens relaxe, parce que les poser comme vrais au sens relaxe le commettrait à les asserter, ce qu'il ne peut pas faire sans incohérence, si du moins il parle sérieusement. Peut-être cet antiréaliste prétendra-t-il, pour trouver une échappatoire, qu'en disant « le mensonge est moralement condamnable », il n'entend pas faire une assertion, mais une prescription sous un mode déguisé. Mais dans ce cas, soit il refusera de dire « il est vrai que le mensonge est moralement condamnable », même au sens relaxe du terme « vrai », et du coup puisqu'aucune assertion n'est faite, le truisme (a) ci-dessus (asserter un énoncé, c'est le présenter comme vrai) ne s'appliquera tout simplement pas. Soit il acceptera quand même l'énoncé « il est vrai que le mensonge est moralement condamnable », mais en précisant que de même que la pseudo-assertion « le mensonge est moralement condamnable » relève d'une façon de parler non sérieuse, et même trompeuse, de la même manière l'énoncé « il est vrai que le mensonge est moralement condamnable » constitue une façon non sérieuse de s'exprimer. Mais alors, la notion relaxe de vérité deviendrait non seulement mince, mais carrément frivole sur le plan philosophique. Dans un cas comme dans l'autre, seule la notion réaliste serait philosophiquement

[18] Tout le chapitre 4 de *Truth* (. 99-124) est consacré à ce genre de discussions.

intéressante et des deux théories engéliennes seule l'approche réaliste serait pertinente.

Une autre raison enfin pour laquelle l'interprétation dualiste est inacceptable pour Engel est qu'elle est incompatible avec la thèse 3, selon laquelle les porteurs de vérité sont des propositions. Pour ceux en effet qui admettent les propositions, un énoncé qui n'est pas véri-apte n'exprime aucune proposition. Si donc les propositions sont les porteurs de la vérité minimaliste (la thèse 3 appartient à la composante minimaliste de l'approche engélienne), alors on ne peut avoir quoi que ce soit de vrai au sens minimaliste si l'énoncé correspondant n'est pas véri-apte au sens réaliste et il n'y a pas lieu au bout du compte de distinguer là deux notions de vérité. Récusons par conséquent ? avec Engel lui-même, j'en suis sûr ? l'interprétation dualiste.

L'autre interprétation que je puisse voir requiert d'accepter que le vrai soit un sous-ensemble du véri-apte. Il n'y a rien de vrai, donc, qui ne soit véri-apte (ou qui ne corresponde à un énoncé véri-apte). La contrainte réaliste d'indépendance vaudrait alors tout autant pour la notion (unique) de vérité. Mais quelle est dans ce cas la pertinence de distinguer les deux composantes de la théorie ? Ce serait, pour autant que je puisse voir, de distinguer entre une caractérisation de la *notion* de vérité et une caractérisation de la *propriété* correspondante. La chose du reste est nettement suggérée par Engel :

> Il y a en fait deux questions : l'une est de savoir si la *vérité* est minimale, l'autre est de savoir si la *véri-aptitude* est également minimale. Il n'est pas clair qu'il s'agisse là de la même question, car la première porte sur notre *concept* de vérité et l'autre sur la *propriété* de vérité. Une réponse affirmative à la deuxième n'est pas donnée par une réponse affirmative à la première.[19]

Les thèses 1 à 4 dans cette optique caractérisent *entièrement* la *notion* de vérité : le prédicat de vérité est celui qui satisfait aux quatre conditions en question. Mais pour qu'une unité quelconque tombe sous ce prédicat (ou exprime quelque chose qui tombe sous ce prédicat), elle doit satisfaire certaines conditions supplémentaires qui requièrent une explicitation d'un autre ordre. Il faudra, par exemple, que des entités correspondantes existent réellement, indépendamment de nos capacités épistémiques.

Cette démarcation entre ce qui caractérise la notion de vérité et ce qui caractérise la propriété de vérité revient à la vieille distinction que traçaient les médiévaux entre une définition *quid nominis* et une définition *quid rei*. La

[19] P. Engel, *Truth*, *op. cit.*, p. 84 (avec les italiques de l'auteur).

première sert à circonscrire l'usage d'un terme donné et ses rapports analytiques avec d'autres termes. Ainsi considérée, la caractérisation minimaliste de la notion de vérité fournie par les thèses 1 à 4 est strictement analytique. Une définition *quid rei*, en revanche, est une caractérisation plus substantielle de la chose même. Ce serait en l'occurrence une caractérisation substantielle de ce que c'est pour un énoncé quelconque que d'être vrai ou faux. Mais s'il s'agit bien de cela, alors la composante réaliste de la théorie engélienne ? les thèses 5 à 7 ? commet son auteur à la recherche d'une définition *quid rei* de la vérité, c'est-à-dire à la recherche d'une théorie *substantielle* de la vérité. Tant que le philosophe n'a pas précisé, notamment, quelles sont les relations qu'un énoncé doit entretenir avec les choses réelles pour être dit vrai ou faux, il n'a pas encore caractérisé de façon éclairante la *propriété* même de vérité. Notre conclusion à ce stade-ci doit être qu'une théorie réaliste de la véri-aptitude appelle une théorie substantielle de la vérité.

Engel, cependant, nie cela, sur la base de deux arguments qu'il nous faut maintenant examiner : « Si la véri-aptitude est une propriété de la vérité, et si la véri-aptitude est robuste », écrit-il, « ne devons-nous pas conclure que la vérité est elle-même robuste et revenir à une conception substantielle ? Non, pour deux raisons ».[20] Le premier de ses arguments est que quoi qu'il en soit des exigences réalistes, nous sommes en mesure de spécifier « les propriétés de base du prédicat de vérité pour chacun des domaines » (éthique, mathématiques, physique etc.) en évoquant seulement les conditions minimalistes ? les truismes notamment de la thèse 1.[21] À quoi je répondrais que oui, nous sommes en mesure de le faire. C'est là la caractérisation *quid nominis*, c'est-à-dire analytique, du prédicat de vérité. Ce prédicat est tel, par exemple, que quiconque affirme qu'il est vrai que p est tenu à concéder que p. Mais que cette caractérisation *quid nominis* soit possible en termes minimalistes n'empêche pas qu'une caractérisation plus substantielle ? *quid rei* ? soit requise si l'on est réaliste dans au moins un domaine. Engel lui-même semble aller en ce sens lorsqu'il explique que le réaliste scientifique peut bien admettre, avec le nominalisme, que « il existe des électrons » est vrai si et seulement si il existe des électrons, mais qu'il nous doit alors quelque explication supplémentaire par delà cette simple équivalence.[22] Je ne prétends pas autre chose : le réaliste ? en physique par exemple ? ne peut s'en tenir au minimalisme, il doit fournir une théorie substantielle.

[20] *Ibid.*, p. 120.
[21] *Ibid.*, pp. 120-121.
[22] *Ibid.*, p. 84.

Engel répliquera peut-être que ce dont le réaliste a besoin à ce stade-ci n'est pas une théorie substantielle de la vérité, mais une théorie des électrons. Telle réponse, cependant, ne suffit pas. Une théorie réaliste de la véri-aptitude des énoncés portant apparemment sur les électrons devrait spécifier d'une façon ou d'une autre les rapports que ces énoncés doivent entretenir avec les électrons pour être vrais, par exemple qu'ils doivent contenir un terme référant à des électrons ou un prédicat qui a des électrons dans son extension. Il faudrait, en d'autres mots, caractériser de façon plus que minimale les relations que doivent entretenir pour être vrais les énoncés véri-aptes avec les choses du monde. Or cette exigence n'est satisfaite ni par la théorie physique des électrons ni par une caractérisation minimaliste du prédicat de vérité ni même par la conjonction de l'une et de l'autre. Le premier argument d'Engel pour refuser de s'engager dans la voie d'une théorie substantielle ne paraît donc pas convaincant.

La deuxième raison est la suivante :

> [...] nous n'avons pas besoin d'accepter que notre critère de véri-aptitude réponde au critère d'une théorie substantielle de la vérité au sens classique que nous avons examiné au chapitre 1.[23]

Il n'est pas nécessaire, en particulier, de ramener la vérité au sens réaliste à une question de correspondance avec la réalité. Le lecteur est alors renvoyé à la critique des diverses théories substantielles qu'Engel a proposée plus haut dans l'ouvrage. Mais il y a là quelque ambiguïté. L'argument en effet peut être concédé si on le prend de manière purement extensionnelle : la théorie réaliste dont nous avons besoin n'est pas l'une de celles qui ont été discutées au chapitre 1 de *Truth*. Mais cela n'empêche pas pour autant ladite théorie de devoir être « substantielle » en ceci qu'elle doit nous dire quant aux conditions sous lesquelles un énoncé peut être déclaré vrai ou faux quelque chose de plus que les truismes minimaux, quelque chose qui ait trait à la constitution de la réalité elle-même ? c'est-à-dire à l'ontologie ? *et* aux relations que les énoncés doivent entretenir avec les choses du monde pour être vrais ou faux. Il lui faudra, en d'autres mots, une sémantique substantielle ; je veux dire par là : *une sémantique couplée à une ontologie*. Que les diverses théories discutées au chapitre 1 de *Truth* échouent à réaliser ce projet n'annule pas l'obligation philosophique à laquelle conduit le réalisme d'Engel, celle de rechercher une telle « onto-sémantique ».

[23] *Ibid.*, p. 121.

Engel, certes, semble penser qu'une théorie correspondantiste de la vérité est tout bonnement impossible et il propose en ce sens plusieurs arguments, que je ne saurais discuter ici de façon précise.[24] Mais pour l'essentiel, ces arguments ne visent qu'une variété particulière de correspondantisme, celui qui est basé sur la notion de *fait*. La considération décisive à cet égard revient à ceci : il n'y a aucune façon de spécifier ce qu'est un fait, ni a fortiori ce qu'est la correspondance avec un fait, indépendamment de la notion de vérité ; aucune théorie de ce genre, donc, ne peut fournir une explication satisfaisante et non-circulaire de la vérité ou de la véri-aptitude. Admettons-le, aux fins du moins de la présente discussion. Mais il n'en découle pas qu'aucune théorie substantielle de la vérité ne soit possible, ni même qu'aucune théorie correspondantiste ne soit possible.

Le problème de fond des théories de la vérité qui reposent sur la correspondance avec les faits est qu'elles prennent les *phrases* pour unités sémantiques minimales. C'est l'approche d'un Davidson, par exemple, ou d'un Quine, ou plus récemment d'un Robert Brandom comme de multiples autres. Engel lui-même du reste paraît souscrire à cette approche dans son livre sur Davidson.[25] Mais c'est là une erreur, à mon avis, et je ne trouve guère étonnant qu'elle mèneà l'impasse en théorie de la vérité. Dans tous les langages que nous connaissons, les phrases sont décomposables en unités de signification plus petites et il y a lieu de présumer, par conséquent, qu'une théorie substantielle de la vérité doit ramener celle-ci en dernière analyse à un jeu de rapports entre certaines composantes subpropositionnelles des énoncés et les choses du monde. À monnayer ainsi la vieille idée de correspondance, on évite les objections qu'Engel adressait aux théories correspondantistes (les apories de la notion de fait, la régression de Frege, le lance-pierre de Davidson, etc.), lesquelles n'atteignent, à ce qu'il me semble, que les approches de la vérité qui endossent ce qu'Engel appelle ailleurs le « holisme de la phrase ».[26]

Pour ne considérer qu'un cas très simple à titre d'exemple, on dira dans l'optique que je propose, qu'un énoncé atomique de forme « A est F » est vrai si et seulement si l'individu dénoté par le nom propre sujet « A » appartient à l'extension du nom commun prédicat « F ». Cela suppose en premier lieu une analyse syntaxique de l'énoncé, capable d'en repérer les composantes subpropositionnelles, d'assigner à chacune une catégorie grammaticale (comme celle de nom propre ou de nom commun) ainsi qu'une fonction dans

[24] *Ibid.*, pp. 14-26.
[25] Voir *Davidson et la philosophie du langage*, *op. cit.*, pp. 8-12.
[26] *Ibid.*

la phrase (comme celle de sujet ou de prédicat). Il y faut aussi, deuxièmement, un domaine d'objets, qui puissent être les corrélats des expressions en question, c'est-à-dire une ontologie. Et l'on doit pouvoir puiser, troisièmement, dans un répertoire des rapports sémantiques possibles entre les expressions linguistiques et les objets du monde (l'exemple considéré n'évoque à ce titre que la dénotation pour les noms propres et l'extension pour les noms communs, mais d'autres rapports, de toute évidence, devront être ajoutés selon les besoins conjugués de l'analyse linguistique et de l'ontologie).[27] La théorie de la vérité, à partir de là, peut ramener celle-ci, pour une phrase p donnée, à un jeu de rapports entre les composantes de p et les objets du monde.[28] Ce genre d'approche est controversée, à n'en pas douter. Elle implique notamment une théorie atomiste de la signification, que plusieurs philosophes aujourd'hui pensent devoir récuser.[29] Mais elle fournit bel et bien l'esquisse d'une théorie réaliste et substantielle de la vérité et pour autant que je puisse voir, rien de ce que dit Engel dans *Truth* ne la met en péril.

Ma conclusion, bref, est la suivante. D'une part, le réalisme de Pascal Engel à l'endroit de la véri-aptitude le commet à rechercher une théorie substantielle de la vérité. Sa critique du correspondantisme associé à l'idée de fait, d'autre part, le commet à récuser toute théorie substantielle qui accorde aux phrases le statut d'unités minimales de signification. Ma suggestion, dans ces conditions, est qu'une approche atomiste du genre esquissé ci-dessus pourrait bien lui permettre d'approfondir et de consolider ce réalisme minimal dont il s'est fait le défenseur.

---

[27] Je pense, notamment, à ce que divers auteurs, de Guillaume d'Ockham à John Stuart Mill, appellent « la *connotation* », qu'ils caractérisent de diverses façons selon l'ontologie qu'ils adoptent. Je me suis particulièrement intéressé pour ma part à l'approche ockhamiste, notamment dans *Les mots, les concepts et les choses. La sémantique de Guillaume d'Occam et le nominalisme d'aujourd'hui*, Paris/Montréal, Vrin/Bellarmin, 1992, pp. 240-247, et dans *Ockham on Concepts*, Aldershot, Ashgate, 2004, pp. 63-83.

[28] La théorie des conditions de vérité de Guillaume d'Ockham, par exemple, telle qu'il la développe dans la deuxième partie de sa *Somme de logique* (trad. fr. par J. Biard, Mauvezin, T.E.R., 1996), constitue précisément une approche de ce genre. J'en ai proposé une reconstruction théorique dans *Les mots, les concepts et les choses*, *op. cit.*, pp. 23-67.

[29] La conception atomiste de la signification, cependant, a vigoureusement été défendue depuis plusieurs décennies par Jerry Fodor, et sur la base d'arguments très forts. Voir notamment J. Fodor et E. Lepore, *Holism. A Shopper's Guide*, Oxford, Blackwell, 1992.

4

Valueless Truth *

PAOLO LEONARDI

By means of the predicate 'is true' we monitor our use of language, thereby claiming truth for, or denying it to, what we say or are said.[1] For instance, we monitor what we say by a tag question like *Isn't it true?* or what we are said by replying *That's not true!* The unit of measure of these evaluations is cases in which we assume that what is said tells (or does not tell) how things are. Truth matches a linguistic representation with a state of affairs.

What we assume to be true, I shall argue, are cases of name placement, i.e. cases in which an object or a kind of object is given a name.[2] What I have in mind are not baptisms, or not only. Introductions, giving an example, and occasionally many other uses of a name can do. Names are tools to investigate the nature of things and by themselves names do not carry any, though a practice in using them carries with it information.[3] If a proper name is attributed to an object, the name ideally distinguishes it from anything else; if a

*Pascal Engel has cooperated very much to the Summer School in Analytic Philosophy I organized for some six years, and one of which was held in Paris. But I remember Pascal since the first ESAP meeting in 1992 in Aix-en-Provence, and remember his kindness then in immediately offering himself an organizational matter which was upsetting a session and a speaker. Pascal is a kind and a curious, jokeful, cultivated, all virtues that come up in discussing with him and in reading his writings.
I have discussed ancestors of this paper in Bologna and in Palermo. I thank you for their remarks Patrizia Violi, Claudio Paolucci, Franco Lo Piparo, Francesca Piazza, Marco Carapezza, Francesco La Mantia, and Pietro Perconti. Some of the ideas here presented I have discussed also in Leonardi 2013a and 2013b.

[1] But 'p is true' and p do not assert the same, see Bolzano 1837: I, 147 and 1849 §13.

[2] Or, cases of name displacement, in which a name is negated to an object or a kind of objects.

[3] Names are like baby's bites – at the core, they trace an interest and an appropriation.

predicative name is applied to an object, it potentially groups it together with other objects and distinguishes the group from other groups.[4] It is uses of a name for the same object or kind of object that develop and transmit concepts and conceptions of the object and the kind.

Minimal and modest views of truth are concerned with the predicate 'is true' and the like, and act as if in asserting *s is true if and only if p* or its modest version there were not already an issue with the truth of *p* – truth comes in before the predicate 'is true'. If minimalism and modesty are pursued to avoid correspondentism, the fact is that a sentence, or a discourse, are no mirror of a state of affairs – if there are atomic sentences there are not atomic state of affairs. Moreover, linguistic expressions are made up of parts to which no thing corresponds in the state of affairs it speaks of (the state of affairs it is used to speak of), the most well known issue being syncategorematic expressions. And that in any state of affairs there are many elements to which no thing corresponds in what speaks of it. 'Mark and Ann were playing chess in the dining room, when I came in', say I. In the dining room there were many other things too, the dining room was located in some house or flat, Ann was drinking a beer besides playing chess, and I have come in with two friends of mine. Etc. This is only a sketch of one of indefinitely many different states of affairs, in which what I say could be deemed true. Rather, by a true sentence we point out some features in a state of affairs.

1.

Minimalism, which Pascal half endorses, would have truth as «a merely "formal" or "logical"» property (Engel 2002: 50) plus some platitudes. The formal or logical properties are fully expressed by the equivalence *the proposition that p is true if and only if p*. The platitudes consist in understanding the schema as saying that a proposition *p* is true (i) if and only if it corresponds to the facts, (ii) if and only if things are the way it says they are, or (iii) because *p*. (Engel 2002: 51)

With some good reasons, because of its problems, minimalism and modesty skip giving an analysis of «the internal structure of the truth-value bearers», which Tarski tackles with his recursive strategy (Künne 2003: 317) and in doing which words and objects get connected.[5] The core of their theory

[4] See Leonardi 2011.

[5] Field 1972 claims that Tarski accounts for the semantic predicate 'is true' by means of the semantic predicates of denotation and satisfaction.

of truth consists of the T-sentences, which Tarski derives as consequences in his own theory. Keeping to the linguistic side, minimalism reduces truth to a formal property and little else, modesty to little more – their accounts leaves truth dangling.

Would, for instance, the biconditional '*The water is sparkling' is true if and only if the water is sparkling* account for the truth conditions of the sentence '*The water is sparkling*'? Surely, if the water is sparkling, 'the water is sparkling' is true, and, if it is not sparkling, 'the water is sparkling' is not true. But what are the conditions for accepting the right element of the biconditional? 'The water is sparkling' is acceptable if and only if the water is sparkling... (How do acceptance conditions differ from truth conditions?) The situation is not very different if we move from a minimalist conception to a modest one, i.e. to one according to which $\forall$x(x is true $\leftrightarrow$ $\exists$p(x=[p]&p)). (Künne 2003: 337, but see the whole account 333-74.)

One could conjecture that the grounds for claiming that the water is sparkling do not call for truth. Writes Horwich:

> In mapping out the relations of explanatory dependence between phenomena, we naturally and properly grant ultimate explanatory priority to such things as basic laws and the initial conditions of the universe. From these facts we deduce, and thereby explain, why for example
>
> Snow is white
>
> And only then, given the minimal theory, do we deduce, and thereby explain why
>
> "Snow is white" is true (Horwich 1990: 111)[6]

We give priority to basic laws and initial conditions of the universe, which in our explanations figure by means of sentences. However, these sentences do if and only if they are *true*.[7] We are rather careful at that, monitoring their case and revising our conjectures anytime we find wanting the basic laws and

[6] Pascal quotes the passage, see 2002: 51.

[7] Horwich 1998 accounts for meaning by introducing acceptance properties, «a small set of properties which [...] explain total linguistic behaviour with respect to that word.» Then, he offers as instances the acceptance properties of 'and', 'red' and 'true'. We «accept '*p* and *q*' if and only if we accept '*p*' and '*q*'»; we accept «to apply 'red' to an observed surface when and

initial conditions of the universe that we have posited, i.e. anytime we suspect them to be false or not precise. Indeed, the relevance of the truth predicate can be inferred from the fact that any biconditional along the equivalence schema above is true if and only if its left element and its *right* element are both true or both false. It does not matter that in the right element does not occur the predicate 'is true'.[8]

2.

We ground truth assuming to be true some sentences in some circumstances. In his definition of truth, Tarski assumes the extension of any predicate to be defined, and hence the truth or falsity of any atomic formula to be established. This is not actually the case. Language is a cognitive tool, and as a matter of fact predicative names are applied to a limited number of things, and their application is always revisable.[9] In any event, we accept *some* contingent truths, which are relevant as proper and predicative names placement relative to *some* circumstances.

The truth of other sentences, as the occasion comes up, is decided by assimilating them and the occasion to, or distinguishing them from, the sentences and the circumstances previously described by the proper and predicative names. Mark is a child, is George a child too? The Earth is a planet, is *Mu Arae e* a planet too? If the cases cannot be assimilated to any previously assumed one, we introduce new sentences – George is an old child, or George is a boy, George is a young man, etc – and assume they properly describe their circumstance. Or, alternatively, we refute assimilating the present case to the previous one – George is not a child. These are mixed waters, where epistemology and semantics mesh together, and they do not concern me here.

Let us call the uses of language I am examining *coordinative uses*. In any such use, language and reality touch each other. The set of cases has neither to

---

only when it is clearly red»; and we accept 'true' when we «accept instances of the schema "the proposition *that p* is true if and only if *p*".» Then the question becomes when do we accept '*p*', '*q*', '*red*', and again '*p*'.

None of these is a basic law or part of the initial conditions of the universe.

[8] Sher and Wright 2007 remark that deflationist views of truth reduce truth to the predicate 'is true'. This choice has, they claim, two drawbacks. It forgets other ways truth surfaces in natural language sentences – for instance, by means of adverbs as 'truly' – and what they call the illocutionary role truth plays in defining assertoric uses of sentences. On the second point, they refer to Frege 1918. On the stroke symbol and assertion in Frege, see also Picardi 1989.

[9] If I were careful, I would have claimed that the application of a predicate is almost always revisable. If something, however, is not revisable maybe we cannot claim that it is not.

be stable nor the same for all of us. The sentences have neither to be elementary as Tractarian propositions were, nor to be fully explicit – 'Boy', 'The boy', 'That's a boy', 'Ann's boy', 'The boy is Ann's', etc, all can do. With such units of measure, we distinguish boys from children and adults (and judge the case in which someone is claimed to be a boy).

Only rarely we decide our coordinative uses. Occasionally, we revise single assumptions, but we do not decide anytime the whole asset of cases. We happen to revise our assumptions without deciding – because we are absent-minded and do not even realize we have changed sentences, instances or views, or a change of views may impose on us. Any change is consequential.[10]

Now, I would push my point linking it with some stands that I feel close to it.

What I have in mind articulates a thing that, in "A Defence of Common Sense" in 1925, Moore *en passant* says, namely that he knows the meaning of the truisms, but not how to analyze that meaning. Moore claims to be using the words with their ordinary meaning. Some truisms – for instance, 'I am a human being', or 'Here is a hand' – place common nouns – respectively, *human being* and *hand*.[11] 'Here is a hand' is not a sentence (a proposition) with an empirical look and a grammatical role, as Wittgenstein would have argued, but a use of the noun 'hand' to which Moore attributes a paradigmatic value, and which he suggests his audience to attribute the same value. The use plays the role of a standard. Any use in which a word and what it is about come together can play that role, and the better the more *perspicuous* it is.[12]

There are three relevant aspects in Moore's case. (i) He commits himself to the existence of what is named, whose nature has yet to be investigated. (ii) The existence of two things is acknowledged at once, the noun 'hand' and

[10]Coordinative uses do not relate to truth-aptness. Truth-aptness is an illocutionary issue, so to speak, whereas coordinative uses are a semantic one. Perhaps, any field of discourse is true-apt, and all judgments but perceptual ones are. That is, sentences about any field of discourse are possibly true or false. Coordination is not about what can be linguistically represented, but about how a linguistic representation acquires content, and the idea is that anchoring a linguistic representation to a state of affairs is what generates its content.

This is how Kant seems to have argued (see Vanzo 2012). Burge 2010 claims that perceptual judgments too are truth-apt – that objectivity begins with perception is a central claim of the book.

[11] 'Here is a hand' is the first premise in Moore's 1939 proof of the existence of an external world.

[12]The placement of a common noun is the placement of a predicative name. And there is also proper name and relational name placing.

I would call 'perspicuous' a use the more easily it is understood by the higher number of people to whom it is offered.

the hand itself. (iii) The previous history of the two things is relevant but inessential. There could be previous concepts and conceptions of the relevant thing that are picked up, or retrieved, together with the suggested standard use of the term – but they may change – or concepts and conceptions of it may develop after the standard, and be transferred by the term which the use anchors.[13]

Schlick, in 1918-1925, advocates a less informal but similar picture, to which Reichenbach later subscribes. Dealing with the introduction of units of measure, they assert that such units are introduced by coordinative definitions, that is by definitions that coordinate physical objects and concepts (I would say 'terms' rather than 'concepts'). Writes Reichenbach:

> In principle, a unit of length can be defined in terms of an observation that does not include any metrica! relations, such as "that wave-length which occurs when light has a certain redness." In this case a sample of this red color would have to be kept in Paris in place of the standard meter. The characteristic feature of this method is the coordination of a concept to a physical object. These considerations explain the term "coordinative definition." If the definition is used for measurements, as in the case of the unit of length, it is a *metrical* coordinative definition. (1928 [1957]: 15)

A coordinative definition transforms a particular length, weight, volume into a standard respectively for length, weight, volume, linking the level of objects with that of language and thought (with words and concepts). The definition supplies no information, but constitutes a tool to collect information. As it is well known, we reflect on our standard and keep looking for better ones. Lateral information and indefinitely many adjustments (how to apply the standard, how to keep properly the physical standard, like the meter bar in Paris, in what circumstances its use can be trusted, etc) point at how to revise the standard itself.[14]

[13]Quine's denial of a distinction between linguistic and factual elements goes with my Moorean understanding. See Quine 1953. However, Quine 1960, and later, turns the problem towards his indeterminacy thesis.

The idea that naming helps recognizing and developing concepts of things is a topic investigated by Markman 1989 and Bloom 2000.

[14] Speaking of the standard meter, Kripke 1972-1980 investigates how the standard is fixed and kept in the Sévres Museum as a case of an a priori contingent truth. That the standard bar is one meter long is one such truth, and it fixes the reference of 'one meter'. Wittgenstein 1953 (§50) too discusses the standard meter case, asserting that the standard meter cannot be said to be

Predicative names introduced coordinating them with some instances are thereby defined and true of the instances.[15] At the same time, as with coordinative definition in physics, the coordination by itself does not endow any articulated content, which comes later investigating what there is thanks to the coordination.

Thirdly, I would compare Wittgenstein's discussion on Moore's truisms in *On Certainty* with my claim. The sentences we assume true are, in my view, partially alike and partially different from Wittgenstein's hinge propositions.

> 519. Admittedly, if you are obeying the order "Bring me a book", you may have to check whether the thing you see over there really is a book, but then you do at least know what people mean by "book"; and if you don't you can look it up, – but then you must know what some other word means. And the fact that a word means such-and-such, is used in such-and-such a way, is in turn an empirical fact, like the fact that what you see over there is a book.
>
> Therefore, in order for you to be able to carry out an order there must be some empirical fact about which you are not in doubt. Doubt itself rests only on what is beyond doubt.
>
> But since a language-game is something that consists in the recurrent procedures of the game in time, it seems impossible to say in any *individual* case that such-and-such must be beyond doubt if there is to be a language-game – though it is right enough to say that *as a rule* some empirical judgment or other must be beyond doubt.

---

one meter long because it plays a grammatical and not an empirical role. Wittgenstein touches the issue in many other places, indirectly already in the *Tractatus*, in conversations with members of the Wiener Kreis, in his works on the fundaments of mathematics – distinguishing all along the logical (grammatical) role of the standard and its empirical application. As I argue in the text Wittgenstein 1969 seems to doubt this distinction, though he does not give it up (see, for instance, §§ 309, 319, 321, 519). Wittgenstein writes that «Not only rules, but also examples are needed for establishing a practice.» (1969 §139) In the examples, words and objects meet, and if we kept only to the linguistic formulation of the rule we would have loop-holes in the practice. On Kripke and on Wittgenstein cf Salmon 1988, Diamond 2001, Pollock 2004, Mácha 2012.

[15] Proper names distinguish their bearer from anything and anyone else and do not categorize their bearer. I am inclined to think that 'This is George' and 'That is not Ann' respectively assert and deny the appropriateness of applying to two individuals the distinctive marks 'George' and 'Ann'.

With some hesitance, Wittgenstein calls the sentences that formulate the recurrent procedures of a language game *grammatical propositions*. Any sentence, however, can play the role of a grammatical proposition, offering a paradigm rather than voicing a rule, and being used as a standard. Playing this role does not conflict with its being also an empirical proposition. Any sentence can play the two roles – tell, imagine, inquire, comment on what is the case and offer a standard for future uses. Any example does. If I am right, there is no problem in telling true a grammatical proposition as Moore does, and in claiming to know it, though not in the sense of being able to justify it. The dilemma between grammatical and empirical propositions is one Wittgenstein has faced throughout. It shows up already in the *Tractatus logico-philosophicus*:

> 2.0211 If the world had no substance, then whether a proposition had sense would depend on whether another proposition was true.

World (and its substance) and language come together in assuming true some uses of a sentence.[16]

Wittgenstein's claim comes very close to mine, substituting 'truth', 'true', and 'assumed to be true' in the quote from *On Certainty*, above, as follows,

> Therefore, in order for you to be able to carry out an order there must be some empirical fact which you assume to be true. Truth itself rests only on what is assumed to be true.
>
> But since a language-game is something that consists in the recurrent procedures of the game in time, it seems impossible to say in any *individual* case that such-and-such must be assumed to be true if there is to be a language-game – though it is right enough to say that *as a rule* some empirical judgment or other must be assumed to be true.

My claim, let me repeat, is that some uses of sentences have to be assumed to be true – for instance, that this is a hand, that the Earth exists by more than five minutes, that the White Mountain exists by more than four minutes, that George is a boy, etc.

[16] Wittgenstein was Kantian enough at the beginning to pursue the idea of conditions of experience as something detachable from experience itself.

Assuming something true, fourthly, is not part of an interpretation – and hence it is not what Davidson aims at when he speaks of retrieving what people hold true. Interpretation reconstructs a language going from words to things, whereas what I am pursuing goes in the other direction.

Rather the case, fifthly, can be compared with Donnellan' referential uses of descriptions. In "Reference and Definite Descriptions", in 1966. Donnellan sketches the referential and the attributive use of a definite description. In a *referential* use, a person has a thing in mind and by the description calls others' attention to it. In an *attributive* use, a person attributes properties or relations to, or looks for, etc, a thing satisfying the description, possibly not having it in mind. Here is an example of the same description once in referential and once in attributive use. I am invited to dinner by a couple of friends. On the coffee table there are some architectural photographs, my host tells me the when and the why of most shots, her preferences in this special category of pictures, etc. Understanding that it was her to take the shots, in leaving, I say «The photographer knows her job!» – I use 'the photographer' to refer to her. At the entrance of a female civil engineering trade exhibition there are some architectural photographs. Suggesting you to have a look, I say «The photographer knows her job!» and add «Can you tell whom she is?» – I use 'the photographer' attributively to denote whoever took the pictures.

Donnellan neatly sketches the attributive and the referential use of a definite description:

> To illustrate this, we can imagine the following games: In the first a player gives a set of descriptions and the other players try to find the object in the room that best fits them. [...] In the other game the player picks out some object in the room, tries to give descriptions that characterize it uniquely and the other players attempt to discover what object he described. In the second game the problem set for the other players (the audience in the analogue) is to find out what is being described, not what best fits the descriptions. (1970: 356; see also Donnellan 1968: 214, n 12.)

Using a description to refer is a game of the second kind.

Section IX of "Reference and Definite Descriptions" assimilates a description in referential use to a Russellian (logically) proper name. As a Russellian proper name does not require that what it names satisfies any description, so a description in referential use does not. It does not even require that what it refers to satisfy its descriptive condition. Almost at the conclusion of the second last paragraph of that section, Donnellan asserts that

> [...] this seems to give a sense in which we are concerned with the thing itself and not just the thing under a certain description ... (1966: 303)

In the referential use, the speaker grasps what she refers to independently from the description she offers and claims that it satisfies the descriptive condition. The descriptive condition advocated, whatever its previous usage, offers a standard, and if the use deviates from the previous one, it is the occasion for a language shift.[17] Donnellan's claim can be extended to predicate. In the referential use of a description, it is the descriptive condition which is directly linked with the particular that is thereby claimed to be an instance satisfying that condition. Then, the same phenomenon happens when a predicate is applied to a thing the speaker grasps independently from what she predicates of it.

3.

How does linguistic representation develop and how does it get its content? How can we assess whether it is affordable? (This issue is distinct from that of how language and things are related.) Everything is grounded, I suggest, on assuming some representations to be affordable, a lighter requirement if we require a minimal content to be relevant at that. A coordinative definition attributes no content. The bar offer no content to the standard meter, it offers its length, whatever it is, as a standard of measure. Another bar will be said, if it is as long, to be long one meter, if it is twice as long, to be long two meters, ... If it is a proper name to be coordinated, the definition further distinguishes a thing from the other ones – my brother and me are distinguished by me being named 'Paolo' and him not being so named. If it is a predicative name to be coordinated, the definition further assimilates things in groups and distinguishes among groups of things – my cat is assimilated to your cat by both being said to be *cats*, both cats are distinguished by Ann's pet, who is said to be a *dog*. By acknowledging my cat as a cat, I am driven to acknowledge your

[17] Writes Kripke:

> In particular, I find it plausible that a diachronic account of the evolution of language is likely to suggest that what was originally a mere speaker's reference may, if it becomes habitual in a community, evolve into a semantic reference. And this consideration may be one of the factors needed to clear up some puzzles in the theory of reference. (1977: 271)

pet as a cat too.[18] Thereby, the realist engagement starts before attributing a nature to things.[19]

Truth as I have discussed it is a property of linguistic representations and concern attributing a predicative name to things. By that property we monitor, in everyday contexts as in more sophisticated ones, the adequacy of linguistic representation. It is not exactly a semantic property, but fixing elements to evaluate truth fixes the semantic of the language.

As most people, I have an instinctive inclination to take truth as correspondence. But truth is not correspondence. Any sentence whatsoever matches little of the circumstance it is about. Any sentence has a structure much simpler than the circumstance it is about has, and at the same time many parts of a sentence do not match anything in the circumstance. «Marco has left with Anna» say I. What Marco does is something much more complex than the sentence I utter. Marco has legs, arms, ears, eyes, nose, etc, his going out is along a path – different in space-time from Anna's path. The name 'Marco' has no semantic parts – 'arc' is not a semantic element of 'Marco', but a phonetic string that distinguishes the name 'Marco' from the name 'Mario', in which figures the string 'ari'. If Marco's forehead has an arc shape, the 'arc' in 'Marco' does not represent it. Leaving is a complex activity which starts in a location and ends in another one, involving a sophisticated motor performance – things which have no elements corresponding to them in 'has left'. One can think that my remark on 'arc' is irrelevant. The problem it poses at the level of individual word cannot be hidden when we move to sentences which contain sentences as elements, and specifically those which are logically easier to deal with, i.e. sentences in which occur a logical connective, or in my example a sentence in which a preposition occurs such as 'with'. What does a logical connective, such as 'and' or 'or' (to use their natural language version), correspond to? What does a preposition such as 'with' correspond to?[20] Even

[18] Russell 1903 §48 writes: « [...] things and concepts. The former are the terms. indicated by proper names, the latter those indicated by all other words.». My point is that names keep indicating things, even when we connect with them a richer content entertaining views, and mastering information, about the nature of the things named.

[19] Could the meaning or content of language be differently accessed? Imagine content were innate. It could be that our words have meaning because God endowed us some ideas. God knows what ideas are appropriate to our world – hence, this is only a indirect link between ideas and things, and it doesn't detach ideas from things, giving ideas a priority. Ideas could be innate because of the biological evolution of our species. But biological evolution tells the experience of the species rather than that of the individual, and again it is does by having ideas directly selected by fitness to the case.

[20] Perhaps, it corresponds to an operation to be applied to the linguistic string itself in which

if we had straight up which linguistic pieces have to match which pieces in a state of affairs, any bit of discourse matches only a limited number of the relevant pieces in a state of affairs. There are indefinitely many circumstances that match what has been said.

Sometimes I imagine that the correspondence problem can be solved positing an injective relation between a sentence and the circumstance it tells about. That is, the match is between the relevant elements making up the sentence and only some elements of the circumstance. If that were right, a sentence would constrain the circumstance it is about, by picking out some elements in it. Then, however, it could be related to indefinitely many different circumstances. Which one is that to which the sentence corresponds?[21] Has Marco left alone or did he go with some other people besides Ann? Did they leave by foot or by car? To go out of town or to another place in town? Etc. Any match leaves the relevant circumstance largely indeterminate.

Hence, the relation between words and objects is more sophisticated than how a correspondence view takes it to be. It is a limited match, which starts from matching two complex units as if each where point form. That does not introduce any indeterminacy, because there is no question about which object the words have to be linked with – the object involved in the link was involved in the linking.

4.

A very short remark on the norm of truth, in closing. If truth is a property of some linguistic expressions, and it is about the adequacy of the application of a predicate, one such application either is true or it is not.[22] That is, truth is a factual property. On the relevance of entertaining proper information about what's the case, we come to value pursuing truth. Pascal dedicates a chapter in his book *Truth* to the norm of truth, a norm which he formulates in two ways, at p. 129:

---

the connective occurs. But it corresponds to nothing in the circumstance.

[21] Austin 1950 introduced demonstrative conventions in his analysis of truth, I believe, to overcome this difficulty. He didn't solve the problem, yet, because he said nothing on how these conventions are supposed to work in selecting the relevant circumstance.

[22] This is a standard formulation of realism by Dummett. An antirealist would add that in the undecided case we cannot claim neither truth value. In my case, I would say that the linguistic representation of the case may be undecided, and hence that in make no sense as yet to imagine the case to be anyway true or false. In my case, the fault, if fault there is, is on language, i.e. we have not yet a proper linguistic representation of the case.

(BT) *For any* p, *one ought to believe that* p *only if* p *(is true).*

(BK) *For any* p, *believe that* p *only if, for all you know,* p *(is true).*

One might agree, but these are norms of belief, which assume truth as a value, and which most likely tell what belief is. But here the norm of truth has truth as object and not as subject.

5. References

J.L. Austin 1950 "Truth" (*Proceedings of the Aristotelian Society / Supplementary Vol.* **24**: 111-28).

P. Bloom 2000 *How Children Learn the Meanings of Words* (Cambridge MA Mit Press/Bradford Books).

B. Bolzano 1837 *Wissenschaftenslehre* (Sulbach Seidel; Engl tr. by G.Rolf *Theory of Science* Oxford Blackwell 1972).

B. Bolzano 1849 *Patadoxien der Unendlichkeit* (Leipzig Reclam; Engl. tr. by D.A. Steele *Paradoxes of the Infinite* London Routledge & Kegan Paul 1950).

T. Burge 2010 *Origins of Objectivity* (Oxford Clarendon Press).

C. Diamond 2001 "How long is the standard meter in Paris?" (in *Wittgenstein in America* T. McCarthy, & S.C. Stidd eds Oxford Clarendon Press: 104-39).

K. Donnellan 1966 "Reference and Definite Descriptions" (*The Philosophical Review* **75**: 281-304).

K. Donnellan 1968 "Putting Humpty Dumpty Together Again" (*The Philosophical Review* **77**: 203-15).

K. Donnellan 1970 "Proper Names and Identifying Descriptions" (*Synthese* **21**: 335-58).

P. Engel 2002 *Truth* (Chesham Acumen).

H. Field 1972 "Tarski's Theory of Truth*"* *(The Journal of Philosophy* **64** 347-75).

G. Frege 1918/19 "Die Gedanke" (*Beiträge zur Philosophie des Deutschen Idealismus*: 56-77; en. tr. by A.M. and M. Quinton "The Thought" *Mind* **65**, 1956: 289-311).

P. Horwich 1990 *Truth* (Oxford Blackwell; 2nd ed. revised Oxford, Oxford UP 1998).

P. Horwich 1998 *Meaning* (Oxford Clarendon Press).

S.A. Kripke 1972 -1980 "Naming and necessity" (in *Semantics of Natural Language D.* Davidson & G Harman eds:. 253-355 and 763-9; published as a book with added a "Preface" Oxford Blackwell 1980).

S.A. Kripke 1977 "Speaker's Reference and Semantic Reference" (*Midwest Studies in Philosophy* **2:**. 255-76).

W. Künne 2003 *Conceptions of truth* (Oxford Oxford UP).

P. Leonardi 2011 "Predication" (in P*hilosophical Papers dedicated to Kevin Mulligan* http://www.philosophie.ch/kevin/festschrift/ pp. 14; printed in *Mind, values and metaphysics / Philosophical papers dedicated to Kevin Mulligan* Berlin Springer 2014).

P. Leonardi 2013a "Moore and Wittgenstein" (*Philosophia* **41**: 51–61).

P. Leonardi 2013b "La sostanza del vero" (in *A plea for balance in philosophy / Essays in honour of Paolo Parrini* R. Lanfredini & A. Peruzzi eds Pisa ETS: 137-50).

J. Mácha 2012 "Language meets and measures reality" (in *Doubtful Certainties. Language-Games, Forms of Life, Relativism* J. Padilla Gálvez & M. Gaffal eds Munich Ontos Verlag: 121-8).

E.M. Markman 1989 *Categorization and Naming in Children: Problems of Induction* (London A Bradford Book/ MIT Press).

G.E. Moore 1925 "A defence of common sense" in *Contemporary British philosophy* J.H. Muirhead ed London George Allen & Unwin: 193-223).

G.E. Moore 1939 "Is existence a predicate?" (*Proc. of the Aristotelian Society - Suppl. vol.* **15**: 175-88).

E. Picardi 1989 "Assertion and Assertion Sign" (in *Teorie delle Modalità* Atti del convegno internazionale di Storia della logica di San Gimignano 5-10 dicembre 1987 a cura di G. Corsi C. Mangione M. Mugnai Bologna Clueb: 139-54).

W.J. Pollock 2004 "Wittgenstein on the standard meter" (*Philosophical Investigations* **27**: 148-57).

H. Reichenbach 1928 *Philosophie der Raum-Zeit-Lehre*, (Berlin and Lipsia Walter De Gruyter; Engl. tr. M. Reichenbach & J. Freund New York Dover 1957).

B. Russell 1903 *Principles of Mathematics* (London George Allen & Unwin).

N. Salmon 1988 "How to measure the standard metre" (*Proceedings of the Aristotelian Society*
New Series **88**:193-217).

M. Schlick 1918-25 *Allgemeine Erkenntnislehre* (Berlin Julius Springer 1918, 1925).

G. Sher and C.D. Wright 2007 "Truth as a normative modality of cognitive acts" (in *Truth and Speech Acts / Studies in the Philosophy of Language* G. Siegwart & D. Griemann eds London Routledge: 280-306).

A. Vanzo 2012 "Kant on Truth-Aptness" (*History and Philosophy of Logic*, **33**: 109-126).

L. Wittgenstein 1921 "Logisch-Philosophische Abhandlung" (*Annalen der Naturphilosophie* **14**: 185–262; Eng. tr. by C.K. Ogden with an introduction by B. Russell *Tractatus Logico-Philosophicus* London Routledge & Kegan Paul 1922).

L. Wittgenstein 1953 *Philosophische Untersuchungen* (Oxford Blackwell; Engl. tr. by G.E.M. Anscombe).

L. Wittgenstein 1969 *Über Gewissheit / On Certainty (*Blackwell Oxford; Engl. tr. by D. Paul & G.E.M. Anscombe).

5

'Happiness is overrated: It's better to be right.' On Truth as Emergence

MAURIZIO FERRARIS

Eating mushrooms in a restaurant involves an act of great faith in truth: the person who picked the mushrooms knew (or, in some unfortunate cases, thought he knew) that they were not poisonous, and this knowledge of his corresponded to a property of the world, namely the fact that the mushrooms were not poisonous. It also involves a no less important act of faith in humanity: people who, usually, we have never seen before and will never see again feed us with mushrooms that may be poisonous, but are not. It is hard to see why one should place an antithesis between the solidarity of the cook who is not poisoning us intentionally by adding cyanide to the mushrooms and the objectivity of the mushroom picker who was not mistaken. It is also hard to see why the cook's solidarity should be bigger and more true than the picker's objectivity if, prescinding from that objectivity, the cook gave us poisonous mushrooms, pursuing the humanitarian ideal of sparing us the inevitable pain of existence.

And yet, these are the assumptions of what I propose we call 'post-realism', i.e. the thesis – which dominated the philosophical debate of the second part of the past century – that reality and truth are historical notions, just as feudalism and courtly love, and that we can do without them, not so much for ontological parsimony but rather for an emancipative goal. Post-realism has two versions, the pragmatist and the nihilist. The first has the merit of being explicit: we must get rid of truth and reality, which (if we move from the prosaic mushroom example to the more sophisticated weaves between knowledge and power) are useless if not dangerous. The second is more insincere,

and argues that we should move beyond the realism / anti-realism issue, since it is not philosophically relevant. In a memorable confrontation with Richard Rorty,[1] Pascal Engel faced the pragmatist version.

In dialogue with Rorty in 2002, at a time when realism was still unpopular, Engel had the merit of reinstating the crucial philosophical opposition between realism and anti-realism and of proposing the theory of truth as correspondence. One can certainly demythologize truth and stop thinking that it has magical properties, as it were. But the best way to demythologize it is not to get fully rid of it, but rather to acknowledge where it lies: it is true that the amanita phalloides is poisonous, and this depends on the amanita phalloides, not on us.

In these pages, I would like to return to that debate by proposing an argument in favour of correspondentism, which I call ' truth as emergence.' In a way, truth pops up like a mushroom, emerging from the world towards other parts of the world – us. Which is the exact antithesis of Rorty's thesis according to which, after all, mushrooms are socially constructed too, and the amanita phalloides can become edible if society wishes so. And yet, a poisonous mushroom is such even if the United Nations Assembly decrees that it is not and the truth – if fortune (or misfortune, because the truth is not always welcome) helps us – can pop up like a mushroom, without anyone constructing or seeking it. Before describing the characteristics of truth as emergence I will outline the characteristics of internalism (i.e. the post- realist thesis that truth is completely internal to conceptual schemes) and externalism (i.e. the commonsense thesis that truth is the encounter between conceptual schemes and something external to them).

1. Internalism

As I have just said, post-realism is internalism: the argument that everything lies within conceptual schemes. This means that if a mushroom is poisonous, it is because of the conceptual frameworks that assess it as poisonous. At the origin of internalism there is a broadly political concern: objectivity is seen as an instrument of domination and an obstacle to solidarity, so that truth is regarded as something potentially dangerous or at least useless. With respect to

[1] P. Engel – R. Rorty, *A quoi bon la vérité*, Paris, Grasset 2005. English translation: *What's the Use of Truth*, New York, Columbia University Press, 2007. For the references mentioned in this article, and for a further clarification of my perspective, I refer the reader to my *Documentality. Why It Is Necessary to Leave Traces*, New York, Fordham University Press 2012.

this state of affairs, internalism plays a dual role. On the one hand, it lightens the weight of truth by making it suspect (truth is socially constructed, so there is nothing absolute); on the other hand, it proposes alternative perspectives: if truth is socially constructed and objectivity is a totalitarian myth, it is better to engage (in the pragmatist version) in more fruitful constructions, such as democracy, or (in the nihilist version) in more daring deconstructions, for example by stating that '2 + 2 = 4' is a proposition of the same family as 'woman is by nature inferior to man.'

Analyzing the reasons of internalism, Engel pointed out that the Bush administration was the promoter of a potentially externalist objectivism, but he also noted that the fact that externalism has bad advocates is not enough to disqualify the appeal to objectivity.[2] And we can say more. At the time when Engel was dialoguing with Rorty, the Bush administration seemed to have abandoned its externalism (whether real or apparent) in order to embrace a radical internalism, arguing – *à la* Rorty, after all – that reality is not absolute, but simply the fixation of 'reality-based communities', where the Empire is able to construct its own reality[3] (but then why pursue externalist degrading practices such as phone hackings?). This was a case of Fichtian internalism that, alone, suffices to make any kind of internalism problematic, including non-governmental and leftist ones.

But in general the whole internalist system seems to describe a wish of the heart rather than a philosophical theory. For example, the argument about the superiority of solidarity over objectivity does not consider the obvious counterexamples, such as the fact that the mafia is an extremely supportive organization that, moreover, relies on objective factors, such as the effectiveness of firearms. And when Rorty argues that 'our responsibilities are exclusively toward other human beings, not toward "reality,"'[4] he seems to be placing human beings in the context of unreality, with the paradoxical outcome that we are responsible only towards unreality.

[2] P. Engel – R. Rorty, *What's the Use of Truth*, p. 74 fn.

[3] I quote from "Reality-based Community" in *Wikipedia*: 'The source of the term is a quotation in an October 17, 2004, The New York Times Magazine article by writer Ron Suskind, quoting an unnamed aide to George W. Bush (later attributed to Karl Rove): The aide said that guys like me were "in what we call the reality-based community," which he defined as people who "believe that solutions emerge from your judicious study of discernible reality." ... "That's not the way the world really works anymore," he continued. "We're an empire now, and when we act, we create our own reality. And while you're studying that reality—judiciously, as you will—we'll act again, creating other new realities, which you can study too, and that's how things will sort out. We're history's actors ... and you, all of you, will be left to just study what we do.'

[4] P. Engel – R. Rorty, *What's the Use of Truth*, p. 41.

The fact that internalism expresses a wish of the heart is the key to everything. On closer inspection, the basic problem of the internalist perspective is that it takes the fact / value dichotomy as valid, and then proposes to cancel the facts (objectivity) for the exclusive benefit of the values (solidarity). Thus there is a world of facts, which is regulated by causes and effects, and then a world of values transcending causes, magically surrounded by freedom. This contradicts everything we know of values: their binding character, their being able to go against our interests, their being much more solid and grounded than our philosophies. This is not to say that we can not change values, but we can be sure that if it depended on us and a our freedom we could change them without too much effort, which is not the case.

Values do not fall from the sky: they emerge from the world. Suffice it to think that the first value, the value of all values, is the real that imposes itself and demands our attention. Any value claims to hold for everyone, and nothing better represents this claim than the presence of something we cannot avoid nor amend: reality. For this reason, ethics is not conceivable without an ontology. Imagine a hyper-internalist world of values without facts. What kind of world would it be? And above all, would those values be such? I do not think so. Let us look at the experiment of the ethical brain, which is a variation of the *Gedankenexperiment* of the brain in a vat. The idea is this: imagine that a mad scientist has put some brains in a vat and is feeding them artificially. By means of electrical stimulation, these brains have the impression of living in a real world: some are evil and some are holy. But are they *really* evil or holy? Can we attribute values to a body-less and world-less brain? Would terms like 'happiness' or 'unhappiness' make sense at all if there were no outside world? I think not.

2. Externalism

The *British Medical Journal* has recently published the results of a somehow Rortian experiment.[5] In the attempt at answering the question 'Do you care more about being happy or being right?' a husband was asked to always agree with his wife (even when he though she was wrong). This seemed to drive the wife crazy, so the experiment ended after twelve days. As the *Los Angeles Times* put it when giving an account of the experiment, 'Happiness is overrated: It's better to be right.' Truth has a peculiar importance: it cannot

[5] BMJ 2013;347:f7398.

simply be given by virtue of an intersubjective consensus, and it is the prerequisite of all our practices. Hence the inevitability of externalism, namely the argument that there are things actually independent of, and external to, conceptual schemes. Dinosaurs existed long before us, and the fact that they have never known to be called 'dinosaurs' does not deprive them of any essential property. Our perceptual apparatuses select a certain colour wave as 'white', but 'being white' is still a property of the snow and not of our eyes (which, we should not forget, are a part of the external world). Not to mention that a certain degree of externalism is the basis for the very notion of 'conceptual scheme': in order to really be a scheme (a form), it needs a content that lies outside itself.

Externalism also regards the sphere of words: 'dog' is external to 'cane' no less than the words 'dog' and 'cane' are external to (i.e. are not identical with) the being they refer to. These considerations suggest that the domain of internalism, which for the post-realist is immense, turns out to be rather small. Not only does the external world comprise natural and ideal objects (unless we want to confuse arithmetic with psychology or sociology), but, in many cases, it also includes social objects – an area where often one regards as 'socially constructed' what, at most, can be considered 'socially dependent'. Again, if – in agreement with Engel – we apply King Lear's principle 'I'll teach you differences,' we will realize that externalism exists in the sphere of social objects as well.

For example, the Ecole des Hautes Etudes en Sciences Sociales is unquestionably 'socially constructed', as we have the written documents proving the origin of the institution. Consequently, there are also responsibilities to be assigned. For example, in an institution that was significantly different from the EHESS, i.e. the Third Reich, Goering, by signing the document for the final solution of the Jewish problem, became responsible for genocide. One can also claim without too much difficulty that anti-Semitism is a socially constructed phenomenon. We have historical data that signal the deportation to Egypt or the Babylonian captivity, then the diaspora. Hence – with political, social and psychological motivations that one might be able to reconstruct – the genesis of anti-Semitism as a reaction to a sense of guilt, as a search for scapegoats, as the pursuit of economic gain, as religious fanaticism, and so forth.

I would have much more difficulty in saying that monotheism is socially constructed. Because not only there is no name or signature (as in the case of the final solution), but there are no generic historical testimonies either, unlike the case of anti-Semitism. One can make conjectures, but they would all be equivalent because we might never have any kind of historical evidence

on the social genesis of monotheism or polytheism. Therefore it is amusing and instructive to see that Hume explains how we went from polytheism to monotheism, while Schelling explains how we went from monotheism to polytheism.

Despite appearances, these difficulties are not empirical but transcendental. I have no difficulty in accepting that the monotheism of Akhenaten was socially constructed, given the historical evidence about a pharaoh's decision to impose (without success, the impact factor of Moses was much higher) a monotheistic worship of the sun. On the contrary, I have great difficulty in accepting the idea that monotheism, polytheism, or religion in general are socially constructed. One might say that Christianity is socially constructed and *a fortiori* Islam and Protestantism are, but I am not so sure about Judaism. Did the Jews know they were constructing a religion? And when did it start? It wasn't even called 'Judaism', and the covenant between God and Israel took place after the religion, at least if you believe in the Bible.

Here we are entering ancient ages, where the notion of 'social construction' seems to be problematic if not altogether ridiculous. Arguing that animals have a social organization is a form of anthropomorphism: the bee queen is not actually a queen. In the same way, one can do nothing but smile at Pliny when he speaks of the religion of elephants. Of course one can see a continuity between the alpha male in wolf packs and the CEOs of multinational corporations or bullies on facebook. But this proves, in fact, that 'alpha male' is *not* a socially constructed notion, since its origin lies in a past in which we cannot – if words have meaning and we are not willing to seriously support the thesis according to which the hermit crab is the ancestor of squatters – speak of society. Indeed, how could something be socially constructed at a time when there is no society in any serious sense of the term? Wolf packs do not bury the corpses of their members, they do not administer justice; they celebrate no weddings and have no taboo against incest or cannibalism. Rather than being 'socially constructed', the burial of the dead, the various forms of union between people, the administration of justice and taboos mark the passage from nature to culture. After them there can be social construction, but not before.

At this point, once there is a society (and a society, at least in its earliest forms, is not something socially constructed, otherwise we would enter the vicious circle of the social construction of society, which is the same circle we find in the social contract), through a gradual process – as gradual as the transition from early hominids to the *directeurs d'études* at the EHESS – we get to social constructions (absolute monarchy, interest rates) and to social justifications or discredits of natural facts. An enlightened culture blames the

alpha male, Clint Eastwood fans appreciate it, but the alpha male is neither socially constructed nor socially dependent, nor mind-dependent. The alpha male is part of nature, since nature admits hierarchical structures and, indeed, is inherently hierarchical – whereas the main effort of culture is to deconstruct this hierarchy.

Now, let us consider the gender issue, one of the flagships and underlying motivations of internalism and social construction. To say that genders are socially constructed is very important from the political point of view, since the strong ideological weight of the category of 'nature' makes it more enticing to say that women or slaves have different *physei*, thereby justifying their subordination. But, if things are as I said, this is only a rhetorical move, which is understandable, but unfounded: the subordination of women and slavery are *socially dependent*.

Philosophically speaking, the opposition to slavery, female subordination etc. is the one to be *socially constructed*. And the most significant thing is that the reasons for the opposition do not depend on the solidarity-related strategies of some benevolent internalism, but on the perception of something that was both social and external to consciousness. At some point, in *some* cultures (and not in others) slavery or the subordination of women appeared unacceptable, and we proceeded to the social construction of anti-slavery and anti-sexism. But these phenomena we now react against were long part of society, along with the alpha male, and belonged to a legacy prior to the formation of society itself – which, by the way, explains why they appear so beastly. From this point of view, history is indeed a revelation in which pieces of a huge non-constructed collective unconscious progressively come forward. And it is very likely that, within a few years, many other pieces of this unconscious will appear, as history goes much faster today than ever before.

3. Emergentism

As I mentioned above, in *King Lear* we find the famous sentence: 'I'll teach you differences'. In *Hamlet* we find another well-known and often quoted passage: 'There are more things in heaven and earth, Horatio, Than are dreamt of in your philosophy'. There are many more differences in things than in the spirits contemplating them. The Inuit people have ten names for the colour white. This is not because the names create the colours, but simply because the colours are there and emerge in the environment, standing out much better as they are all together, so that their comparison and differentiation become eas-

ier. The fifty shades of gray we see do not depend on the famous pornographic novel, but on the fact that gray is in fashion, and this has made it easier to recognize different shades of this colour; these shades certainly existed prior to the names, as shown by the fact that so many colours are named after flowers. This is the fundamental intuition behind emergentism.

In the first section I showed the unsustainability of a generalized internalism. In the second I showed you how the scope of externalism is much broader than we are willing to admit. At this point, however, there is a rather obvious question, which concerns truth. Internalism erases the notion of 'truth', making it indistinguishable from error. On the contrary, externalism gives great importance to truth, but at the same time it comes across the difficulties of the theory of correspondence, which Engel rightly considers essential but problematic. In fact, there is an inherent difficulty in the idea that the mind relates to the world producing a magical event that we call 'truth.' Now, the magical bit is already greatly reduced if we integrate correspondentism with coherentism, instead of opposing them.

It may be true that if we look at *our* body, *this* paper, *this* fire, we might be overwhelmed with sceptical doubts. But these doubts, so plausible when we are alone, are much reduced in a sphere of interaction and interobservation. Typically, when a philosopher wants to be a sceptic, he explains his scepticism by questioning the existence of things that are on his desk, and not those found on a restaurant table surrounded by diners (with a form of coherentism that, in fact, confirms correspondentism) interacting with one another and proving the existence of the external world. It may be objected that the interaction between coherentism and correspondentism is a antisceptical *ontological* argument, that still does not solve the epistemological difficulties of correspondentism: in fact, how does the mind to faithfully represent the world? I would like to respond to this objection with the theory of emergentism, which means the following: the mind relates to the world without difficulty because, first of all, it does not represent it, but rather *records* it and, secondly, because in most cases it is not we who seek the world, but the world seeks us, encountering us and often upsetting us.

Let me try and clarify what I mean. Austin rightly said that, just like with marriage, it takes two to make a truth. We could push the metaphor a little further noting that, just like the spouses, the two poles of truth are rarely equivalent, if ever. There is a solemn concept of truth, the one that is sanctioned by the Nobel laureates in physics, in which one partner chased the other across seas and mountains, and sees truth as the culmination of a romantic epic. But there is also an ordinary concept of truth in which the partner has found a soul

mate next door, without any effort. Or one may realize too late that the soul mate was the one who wanted to get married at all costs, and that the partner was not even that much of a soul mate, after all. Of course these are anthropomorphisms, but they clearly illustrate why certain things always appeared to be obviously true without us ever reflecting on them. It also explains why unexpected or unpleasant truths appear before us, with irrefutable evidence, and without us ever seeking them.

Now, the mind does not necessarily have to represent the world for the encounter between mind and world to take place in the form of correspondentism. The Aristotelian theory of knowledge, which lies at the basis of correspondentism, is not a representational theory – a sign that correspondentism in itself implies by no means representationism. Aristotle's thesis is that the form of things is placed in the soul, without the substance, but that does not mean that the forms are present in analogical form: the soul does not turn green or square when it sees something green or square. That this is not a kind of representation is made clear by the fact that Aristotle, like all ancient philosophers, does not compare the soul to a dark room or a canvas, but to a wax tablet: a writing surface on which thoughts and feelings are imprinted. Note that the Greek writing was alphabetical, not ideographic, and what was imprinted were not images, but the symbolic or stenographic recordings of things. This is even more evident in Plato, who argues that first there is a writer, which only later is joined by a painter who illustrates impressions (in terms of *reconstruction* of experience, not of experience itself, one imagines).

These correspondist theories assume a theory of truth as recording, not as a representation. A trace is recorded, and the gradual accumulation of traces produces knowledge, which can be adequate even if it is not necessarily representative (it is not similarity that makes us think that when we create a mental image of our parents we are thinking of our parents!). It is essential to note that statements are not 'representations' of states of affairs: there is no similarity. We have no difficulty in thinking that our inner painter does not exactly belong to the figurative school: a state of things, which is imprinted in many different forms, emerges. What we cannot do without and is absolutely necessary is the recording that allows what emerges from the outside to be imprinted.

In the frame of the emergencist theory of truth – which, I repeat, cannot be considered separately from correspondentism and coherentism – there may well be competence (a true ontological relationship with something) without understanding (an epistemological relationship). Objects exert a peculiar affordance towards us and interact with us with an 'invitation' that, in the case of artefacts, was not even present in the mind of the inventor (the person who

invented coffee cups did not foresee their use as pen holders, and the person who invented the cell phone did not foresee its evolution into a typewriter and archive). The gradualist theory of knowledge in Leibniz illustrates this point very well: we have obscure perceptions, clear but confused, and only occasionally clear and distinct ones. As we can see, we are dealing with an evolutionary theory of truth, which regards representationalism as an emergence that is rather sporadic in the cognitive process.

This competence without understanding appears in a countless number of demonstrations in the constant interaction not only between human beings (who share the same world, but look at it from different perspectives), but also between beings who have totally heterogeneous perceptual apparatuses and conceptual schemes – or none at all. It would obviously be difficult to argue that this interaction is made possible by the sharing of conceptual schemes or representations. What kind of representations could I share with a bat when I am trying to dodge it, while helping it understand where the window is, so that it can go out? Once we have made all these considerations, we will understand that the concept of evidence has nothing mystical or subjective about it. The 'feeling of evidence' is certainly something that may accompany wrong evidence – no one has ever denied that error is possible. Rather than the sign of truth, evidence must be considered (along with surprise and disappointment) as belonging to the realm of all those experiences that demonstrate the emerging nature of the real, its coming from the world toward the subject, and not the opposite. This can undoubtedly be a source of bad surprises, but it is also true that without the world words like 'happiness' and 'unhappiness' would not make sense. Indeed, 'happiness is overrated: It's better to be right.'

6

Truth and Excluded Middle in *Metaphysics* Γ 7

PAOLO CRIVELLI

In chapters 7 and 8 of book Γ of the *Metaphysics*, the last two chapters of the book, Aristotle examines the Principle of Excluded Middle. He offers several arguments in its support. The purpose of this study is to reconstruct and evaluate the first of these arguments, which is based on a definition of truth and falsehood.

What principle is at stake? When in *Metaphysics* Γ he discusses a principle or principles which commentators normally call 'the Principle of Excluded Middle' (henceforth 'PEM'), Aristotle uses variants of two formulations:

> [a] It is not possible for there to be anything in the middle of a contradiction[1]

and

> [b] It is necessary either to affirm or to negate any one thing of one thing[2]

Elsewhere in the *Metaphysics* and in other works, Aristotle uses mainly variants of [b].[3] Only once, in the *Physics* (5.5, 235^{b} 15–16), does he employ the formulation 'Everything must either be or not be', which may be plausibly cashed out as 'Everything must either be so-and-so or not be so-and-so' (where 'so-and-so' can be replaced with any general term).

Formulations [a] and [b] might induce one to believe that in Aristotle's view PEM is a linguistic or 'logical' principle,[4] i.e. a thesis that concerns exclusively linguistic expressions or speech-acts: either the claim that there is no linguistic expression intermediate between affirmative and negative declarative sentences or the claim the only truth-evaluable linguistic expressions are affirmative and negative declarative sentences. Such an exegesis however sits uneasily with the fact that at several points of his discussion Aristotle appears to treat the denial of PEM as an ontological claim. (1) At the end of his first argument in support of PEM (1011^{b} 23–9), Aristotle describes (1011^{b} 28–9) the person denying it as committed to something that neither is nor is not, i.e. something that neither is so-and-so nor is not so-and-so. (2) In his second argument in support of PEM (1011^{b} 29–1012^{a} 1), Aristotle distinguishes two ways of understanding the position that there is something in the middle of a contradiction: either the thing in the middle of a contradiction is like something grey between black and white or it is like something that is between man and horse by being neither a man nor a horse. He goes on to argue that things in such a condition would be exempt from change and claims that such a conception is untenable. Here, the thing that is supposed to be in the

[1] Cf. 1011^{b} 23–4; 1011^{b} 30; 1011^{b} 35; 1012^{a} 26.

[2] Cf. 1011^{b} 24; 1012^{a} 2–3; 1012^{b} 11–12; 4, 1008^{a} 3–4.

[3] Cf. *Int.* 13, 22b12–13; *APo.* 1.1, 71^{a} 14; 4, 73^{b} 23; 11, 77^{a} 22; 77^{a} 30; 32, 88^{b} 1; *Metaph. B* 2, 996^{b} 29; Frede (1985), 79–80.

[4] Cf. Cavini (2007), 147.

middle of a contradiction does not seem to be a linguistic expression intermediate between affirmative and negative declarative sentences; rather, it seems to be an entity in a condition that in some sense falls between those of being so-and-so and not being so-and-so. (3) In his fourth argument in support of PEM (1012^{a} 5–9), Aristotle argues that one cannot assert that the principle that nothing falls in the middle of a contradiction fails only for a restricted area: if one takes this principle to fail, one must go for a universal failure. The person defending such a position is therefore committed to the claims that one will neither be right nor not be right and that 'there will be something outside what is and what is not [*παρὰ τὸ ὂν καὶ τὸ μὴ ὄν*]' (1012^{a} 7–8). Again, the things supposedly in the middle of a contradiction seem to be entities in a condition that in some sense falls between those of being so-and-so and not being so-and-so. (4) In the chapters of *Metaphysics* Γ that precede those dealing with PEM, Aristotle examines the Principle of Non-Contradiction, which he expresses both by an 'ontological' formulation ('It is impossible for the same thing to hold and not to hold of the same thing at the same time and in the same respect')[5] and by a linguistic or 'logical' formulation ('It is impossible to affirm and negate truly the same thing').[6] It wold be surprising if in his discussion of PEM Aristotle were to adopt exclusively linguistic or 'logical' formulations.

Formulation [b], 'It is necessary either to affirm or to negate any one thing of one thing', undeniably concerns linguistic expressions or speech-acts. But the evidence just reviewed makes it reasonable to regard formulation [a], 'It is not possible for there to be anything in the middle of a contradiction', as an ontological principle. When he uses formulation [a], Aristotle probably does not mean that there is nothing in the middle of a contradictory pair consisting of an affirmative declarative sentence and the corresponding negative declarative sentence, but that there is nothing in the middle of a contradictory pair consisting of the situation that consists in something being so-and-so and the situation that consists in that thing not being so-and-so. If this is right, by employing formulation [a] Aristotle commits himself to all instances of the schema 'Everything either is so-and-so or is not so-and-so'. This solution is corroborated by a passage from *Metaphysics I* 4: '… there is nothing in the middle of a contradiction, but there is in the case of some privations: for everything is either equal or not equal, but not everything is either equal or unequal' (1055^{b} 8–10). In this passage, a claim expressed by means of a

[5] Γ 3, 1005^{b} 19–20, cf. 4, 1006^{a} 3–4.

[6] Γ 4, 1008^{a} 36–1008^{b} 1, cf. 1007^{b} 21–2; 1007^{b} 29–30; 1007^{b} 34; 6, 1011^{b} 20–1.

version of formulation [a] is justified by a claim expressed by an instance of 'Everything is either so-and-so or not so-and-so', which may be regarded as a mere stylistic variant of the corresponding instance of 'Everything either is so-and-so or is not so-and-so'. Note that in the *Categories* (10, 12^b 6–15) Aristotle holds that the relation of contradictoriness obtains not only between linguistic expressions like 'is sitting' and 'is not sitting', but also between what is 'under [ὑπό]' (12^b 6, 12^b 9, 12^b 14) these linguistic expressions.

The first argument for PEM, which is based on a definition of truth and falsehood, is as follows:

T1 ἀλλὰ μὴν οὐδὲ μεταξὺ ἀντιφάσεως ἐνδέχεται εἶναι 1011^b23
οὐθέν, ἀλλ᾽ ἀνάγκη ἢ φάναι ἢ ἀποφάναι ἓν καθ᾽ ἑνὸς ὁτιοῦν.
δῆλον δὲ πρῶτον μὲν ὁρισαμένοις τί τὸ ἀληθὲς καὶ ψεῦδος. $^b25$
τὸ μὲν γὰρ λέγειν τὸ ὂν μὴ εἶναι ἢ τὸ μὴ ὂν εἶναι ψεῦ-
δος, τὸ δὲ τὸ ὂν εἶναι καὶ τὸ μὴ ὂν μὴ εἶναι ἀληθές, ὥστε
καὶ ὁ λέγων[7] εἶναι ἢ μὴ ἀληθεύσει ἢ ψεύσεται· ἀλλ᾽
οὔτε τὸ ὂν λέγεται[8] μὴ εἶναι ἢ εἶναι οὔτε τὸ μὴ ὄν. 1011^b29

> Nor is it possible for there to be anything in the middle of a contradiction, but it is necessary either to affirm or to negate any one thing of one thing. First, this is clear to those who define what truth and falsehood are. For, to say that what is is not, or that what is not is, is false; to say that what is is, and that what is not is not, is true, so that it's he who says that something is or that it is not who will be right or wrong: but neither what is nor what is not is said not to be or to be. (Arist. *Metaph.* Γ 7, 1011^b23–9)

The difference between the 'or' in the definition of falsehood and the 'and' in the definition of truth is probably a purely stylistic matter.[9] The main difficulty posed by T1 is that it is hard to see how a definition, and in particular a definition of truth and falsehood, can serve the purpose of supporting a substantial thesis like PEM.

[7] The reading 'καὶ ὁ λέγων' is attested in E and J; A^b has 'ἐκεῖνο λέγων' (the reading printed by Brandis (1823), 83 and favoured, but not printed, by Schwegler (1847–8), III 182); Alexander (*in Metaph.* 328, 25) seems to have read 'καὶ ὁ λέγων το υτο' (printed and defended by Bonitz (1848–9), I 79 and II 212).

[8] E and J read 'λέγει', 'λέγεται' is in A^b.

[9] Cf. Bonitz (1870), 357^b 20–4; Cavini (1998), 12.

A pragmatic reconstruction. A first attempt at reconstructing the argument is based on the assumption that it has a rather pragmatic character, i.e. linked to the practice of conversation. The definition of truth and falsehood relies on the assumption that the only declarative sentences that can be true or false are affirmations and negations. For: to say of what is that it is not or of what is not that it is is to *negate* being of what in fact is or to *affirm* being of what in fact is not; to say of what is that it is or of what is not that it is not is to *affirm* being of what in fact is or to *negate* being of what in fact is not. Since the only cases contemplated by the definition are that of affirmation and that of negation, and since the definition presupposes that all possible cases are contemplated (for a definition that says nothing about some of the possible cases would be faulty), affirmations and negations are the only sentences that can be true or false. Thus, if anyone wants to produce a declaration, i.e. a truth-evaluable sentence,[10] he or she will have to produce either an affirmation or a negation. Hence there is no intermediate between an affirmative and a negative declaration. Such a claim may be regarded as supporting PEM, in particular of the principle expressed by the second of the two formulations mentioned at the beginning of T1: 'It is necessary either to affirm or to negate any one thing of one thing' (1011^{b} 24).

This reconstruction faces some objections. (1) It credits Aristotle with a defence of a version of PEM which is far from the ontological version which there are reasons to attribute to Aristotle (i.e. a claim to the effect that everything either is so-and-so or is not so-and-so). (2) It does not make much of the last part of the text, i.e. of the remark that 'neither what is is said not to be or to be, nor what is not [*sc.* is said not to be or to be]' (1011^{b} 28–9): this remark does not immediately lend itself to be read in a way that agrees with the reconstruction under consideration. (3) The version of PEM defended by Aristotle according to the reconstruction under consideration is disappointingly weak because it amounts to the claim that every declarative sentence is either an affirmative or a negative declarative sentence. This claim enjoys the double drawback of being false (because some declarative sentences, e.g. disjunctive and conditional ones, cannot be classified as affirmations or denials) and of clashing with Aristotle's own pronouncements in *de Interpretatione* (5, 17^{a} 8–9, 17^{a} 20–2), where he mentions affirmation and negation as the two types of *simple* declarative sentence while allowing for the existence of other declarative sentences (those which are one by composition and thanks to the presence of some connector).

[10] Cf. *Int.* 4, 17^{a} 2–7.

A reconstruction based on the Principle of Bivalence. Some commentators put forward an interpretation that does not incur the difficulties faced by the one which has just been considered and relies on a variant of the principle normally called 'the Principle of Bivalence' (henceforth 'PB').[11] PB states that every declarative sentence is either true or false.[12] The variant of PB on which the argument relies is the claim that 'he who says that something is or that it is not will be right or wrong' (1011^{b} 28), i.e. the claim that both someone who produces an affirmation by saying about something that it is so-and-so is either right or wrong and someone who produces a negation by saying about something that it is not so-and-so is either right or wrong (here 'so-and-so' can be replaced with any general term).

The easiest way to see how this interpretation goes is to present it as a reductio ad absurdum of the assumption that there is an exception to PEM in its ontological formulation, i.e. as a reductio ad absurdum of the assumption that there is an exception to the claim that everything either is so-and-so or is not so-and-so. Thus, suppose there to be such an exception, i.e. that there is an object x that neither is so-and-so nor is not so-and-so. Consider anyone who produces an affirmation by saying about x that it is so-and-so: this person will be neither right (because, according to the definition of truth and falsehood,[13] in order for him or her to be right, x should be so-and-so, while x by hypothesis is not so-and-so) nor wrong (because, according to the definition, in order for him or her to be wrong, x should not be so-and-so, while by hypothesis it is not the case that x is not so-and-so). This clashes with the version of PB on which the argument relies, which requires that someone who produces an affirmation by saying about something that it is so-and-so is either right or wrong. Consider then anyone who produces a negation by saying about x that it is not so-and-so: this person will be neither right (because, according to the definition of truth and falsehood, in order for him or her to be right,

[11] Cf. Alex. Aphr. *in Metaph.* 328, 19–329, 4; Schwegler (1847–8), III 182; Bonitz (1848–9), II 212; Ross (1924), I 284–5; Kirwan (1971/93), 117–18.

[12] Aristotle characterizes declarative sentences as the sentences of which truth and falsehood hold (cf. *Int.* 4, 17^{a} 2–3). This characterization may be taken to require merely that truth and falsehood hold *only* of declarative sentences; it need not be taken to require that either truth or falsehood holds of *every* declarative sentence (cf. Crivelli (2004), 86–7). Thus, the version of PB in the main text above need not be regarded as a logical consequence of the characterization of declarative sentences as the sentences of which truth and falsehood hold.

[13] The exegesis under consideration assumes that Aristotle's definition of truth and falsehood involves a predicative elliptical use of 'to be', i.e. a predicative use of 'to be' where the predicated general term is omitted for the sake of generality. Such a reading of Aristotle's definition is endorsed by several commentators: cf. Sommers (1969–70), 281–2.

x should not be so-and-so, while by hypothesis it is not the case that *x* is not so-and-so) nor wrong (because, according to the definition, in order for him or her to be wrong, *x* should be so-and-so, while *x* by hypothesis is not so-and-so). This also clashes with the version of PB on which the argument relies, which requires that someone who produces a negation by saying about something that it is not so-and-so is either right or wrong. Thus, the variant of PB on which the argument relies rules out an exception to PEM in its ontological formulation. In other words, the variant of PB on which the argument relies requires that everything either be so-and-so or not be so-and-so. The second branch of the argument, which concerns someone who produces a negation by saying about *x* that it is not so-and-so, is redundant: the first branch of the argument suffices. The second branch is offered merely because producing only the first would give the wrong impression that the argument can go through only by considering the case of affirmations.

This interpretation of Aristotle's argument has several strengths: it is close to the actual wording of the argument's second part and it yields as a conclusion an ontological version of PEM, i.e. the claim that everything either is so-and-so or is not so-and-so. But it also faces some objections. Specifically, the interpretation under consideration crucially relies on a variant of PB, which invites two objections. (1) Aristotle himself in chapter 9 of *de Interpretatione* denies PB while accepting PEM (at least according to the most widespread interpretation of this chapter):[14] it would be awkward on Aristotle's part to argue for PEM on the basis of PB. (2) It is not clear that Aristotle's argument would be effective against someone who denies PEM: such a person would probably have no qualms rejecting also PB.[15] The first criticism may perhaps be dealt with by noting that in *Metaphysics* Γ there is no indication of an exception to PB such as the one usually found in *de Interpretatione* 9: this might be an indication that *de Interpretatione* 9 is a late piece and that at the time when he wrote *Metaphysics* Γ Aristotle endorsed PB. As for the second criticism, one might try to answer it by claiming that the effectiveness of a defence of PEM based on PB can only be evaluated by taking into account the motivation that one's antagonist might have for rejecting PEM. Aristotle mentions three reasons that might induce someone to reject PEM (1012^{a} 17–28): giving in to eristic arguments, demanding a reason for everything, and a metaphysical view such as that of Anaxagoras (in a situation of complete mixture, things are allegedly neither good nor not good). In the case of the third

[14] I defended this interpretation of *de Interpretatione* 9 in Crivelli (2004), 198–233.

[15] Cf. Kirwan (1971/93), 117–18.

type of motivation, one might expect that someone rejecting PEM might still want to endorse PB (because bearers of truth or falsehood might be deemed to be foreign to the condition of complete mixture envisaged by Anaxagoras). This reply is however not convincing because it leaves the other motivations mentioned by Aristotle unaccounted for.

A new reconstruction. We have considered two reconstructions of Aristotle's argument in T1. The first reconstruction does not fit in well with the argument's final part; the second saddles Aristotle with an argument that relies on PB, a principle at least as controversial as PEM. It is reasonable to search for a new exegesis that fits the whole of Aristotle's formulation while crediting him with a plausible argument.

Suppose that there were a condition, call it 'M', which is 'in the middle of a contradiction' (1011^b 23), i.e. intermediate between the condition of being so-and-so and the contradictorily opposite one of not-being so-and-so. The opposition between the condition of being so-and-so and that of not-being so-and-so does not have to do with the attribute so-and-so: both conditions are ways of being related to the attribute so-and-so. The opposition between the two conditions depends on the fact that their constitutive relations to the attribute so-and-so are themselves opposed: things in these conditions are related to the attribute so-and-so in opposite ways. For this reason condition M, which is supposed to be intermediate between the two opposed conditions, consists in being related to the attribute so-and-so in a way that is different both from that of being so-and-so and from that of not-being so-and-so.

Given that condition M exists, there must also be a predicative expression, say 'neither-is-nor-is-not so-and-so', that corresponds to condition M in that it is used to say of things that they are in condition M. This predicative expression, 'neither-is-nor-is-not so-and-so', would then be truly applicable to any entity in condition M. We thus have three different conditions, namely being so-and-so, not-being so-and-so, and M, and three corresponding predicative expressions, namely the affirmative predicative expression 'is so-and-so', the negative predicative expression 'is-not so-and-so', and the intermediate predicative expression 'neither-is-nor-is-not so-and-so'. Just as the difference between the three conditions is determined (not by the attribute so-and-so, but) by their different constitutive relations that combine with the attribute so-and-so, so the difference between the three predicative expressions is determined (not by the general term 'so-and-so', but) by the predicative links that combine with the general term 'so-and-so', namely the affirmative pred-

icative link '… is …', the negative predicative link '… is-not …', and the 'intermediate' predicative link '… neither-is-nor-is-not …' Being constructed around the 'intermediate' predicative link '… neither-is-nor-is-not …', which is different both from the affirmative '… is …' and from the negative '… is-not …', the intermediate predicative expression 'neither-is-nor-is-not so-and-so' is neither affirmative nor negative. Thus, the intermediate predicative expression 'neither-is-nor-is-not so-and-so' is different both from the affirmative predicative expression 'is so-and-so' and from the corresponding negative predicative expression 'is-not so-and-so'. A clear indication of this difference is given by the fact that if something were in condition M, it could be truly described by 'neither-is-nor-is-not so-and-so', but would neither be so-and-so nor not be so-and-so and therefore could not be truly described by means of the affirmative predicative expression 'is so-and-so' nor by means of the negative predicative expression 'is-not so-and-so' (cf. 1011^{b} 28–9).

An application of the predicative expression 'neither-is-nor-is-not so-and-so' could then be described as an exception to the claim that 'it is necessary either to affirm or to negate any one thing of one thing' (1011^{b} 24). However, the only cases contemplated by the definition of truth and falsehood are those of affirmation and negation. Since the definition presupposes that all relevant cases are contemplated (for a definition that says nothing about some relevant cases would be faulty), affirmations and negations are the only sentences to be considered when issues of truth and falsehood come up: the definition entails that 'it's[16] he who says that something is [*sc.* affirms] or that it is not [*sc.* negates] who will be right or wrong' (1011^{b} 28). Hence, according to the definition, the only predicative expressions are affirmative ones and negative ones, so there is no place left for an intermediate predicative expression that is neither affirmative nor negative. Hence the definition of truth and falsehood tells against the existence of a condition M 'in the middle of a contradiction' (1011^{b} 23), i.e. intermediate between the condition of being so-and-so and the contradictorily opposite one of not-being so-and-so. Therefore everything either is so-and-so or is not so-and-so.

This reconstruction has the advantage of fitting the whole formulation of the argument and assigning a role to each of its clauses. Its drawback is that it relies on a premiss that does not appear in the text, i.e. the assumption that if there were a condition M which is different both from being so-and-so

[16] I regard the occurrence of 'καί' at 1011^{b} 28 as emphatic: it indicates that it is just the person who is making an affirmation or a negation who speaks truly or falsely. For the emphatic use of 'καί' (whereby it may also be rendered by 'just'), see LSJ *s.v.* 'καί' B 6; Denniston (1954), 320–1.

and from not being so-and-so, then there would be a predicative expression 'neither-is-nor-is-not so-and-so' that could be used to offer a true description of any entity that enjoys condition M. The absence of this assumption from the argument is somewhat disturbing in view of its crucial importance within the argument it contributes to.

1. References

Avgelis, N. and Peonidis, F. (eds.) 1998, *Aristotle on Logic, Language and Science*, Thessaloniki.

Bonitz, H. (ed. and comm.) 1848–9, *Aristotelis Metaphysica*, Bonn.

— 1870, *Index Aristotelicus*, Berlin.

Brandis, C. A. (ed.) 1823, Aristotelis et Theophrasti *Metaphysica*, Berlin.

Cavini, W. 1998, 'Arguing from a Definition: Aristotle on Truth and the Excluded Middle' = Avgelis and Peonidis (1998), 5–15.

— 2007, '*Principia contradictionis*. Sui principi aristotelici della contraddizione (§§ 1–3)', *Antiquorum philosophia* 1: 123–69.

Crivelli, P. 2004, *Aristotle on Truth*, Cambridge.

Denniston, J. D. 1954, *The Greek Particles*, 2nd edn, Oxford.

Frede, D. 1985, 'The Sea-Battle Reconsidered: A Defence of the Traditional Interpretation', *Oxford Studies in Ancient Philosophy* 3: 31–83.

Kirwan, C. (trans. and comm.) 1971/93, *Aristotle, Metaphysics, Books* Γ, Δ, *and E*, 2nd edn, Oxford.

Ross, W.D. (ed. and comm.) 1924, *Aristotle, Metaphysics*, repr., Oxford 1975.

Schwegler, A. (ed., trans., and comm.) 1847–8, *Aristoteles, Die Metaphysik*, Tübingen.

Sommers, F. 1969–70, 'On Concepts of Truth in Natural Languages', *Review of Metaphysics* 23: 259–86.

7

Littérature et vérité. Engel lecteur de Benda

FRÉDÉRIC NEF

Le logicien Boole, qui fonda la logique algébrique, écrivit en 1854 *An Investigation of the Laws of Thought.* Pascal Engel substitue aux lois de la pensée, celles qui permettent de déduire, d'inférer, de généraliser, les lois de l'esprit qui gouvernent de manière plus large la recherche intellectuelle, qui recouvrent ce que l'on entend de nos jours par 'normes épistémiques', c'est-à-dire des normes de la connaissance désintéressée, scientifique notamment : objectivité, détachement… Pascal Engel est à la fois un philosophe de la science, un épistémologue, qui explicite ces normes scientifiques et s'interroge sur leur nature, le type de contrainte qu'elles exercent et un philosophe de l'esprit qui poursuit une enquête patiente sur les normes de la connaissance en général, sur la vérité et l'expression vérace des croyances véridiques. C'est dans cette double perspective que se situe la publication de son dernier livre, sur Benda. On s'expliquerait mal en dehors de cette double continuité le passage d'une réflexion sur les croyances vraies à une méditation sur la littérature par le biais d'une lecture de Benda. Le philosophe et logicien Michael Dummett a écrit un livre sur le tarot, d'autres philosophes ont écrit sur le catch (A. Philonenko) ou les conneries (H. Frankfurt) mais Engel ne se situe pas dans cette optique de défi intellectuel : Julien Benda est un personnage hors du commun, mais ce n'est pas ce qui intéresse Engel : il ne cherche pas le *tour de force*.

Pascal Engel souhaite en effet étendre à la littérature un travail sur les normes de la connaissance, du travail intellectuel, commencé sur les sciences (théorie de la justification), la philosophie spécialisée (philosophie de l'esprit et du langage). Cela fait partie d'un mouvement certes marginal mais important : des philosophes spécialistes de l'esthétique comme Roger Pouivet se sont illustrés déjà par ce genre de démarche. L'idée que l'on apprenne à aimer

dans les romans (idée assez voisine d'une conception normative de la littérature) est une idée couramment débattue – on s'accorde souvent de nos jours sur le fait que la littérature est une sorte de base de données encyclopédiques sur la manière de monter à cheval, de nouer sa cravate, tout autant que de faire une scène de jalousie réussie ou de souffrir avec distinction de l'ingratitude de nos collègues (en ce sens un des écrivains favoris de Engel, Woolhouse est une mine). Pascal Engel se situe *grosso modo* dans ce mouvement mais il s'y inscrit à partir de sa position propre : il ne s'intéresse pas tant à la vertu éducatrice de la littérature qu'à sa fonction de connaissance, de vérité. On connaît les travaux de Pascal Engel sur la vérité et on retrouve dans l'analyse de la littérature les mêmes interactions entre croyance, normes et vérité.

Mais alors, cela posé, pourquoi Benda ? Pourquoi pas Mallarmé ? Melville ? H. James (que Engel aime tant) ? Samuel Johnson (que Engel me fit lire). Pourquoi un auteur vieillot au style suranné, voire ampoulé ? Pascal Engel reconnaît d'ailleurs la faible valeur du versant strictement littéraire, surtout narratif, de l'œuvre de Benda, marqué par la préciosité et l'artifice. Je crois que Pascal Engel s'est attaché à Benda, parce que la situation du philosophe qui croit à la norme de vérité dans le contexte de la pensée actuelle, constructivisme, contextualisme, perspectivisme, relativisme du genre (*gender studies*) ou de la culture, est analogue à celle de Benda qui dans les années 30 ne croit pas à la toute puissance de la littérature, prônée à la NRF et qui dans les années 50 s'oppose à la funeste théorie sartrienne de l'engagement. Engel traite dès lors le versant théorique et polémique de l'œuvre de Benda (sans oublier des textes comme *Les mémoires d'un enterré vif*). Engel donc ne choisit pas seulement Benda parce qu'il serait un écrivain qui croit à la vérité, mais parce qu'il le traite pratiquement comme un égal, à la fois humainement, par cette faculté de résistance (qui provient autant d'une allergie musilienne à la sottise que de la vertu intellectuelle, empressons-nous de l'ajouter) et intellectuellement par l'efficace décorticage des vices intellectuels de ses contemporains (vices des écrivains et des philosophes pour Benda, vices des philosophes pour Engel). Est-il besoin de saluer tout ce versant bathologique de Pascal Engel ? En ce sens Pascal Engel est à la foi le descendant de la tradition classique anglaise, de Swift, Pope, Samuel Johnson et de la tradition des moralistes français. Dans un autre sens, il prolonge la lignée des Taine, Cournot, Boutroux, Meyerson tragiquement négligés des deux côtés de l'Atlantique.

Je dois confesser que ayant lu Benda très tôt j'en avais conclu, dans mes catégories de jeunesse, à la fois qu'il était un écrivain réactionnaire et de droite (l'époque, qui succédait immédiatement à la Révolution Culturelle était à tout ce qu'il détestait et on ne faisait pas couramment cette différence entre être de

droite et être réactionnaire qui plus tard pouvait nous dédouaner aux yeux des progressistes et des humanistes du centre droit, alors que nous étions déjà plus ou moins secrètement des réactionnaires de gauche). Le livre de Louis-Albert Revah (*Julien Benda*) me l'avait rendu très antipathique tout en attirant mon attention sur sa judéité, et seul *la France Byzantine* resta pour moi au fil de ces années littéralement un livre de chevet. L'ouvrage de Pascal Engel, comme tant d'autres de ses livres, m'a ouvert les yeux et enlevé ce qui m'a toujours empêché d'y voir clair dans les multiples facettes du génie de Benda. Mais un peu d'obscurité demeurent et les questions se pressent : Qui est Benda ? Un non conformiste ? Un républicain de droite ? Il n'est pas sûr qu'on puisse répondre facilement à ces questions et en tout cas ce ne sont pas celles que Engel se pose.

Disons quelque mot sur Benda, puisqu'il est à peu près totalement oublié. Benda est né en 1867, mort en 1956. Cette longévité l'a rendu contemporain de l'affaire Dreyfus, de la guerre d'Espagne, des deux guerres mondiales et de la guerre froide. Il connut la réaction spiritualiste dans sa jeunesse, l'apogée du système NRF (ah ! la « *Kommandantur* de la rue Bottin » (E. Martineau) à l'âge mûr, et la capitulation des intellectuels français dans leur grande majorité, y compris Sartre, devant le stalinisme dans sa vieillesse. Il commença en bourgeois, en rentier, fut ruiné à la cinquantaine après une jeunesse aisée, et il finit en communiste, approbateur docile de procès manipulés par Moscou, ce qui pour un fanatique des cocktails en son jeune temps, de la vérité et de la raison n'est pas totalement surprenant. Ami un temps de Péguy, soutenu par Paulhan à la NRF, symétrique de Thibaudet, Benda n'était pas un solitaire vaticinateur et ronchon, une sorte d'arbitre auto-proclamé des mœurs littéraires de son temps animé par une rage équivoque en faveur de la vérité, quoique cette position puisse, si elle n'est pas haineuse, comme chez Marcel Aymé, dans le *Confort Intellectuel* – qu'il est de mauvais ton d'apprécier – avoir quelque charme.

Il est parfaitement distinct aussi de la critique radicale de la société moderne, et qui englobe une critique tout aussi radicale des mœurs et des théories littéraires chez ceux que l'on appelle à la suite d'Antoine Compagnon les 'anti modernes' (Bloy, Bernanos…) bien que ceux ci se désintéressassent des jeux littéraires qu'ils trouvaient futiles et peu dignes d'effort et qu'ils versassent parfois dans la mystique que haïssait Benda en bon rationaliste. Benda est aussi différent d'un Thibaudet (son exact contemporain), élève de Bergson auteur du *Bergsonisme* (1924), Thibaudet qu'il critique souvent, mais qui partage avec lui tant de choses, séparé cependant de lui donc par son bergsonisme. Ce dernier mot, 'bergsonisme', livre une dernière clé pour ce rappel

rapide de qui fut Benda : cette mode intellectuelle a représenté pour Benda l'entièreté des vices intellectuels. Bien entendu, on reviendra sur ce point capital, la bergsonophobie de Benda, mais il faut d'ores et déjà insister sur le fait que Benda attribue à l'influence de Bergson l'anti-intellectualisme de son époque et son influence négative sur la littérature, la critique et la théorie littéraire. Benda n'est pas loin de faire de Bergson un philosophe romantique (en tous les cas cela s'applique assez bien aux bergsoniens, que la durée et l'intuition enivrent). Bergson est certes une des pires catastrophes intellectuelles qu'a connu la France (avec probablement la nomination du logicien Ramus en 1551 au Collège de France), et Benda ne fut pas le seul à s'en apercevoir (sur ce point il était allié, 'objectivement' comme disent les marxistes, à certains catholiques, par exemple Maritain et Blondel, quoique le spiritualisme de ce dernier s'oppose au sien propre, athée) mais il fut probablement le plus acharné dans sa critique. Hélas, l'irrationalisme de Bergson a gagné, mais un vaincu comme Benda n'en a que plus de valeur : il s'est opposé de manière résolue au ressac obscur, vague et prétentieux qui devait emporter une grande partie de la philosophie française. Comme le dit si bien Sollers dans l'*Eloge de l'infini* : « il y a eu des tripotées de médiocres dans la philosophie française ».

Revenons donc à la connaissance littéraire et à sa radiographie par Engel à travers Benda. Celui-ci rejette de manière quasiment véhémente le culte de la littérature rendu par la grande majorité des intellectuels français, favorisé par le mélange de littérature et de philosophie du style bergsonien et le caractère littéraire de presque tous ses concepts, vagues et chatoyants. On pourrait rétorquer que cela remonte même à l'âge classique et par exemple aux écrivains du Grand Siècle, comme Pascal, Bossuet ou Corneille. Pour Benda il n'en est rien. A travers Corneille ce sont des valeurs héroïques auxquelles on rend hommage, tandis qu'à travers Rimbaud c'est à la figure de l'écrivain absolu qu'on rend hommage (quand, encore pire, on ne rend pas hommage à Rimbaud à travers Rimbaud, culte de la singularité purement tautologique et vide de contenu). On tient le rationalisme comme opposé à la littérature et l'anti-intellectualisme est de règle, comme on peut le voir dans les pages (injustes) que Benda consacre à Claudel dans *La France Byzantine*. Depuis, de nouveaux abysses de l'absence d'intellect ont été explorés par les bathyscaphes de l'autofiction, de la post littérature, du métissage des cultures. Le culte de la littérature va donc de pair avec la salutation du vide et le conformisme moral et politique de la pensée.

Selon Engel, Benda conduit cette offensive contre les irrationalistes littéraires parallèlement à celle contre les philosophes irrationalistes. D'un côté Mallarmé, Valéry, Gide … de l'autre Bergson, Lavelle, Sartre, Jean Wahl …

Engel désigne Benda comme le 'tonton flingueur du bergsonisme'. Il est vrai qu'il y a chez Benda une tournure obsessionnelle dans sa haine de Bergson – « il y a plus de vingt ans que Benda me poursuit avec une haine et un acharnement pour lesquels tous les moyens sont bons. Je crois qu'il y a des hommes qui veulent le mal pour le mal » (Bergson in Engel, p. 84). On doit avouer que la philosophie française n'a guère eu de chance. Avec Bergson le seul grand métaphysicien avec Malebranche a versé dans le mysticisme mou, compatible avec la séparation de l'Eglise et de l'Etat, et dans l'irrationalisme. Mais Benda n'est pas spécialement concerné par le triste destin de la métaphysique française ; il critique la notion d'intuition de manière générale et finalement l'intuition mystique bergsonienne va au-delà de la métaphysique, beaucoup plus du côté de Plotin que du côté de Platon et cet au-delà de la métaphysique diffère pour Benda d'une « mysticité étrangère aux plaisirs des sens telle que l'ont pratiquée les grands mystiques, tels Böhme ou Madame Guyon » (*La France Byzantine*, p. 85). Et : « C'est une mystique sensuelle, un peu coquette, dans laquelle on fait carrière » id.)

Julien Benda a pu critiquer l'irrationalisme bergsonien *et* celui des existentialistes. La terrible catastrophe du tsunami bergsonien a détruit à la fois les infrastructures et les superstructures de la philosophie probablement pour plusieurs centaines d'années. En un certain sens Sartre a visé le même objectif de la destruction de la philosophie normative (éthique, métaphysique, logique) mais il a échoué sur le plan philosophique et triomphé provisoirement sur le plan littéraire et politique. Tout d'abord en un certain sens c'est un meilleur philosophe que Bergson (sauf peut-être si on met à part les *Données immédiates de la conscience*, mais Sartre aussi a commencé par de la psychologie rigoureuse cf. la période 1936-1940 de *l'Imagination* à *l'Imaginaire*). Mais il souffre de ses limitations : pas de philosophie de la connaissance ou de la science, ni éthique ni morale. Bergson a été accusé par les misogynes d'être un 'philosophe pour les dames' et Sartre est un philosophe pour les jeunes gens. Comment la philosophie française pouvait-elle se relever de ce double coup du sort ? La réponse est simple : elle ne s'en est jamais relevée. Pascal Engel est un philosophe qui mesure la radioactivité dans un champ de ruines. Il est plus désolé que méchant.

Une thèse intéressante de Benda que reprend et développe Engel est que l'existentialisme est un bergsonisme. Si on compare les univers thématiques et sensibles, on serait évidemment étonné que par delà cette identité supposée il y ait une telle différence : au visqueux, à l'informe s'oppose la boule de neige, la durée fuyante et la neige elle-même. McTaggart a discuté du caractère consolant que l'on peut attribuer à telle ou telle thèse philosophique. On peut

soutenir que Bergson souhaite retenir ce qui est consolant (d'où l'intérêt pour le métapsychique, le spiritualisme pour les nuls), et je fais l'hypothèse que les dames qui allaient l'écouter au Collège de France furent efficacement consolées. Sartre (comme Schopenhauer), au contraire, choisit chaque fois l'hypothèse qui désespère le plus, ce qui est aussi contestable et aussi peu rationnel – la bonne réponse c'est que l'on doit être indépendant de la consolation : C'est la religion qui est consolante, procure le confort spirituel ; la philosophie n'est de droit ni consolante, ni désespérante. Cependant tous nos jeunes gens le dimanche après-midi sur les ondes tentent de nous réconforter avec Nietzsche, Pascal etc. Bergson est à l'origine de cette déviance sentimentale et romantique de la philosophie.

Sur le fond, qu'est-ce qui permet à Engel de rapprocher Sartre et Bergson aussi étroitement ? On peut citer : d'une part leur philosophie de la liberté, cette dernière n'étant pas conçue comme un libre arbitre générateur de choix, comme chez Descartes et Leibniz, mais comme un 'acte créateur' et d'autre part l'opposition de la pensée et de l'existence. Ce qui les unit négativement c'est leur impossibilité de penser les catégories modales. On comprend qu'une philosophie comme la philosophie française ait pratiqué jusqu'à assez récemment et d'ailleurs aussi dans sa version analytique[1] un rejet assez tranché de la pensée modale, ce qui l'isole complètement des autres métaphysiques[2]. On voit donc que la lecture raisonnée de Benda nous permet de comprendre la situation actuelle de la philosophie française : il a vécu assez longtemps pour voir s'insinuer et s'instituer le bergsonisme et se combiner avec lui la pensée dite existentialiste. Au fond, ce qui vient après, la philosophie des années 60 et 70, tout de suite après sa mort, n'ajoute rien de décisif : qu'est-ce que Deleuze sinon un Bergson structuraliste ? Qu'est-ce que Derrida sinon un Sartre textualiste ou nihiliste ? Benda permet donc de comprendre mieux pratiquement un siècle de philosophie française. Mais à côté de cet apport radicalement critique, assez décourageant, il y a une autre facette de son travail, qui ouvre des perspectives positives : la recherche de la nature des normes et valeurs et de la connaissance intellectuelles et plus spécialement littéraires.

Pour Benda les valeurs éternelles sont soit intellectuelles, soit morales ; la valeur intellectuelle suprême est la raison et la valeur morale suprême est la justice. (p. 147, op. cit.). Ne nous attardons pas sur notre déception première : quoi de plus vague apparemment que la raison, quoi de plus dangereux que

[1] Cf. Stéphane Chauvier *Le sens du possible*, Vrin, Paris, 2010, un retour à la critique bergsonienne des modalités, une désinvolture toute française.

[2] Il faudrait nuancer évidemment ces affirmations. Pascal Engel a jadis montré que la philosophie française rationaliste avait développé une pensée des modalités (par exemple Renouvier).

la justice, combien de massacres ont été commis au nom de cette valeur ? Il est vrai que Benda ajoute la vérité scientifique à côté de la raison, mais c'est également ambigu. Quelle est l'interprétation de la mécanique quantique qui est *la* vérité scientifique ? Parmi les dizaines de systèmes de logique, lequel représente la vérité scientifique ? Ces valeurs 'cléricales' ne sont donc pas très convaincantes. Benda a raison de faire des valeurs fondamentales des valeurs éternelles, mais les valeurs éternelles qu'il choisit ne sont pas de nature à faire tenir tout l'édifice des normes et des valeurs. D'ailleurs Pascal Engel reconnaît que « Ni la vérité ni la raison ne sont en elles-mêmes des valeurs » (p. 160, op. cit.) et affirme qu'elles sont des propriétés descriptives des énoncés et des théories.

Engel inscrit sa défense, son apologie de Benda dans le cadre d'une théorie générale de la littérature, différente de ou opposée à la théorie structuraliste ou existentielle. Il le place à égale distance de l'esthétisme et du moralisme et le défend contre l'accusation d'intellectualisme qui lui fut souvent adressée. Le point sur lequel il est difficile de défendre un Benda c'est son silence sur le double système concentrationnaire qui se mit en place à son âge mûr. On peut comparer ce silence avec celui à l'égard du colonialisme il se tait devant les situations réelles et un irrationaliste, nihiliste comme Gide lui en remontre. Benda est plus attentif à la violence d'un style qu'à la violence de masse. C'est la raison pour laquelle on ne peut croire complètement à son 'culte buté de la vérité' dans le domaine de la politique. La partie de l'ouvrage de Engel sur Benda et la politique est par là même peut être la plus passionnante, car elle révèle l'ampleur du débat sur la position de Benda, mais elle est aussi probablement la plus décevante en ce qui concerne la personnalité de Benda. Comment croire à une théorie politique dont son auteur finit au Parti Communiste dans les années de plomb ? Comment a-t-il pu croire que c'était une protection ? Si l'on compare avec George Orwell, ou même avec Raymond Aron, cela devient cruel. Mais en France dans les années 50 il ne fut pas le seul parmi les intellectuels à soutenir activement la Russie stalinienne, bien que cela ne l'exempte nullement, lui qui avait revendiqué, à raison, la lucidité politique la plus exigeante. Il n'a pas eu l'itinéraire d'un professeur d'épistémologie à la Sorbonne qui se rendit coupable d'un « Staline savant d'un type nouveau » et c'est à l'aune de tels désastres moraux qu'il faut le juger.

Ce livre sur Benda a le mérite de situer dans le contexte actuel (malgré la fragilité de la comparaison actuelle entre notre époque et les années 30) l'ensemble des idées, elles-mêmes profondément contradictoires (ce qui en fait d'ailleurs l'intérêt). Il est certain que Benda ne mérite pas tant d'éreintement, et que Engel le défend souvent fort bien contre les attaques qu'elles soient de

droite ou de gauche, mais mérite-t-il vraiment autant d'éloges ? Le versant défense est plus réussi que le versant apologie, et ce n'est pas la faute d'Engel : les critiques adressées à Benda étaient particulièrement malveillantes, de mauvaise foi et souvent stupides, mais ses idées n'étaient pas toujours forcément très profondes (il était plus brillant que profond), ce qui était normal vu son absence de formation solide, si on le compare par exemple à Louis Rougier (un autre électron libre de droite), remarquable en économie et en philosophie[3]. Le problème de Benda c'est qu'il est un homme de lettres, un homme de cabinet, à une époque où Malraux, Gide, Leiris, Hemingway, Orwell ou s'engagent dans des résistances militaires au totalitarisme ou visitent les lieux d'internement et d'exploitation coloniale. On peut, même quand on est Benda, critiquer certes le style de Gide, mais ce dernier a été en Afrique, en URSS, et a été chroniqueur de procès d'assise (sans parler de sa défense de l'homosexualité dans *Corydon*). Au risque de choquer je dirais qu'un écrivain n'a pas à soutenir *ex professo* une théorie morale et politique correcte et à s'y conformer exactement. Si c'était le cas, on condamnerait peut-être par exemple Renaud Camus et Richard Millet, sans lire bien sûr leur œuvre : il suffit de les accuser d'être réactionnaires, alors que leur œuvre émerge et domine la bien pensance littéraire de leurs négligeables accusateurs.

Le livre de Pascal Engel contient deux versants, ou deux parties distincts. Ce qui concerne le combat de Benda contre le bergsonisme permet d'apporter des pièces importantes au procès contre la philosophie française des années sombres. Les légendes rassurantes sur Politzer, Nizan en sortent écornées et c'est justice. Le caractère spiritualiste de cette philosophie (avec la réaction générale contre le positivisme de Comte et Taine, notamment du côté de l'inévitable Bergson) est dénoncé efficacement. En ce qui concerne Bergson lui même, Benda, quoiqu'injuste, détruit le mythe du grand philosophe français. Cependant ce qui concerne le second versant, la politique et la morale ne permet pas, c'est mon sentiment de sauver complètement Benda. Je pense que la *France Byzantine* est un livre admirable et son côté ultra réactionnaire ne gêne plus, grâce en partie à Antoine Compagnon : quand on a lu ces deux admirables écrivains que sont Léon Bloy et Joseph de Maistre on ne s'effarouche pas facilement devant les excès rhétoriques de Benda, mais je ne trouve pas l'équivalent chez Benda dans le domaine politico-moral de qui est si impor-

[3] Louis Rougier à la différence de Benda était au courant de la pensée de Bertrand Russell (cf. son livre *Le langage et la métaphysique*), capable d'intervenir dans le débat sur le néo-thomisme (cf ; son *Scolastique et Thomisme*), parfaitement informé de la mécanique quantique. En économie, il a rejoint la Société du Mont Pélerin (où s'illustrèrent Maurice Allais, Hayek, Bertrand de Jouvenel, Popper, Polanyi, Von Mises ….).

tant dans le domaine de la polémique contre les Grandes Têtes Molles de notre Temps. *La France Byzantine* est un livre efficace (quoique profondément faux dans les détails) par ce qu'il fait le travail d'une lecture fouillée des doctrines, des styles, mais il n'y a pas l'équivalent chez lui sur la fascination de la France pour le totalitarisme, brun ou rouge, dans les années 30 – il faudra attendre Aron, Furet, et Castoriadis pour décrypter cet aspect complémentaire de l'enfermement hexagonal.

Le livre d'Engel sur Benda c'est Engel se cachant derrière Benda (comme l'a remarqué Roger Pouivet dans un récent compte-rendu). On se demande périodiquement à la lecture qu'est ce qui a pu motiver un philosophe de la connaissance d'une part dans une telle entreprise, au carrefour de la littérature, de la morale et de la politique et d'autre part pourquoi ce choix de Benda. Simone Weil, Bernanos, Péguy, étaient tout de même plus intéressants historiquement et littérairement (mais ils sont marqués par le sceau désormais infâme du religieux, quoique Péguy et Simone Weil eussent refusé les sacrements). Les écrits de combat de Bernanos offrent plus de matière et ils présentent de magnifiques *changements de cap*, de *la Grande peur des Bien pensants* (1931) aux *Cimetières sous la Lune* (1938). La prise de parti pour les républicains espagnols d'un ancien de l'action française est tout de même plus passionnante que les disputes internes à Gallimard d'un homme de lettres. Il y a donc un mystère Benda dans la pensée de Pascal Engel. Toutefois ce mystère se dissipe si l'on note l'aspect normatif de la pensée de Benda qui s'accorde bien avec le projet de Pascal Engel de dégager les normes de la connaissance littéraire. Tout cela ne rend pas ce livre moins passionnant, moins important, mais c'est Engel qu'au final nous apprécions, par son aspect de procureur incorruptible et courtois, érudit et implacable – et pas toujours Benda dont l'unité de grand bourgeois communiste échappe quelquefois.

8

Can we solve the paradox of fiction by laughing at it?

CAROLA BARBERO

I have known Pascal Engel for several years and I have always appreciated, in addition to his bright intelligence, his strong sense of humor and his subtle irony. But it was only in May 2010, when I was invited by Pascal to give a lecture in his course on the philosophy of laughter,[1] that I had the opportunity to discuss with him laughter (in all its variants) and philosophy, obtaining useful suggestions for my research – and laughing a great deal.

My starting point consisted in the emotions we feel when dealing with a work of fiction, in this particular case the laughter that some literary or cinematographic works evoke in us. The laughter-fiction relationship seemed to me (and still seems) interesting enough to push me to go to Geneva to talk with Pascal Engel and his students. Why? Because the comic, unlike what happens with tragic works or scary ones, apparently presents no issues. This is what I found fascinating and in need of further study.

Let's think about what happens when we are told a joke. We laugh, and that's it. No one would ever think of asking us "why are you laughing?" or "are you laughing for real?". Unlike other emotions (like sadness and fear),

[1] The lecture, given at Uni Bastions, Université de Genève on May, 4th 2010, was entitled: "L'humour et le paradoxe de la fiction".

in fact, laughter does not seem to establish any paradox[2] relatively to the fictional stories we call jokes, although they clearly are not real events or objects.[3] But if we *really* laugh for something we know is fake, then it is not true that, in order to feel emotions toward an object, we must believe in its existence. Thus the paradox vanishes.

However, it is legitimate to wonder why we have questioned for so long the authenticity and rationality of emotions directed at objects that arouse fear or pity, while instantly recognizing the legitimacy and authenticity of the emotions we feel for the fictional objects and events that make us laugh. Perhaps the reason is simply that, since fear and compassion are negative emotions, and therefore have a *high cost*, we tend to dispense them at our discretion in those situations in which it seems to be actually worth it (i.e. in real situations). Instead, as laughter is always a source of income, we accept it in all its forms (whether it relates to real objects or fictional ones).

From this, we may conclude then that while fear and tears are authentic or justified only when caused by real objects, laughter is always true, regardless of the type of object causing it. But this argument is unacceptable: if we admit that the type of object is crucial to determine the authenticity of the emotions it arises, then we cannot make a distinction according to the type of emotion. Either only real objects can cause genuine emotions – so that both a melodramatic novel and a joke cause false ones – or all kinds of objects (real, fictional, past, dreamed, etc.) can arouse in us authentic emotions (which, of course, vary depending on the type of objects to which they are addressed). The topic of laughter clearly invites us to choose the second option.

Let's briefly recall the subject matter and see to what extent it can be characterized as a good answer to the paradox of fiction. It is a simple *modus ponens*: if we really laugh when we are told a joke, it is not true that, in order to feel authentic emotions, we must believe in the existence of what we are told (as no one believes that jokes are true stories); when we are told a joke we really laugh, therefore it is not true that in order to feel authentic emotions we must believe in the existence of what we are told.

Moral of the story: if instead of considering *Anna Karenina* we had focused on any one joke, it probably would have taken much less to find a solution to the paradox of fiction. Take the following joke:

[2]Here the reference is the famous paradox of fiction, placed at the center of philosophical debate since the publication of the article by Radford (1975).

[3]On the importance of laughter and jokes in order to demonstrate the absurdity of the paradoxes arising in relation to fiction, see Ferraris (2009, 77): "Jokes are the shining example of laughter that is completely independent of the truth or falsity of the things described".

> A man walks into a pet store and asks to see the parrots. The store owner shows him two beautiful ones out on the floor: "This one is $ 5,000 and the other $10,000," he says.
> "Wow!" says the man. "What does the $5,000 one do?"
> "This parrot can sing every aria Mozart wrote," says the store owner.
> "And the other?"
> "He sings Wagner's entire Ring cycle. There's another parrot out back for $30,000."
> "Holy moley! What does he do?".
> "Nothing that I've heard, but the other two call him 'Maestro"'[4]

We laugh without thinking about the legitimacy or rationality of doing so (or better, if we ask ourselves if our laughter is legitimate, we will probably laugh even more). This clearly highlights how, in order to feel authentic emotions, it is sufficient to have an object toward which they are directed, without this necessarily being a real object. Of course reality can be full of very real anecdotes that make us laugh and cry, but this does not mean that fiction is unable to elicit authentic emotions. It simply means that reality, understandably, has its share in provoking an emotional response in us.

But what is it that makes us laugh at a joke? What, exactly, is the object or event that makes us laugh? Obviously much depends on the skill of the person who tells the joke, her ability to involve us building a well-structured story, with the necessary pauses, gestures, looks and everything else. Let us assume that our narrator is very good. What's funny about the story of a man who goes into a store to ask about parrots? First of all, there's nothing funny and this already augurs well. The guy enquires about the prices of the birds and the reasons given by the trader to justify them are most striking. But the argument advanced in favor of the most expensive one is the spring that triggers the laughter: it is a fallacy of relevance, more precisely a fallacy *ad auctoritatem* which is an invalid argument in which a thesis is accepted only on the basis of the (alleged) prestige of those who propose it.

There is nothing wrong in invoking the authority of an expert, but it is wrong to use the respect for such authority as the *sole* evidence in support of a

[4] Although it is widely accepted that jokes do not have an author in the proper sense, but are rather just *discovered* – as claimed by Ferraris (2009, 77): "[...] just like myths, jokes do not have authors" – we would like to report the text from which we took the joke because it is smart and funny, managing to set out the main issues and themes of philosophy through jokes and paradoxes. It is Cathcart and Klein (2007, 44).

thesis. Why has the trader decided that the third parrot had to be the most expensive? Because he listened to the other two parrots. What makes us laugh is the fact that the price has not been decided on the basis of some characteristic of the parrot, but only by making reference to the fact that his fellow parrots, already very gifted and talented, call him "Maestro". It makes us laugh because, obviously, this is not a good reason to justify $30,000 (without thereby arguing that the less expensive parrots are liars). An important element that characterizes this type of fallacy (as the joke makes clear) is that often the personality to which reference is made so as to justify the validity of an argument does not seem to be a legitimate authority.[5] That's why the reasoning of the trader makes us laugh.

So we laugh because of the final answer of the trader to the customer, even though we know perfectly well that neither the former nor the latter exist – let alone the parrots. If we were to outline what happens to us, we could propose something like this, which yet would seem absurd:

> X laughs for the answer of the trader and X knows perfectly well that the trader is a fictitious entity;
> Believing in the existence of what makes us laugh is a necessary condition for the corresponding emotion;
> X does not believe in the existence of fictitious entities.

Such scheme seems absurd because whether the trader exists or not is absolutely irrelevant with regard to the authenticity of the emotions we feel. It would obviously be different if the client were our father and the joke, far from being a joke, was a true story: our father could be the customer entering the store and being fooled by the trader to pay 30,000 dollars for a mute parrot. If, after being *robbed* by the trader, our father told us this story, we would

[5] It is not a coincidence that this type of fallacy frequently occurs in commercials where the only guarantee of the quality of a product is the celebrity spokesperson. Here, of course, it all depends on the type of product you want to advertise and the relevance of the authority you choose. Models are often chosen to advertise beauty products, sportsmen for health products, "beautiful and damned" actors for spirits, etc. and the reasons for these choices are obvious. A model, for example, guarantees for cosmetics and moisturizers because, being beautiful, she is also supposed to have the authority to pass judgment on the validity of these products. The point is that it is unclear what it means to be the most reliable authority to justify the conclusion that has been reached or we want to reach. On what grounds should I believe that this brand of products is valid? Because I am told so by a beautiful model: X is true because P tells me so. But does P really know something of cosmetics, or has she merely been paid to ensure, with her image, the quality of a series of products she knows nothing about? That is the question on which the fallacy *ad auctoritatem* is based.

not laugh so much (although we would still laugh a little bit, we must admit it: maybe we would refrain from doing so simply because it is not very nice to laugh at the misfortunes of others) and we would sue the trader for taking advantage of him.

We laugh so heartily *because* we know it is a joke telling the story of fictional objects and events: their status as fictitious objects, far from making our emotions less authentic, explains and justifies them.[6] Laughter, and more generally comedy, thus resolves the paradox of fiction and demonstrates its groundlessness.

1. Laughter as a solution to the paradox of fiction

Why did not we think of that before? Why do we concentrate on negative emotions, asking whether or not they are authentic when not directed toward existing objects, when it would be enough to have a laugh to make all paradoxes vanish? In fact, it would have been enough to think of a hypothetical paradox of fiction based on comedy to figure out where to find the solution: just as it is not necessary to believe in the existence of what makes us laugh in order to laugh out for real, so it is not necessary to believe in the existence of what makes us cry in order to cry our heart out. But if the paradox of fiction has no reason to exist as regards comedy, then it is unclear why it would still stand as regards tragedy. And, as we have seen, it would not be a good argument to claim that the paradox of fiction has the right to exist only in relation to tragic works (and this regardless of the fact that the first book of Aristotle's *Poetics*, on tragedy, did not go astray unlike that on comedy).

In fact, the problem with the emotions we feel for the non-existing characters of novels or films seems to emerge if and only if we are talking about the so-called "negative emotions". Why cry for someone who does not exist? Why be afraid of a vicious murderer who only exists in fiction? On the other hand, though, when we are told a joke or watch a movie with Mr. Bean, we laugh without questioning the authenticity of the emotions we feel. In the case of comedy apparently no problem arises, although we know perfectly well that even in that case our emotions are not directed toward objects that exist in the world of space and time. But why should we doubt the sincerity of the tears we shed for Anna Karenina while not doubting at all the authenticity of the laughter aroused by Mr. Bean?

[6] On true emotions we feel for fictitious objects I refer the reader to Barbero (2013), pp. 45-58.

We could answer this question by arguing that jokes or comedies do not really cause emotions, but merely states of mind or moods.[7] The difference between laughter (for Mr. Bean) and sadness (for Anna Karenina) lies in the fact that in the latter there supposedly is what we might identify as a cognitive component, in virtue of which we say that our emotions are directed toward an object, wonder whether it is reasonable to feel emotions for objects that do not exist and ask ourselves if these objects could move us to act or behave in certain ways. It is allegedly under this cognitive component that the paradox occurs in cases of sadness and fear felt for fictitious objects but not in cases in which such objects arouse laughter and joy. This explains, in theory, the reason why there is a paradox of tragedy but not a paradox of comedy[8] relatively to the existence of fictitious objects.

It might seem like a good solution, but it is not, since it is based on the highly questionable assumption that laughter and happiness are states of mind devoid of cognitive content. What does it mean "to be devoid of cognitive content"? Does this mean that when I laugh at Mr. Bean it is a bit as if I was in a state of euphoria (while when I cry for Anna Karenina, there are characteristics of Anna and the events she is involved in that make me sad)? It really seems implausible. Suffice it to say that if we see a person who laughs out loud on the couch and ask her "why are you laughing?", she could answer us "for no reason" (meaning "my laughter does not have a cognitive content") - and then we would rightly think she is euphoric (just as we think that those who cry for no reason are depressed). But if she answers that she is laughing at a Mr. Bean gag, then we will probably think that there is a reason (i.e. an object) for which she laughs: Mr. Bean, in fact. It is therefore not possible to make a distinction between tears and laughter for fictional objects by simply referring to the cognitive content supposedly possessed by the first, but not the second. In fact, as we have seen, laughter also has a specific cognitive content. The person who laughs at the scene where Mr. Bean tries to dive from the trampoline is neither euphoric nor generally happy: she is laughing because she just saw a funny scene with a guy making a thousand grimaces and trying to dive off a diving board.[9]

[7] The position that laughter is not exactly an emotion but a simple state of mind was defended by Stuart Brock during a series of conversations with him about these topics.

[8] Later we will see how another paradox can be found in comedy. It does not regard the status of the fictional objects our emotions are directed to (which, as we have seen, is not a problem), but the circumstances for which in comedies, in general, we laugh at the misfortunes of others (which, in normal life, we usually do not do).

[9] *The Curse of Mr. Bean*: http://www.youtube.com/watch?v=K_bX_jX9O8w.

There is no paradox of fiction in the case of comedy, not so much because laughter and happiness are not emotions, but because they are not negative emotions. Is the *cost* of the emotions that determines the level of ontological concern: this is why sadness and fear raise many issues, while laughter does not (let alone giving rise to paradoxes). If we have to *pay* personally by crying or being scared, we want to know why exactly we despair, while trying to figure out what it means to pity or fear a fictional character and how this is different from feelings we have for real people and situations. Instead, for those emotions that only bring advantages (like the good humor and joy that comedies often give us), we do not bother asking questions and just enjoy them. However, seeing a paradox only where it suits us is never a good move, especially if we are interested in taking into account the issue *parte objecti:* we deal with fictional objects both in comedy and in tragedy, so either we admit that in both cases our emotions are genuine as directed to those objects, or we refuse to regard what we feel in those cases as emotions in the true sense of the word.

I am committed to defending the position that the emotions we feel for fictitious objects are authentic and rational both in the case of tragedies and in the case of comedies (because in both cases we are dealing with fictitious objects).[10] The theory of the object identifies an object (a fictitious object, be it Anna Karenina or Mr. Bean) as the cause of a specific emotion (sadness or happiness), thus enabling us to dissolve the paradox of fiction.[11]

An emotion, to be authentic and rational, merely needs to be focused on an object (and not, as the fictionalists obstinately assert, an *existing* one).[12] When we laugh at Mr. Bean all we need is to believe that there is an object with certain characteristics involved in events such as to provoke in us emotions like enjoyment and happiness. With these assumptions, it is clear that the paradox does not arise: we believe Mr. Bean is ridiculous for some of his features, but we do not believe that Mr. Bean actually exists (meaning the character, of course, because the actor Rowan Atkinson exists in all respects).

Another possible objection to the idea that the emotions we feel for comedies can be a proof of the groundlessness of the paradox of fiction might consist in pointing out that laughter arises no paradoxes for the simple reason that it is not a serious thing. After all, one does not laugh that often (only children and madmen do it on a frequent basis) and above all it is never really

[10] Barbero (2010).
[11] See Meinong (1904).
[12] See Walton (1978, 1990, 1997).

clear what there is to laugh about. *Risus abundat in ore stultorum*, said those who believed that the outward manifestations of joy and laughter, as well as the body with all its demands, should be silenced so as not to harm the soul and the spiritual dimension of individuals in general. How can we forget that in *The Name of Rose*,[13] Jorge of Burgos commits the most atrocious crimes precisely to keep the last surviving copy of the second book of Aristotle's *Poetics* on laughter and comedy hidden? Of course, there laughter and comedy were condemned only because, if indeed, as Aristotle argued, it was possible to laugh at everything, then there was a risk that one could even laugh at God. It was therefore a moral condemnation of laughter: if you can laugh at everything, then there is nothing absolute, and everything depends on individual choices. However, it is not the moral side of emotions that is discussed here, but the purely ontological side. In this respect, laughing because of a comedy is an emotion just like crying for a tragedy. From an ontological point of view, in fact, laughing at Mr. Bean is not significantly different from crying for Anna Karenina: in both cases we have an object capable of arousing certain emotions in us.

So the fact that laughter has been regarded as a manifestation of the devil (as suggested by the *doctor Mellifluus* Bernard Clairvaux), a sign of stupidity, a loss of control or rationality and so on,[14] is not important for us here, as we are interested in the object of laughter and not laughter itself. Another interesting case is that of the laughing object, which could be seen as a sort of mid-point: what about a laughing statue? The starting point is offered by the famous film *Scusate il ritardo*,[15] in which Massimo Troisi explains why the real miracle would be a Madonna laughing (and not crying). Mind you, an inanimate being (object) such as the statue of the Madonna weeping is already quite a miracle, but the idea that it could possibly laugh would make it – as Troisi says in the movie – much more miraculous.

Why? For three reasons: first, because it is more difficult to *pretend* to laugh than to cry (try to pretend to laugh, if you are not a professional actor it will be really hard, while you can easily pretend to be sad by looking *down*, talking little, etc.). Secondly, because laughter requires more facial changes than crying (which, at most, requires a few tears in the eyes). Finally, it would be a super miracle because while it is assumed that the Madonna may have many reasons to cry (basically the evils of the world and the wickedness of

[13] U. Eco, *The Name of the Rose*, Boston, Houghton Miffin Harcourt, 1983.

[14] For an interesting history of laughter, see Minois (2000).

[15] M. Troisi *Scusate il ritardo*, with M. Troisi, G. De Sio, L. Arena (Italy, 1982).

men), we do not believe that she has that many reasons to laugh (in fact why would she laugh? because she finds us funny? because our lives are more ridiculous than a hilarious joke? Because we are like her and her son, the only difference being that we do not have a fast track to reach the Father?). However, in both cases (of the Virgin laughing and weeping) it would be a full-blown miracle, since inanimate objects, as is well-known, do not have emotions. Nevertheless, if they could, we might assume that they would have emotions in response to (or even just awareness of) those we feel for them, as it normally happens in relations between human beings.[16]

The Madonna crying (or laughing) would be a curious phenomenon to be tackled because of its intermediate status between the object of the emotion and the subject feeling it, and yet for many reasons, not least that of common sense (which, especially in philosophy, always comes in handy) I'll discuss it no further. Let us return then to our viewer, who laughs heartily at a Mr. Bean gag, or to refer to a classic of comedy, let's say she is watching a movie of Laurel and Hardy. What is she laughing at? What's so funny in a man breaking through the floor with a simple hop and ending up downstairs?

2. The concept of "humor"

In order to understand what's funny about what makes us laugh, it is necessary to dwell on the concept of "humor". What is the basis of humor? Why do we find something funny or entertaining? What does it mean to say that something makes us laugh? Is there a definition of comical? The question can be tackled from two different points of view: *parte objecti* and *parte subjecti*, because it is one thing to ask what features an object must have in order to be funny, but asking why a person finds something funny or amusing is another thing. However, it is clear that these two distinct levels affect each other. In fact, it often happens that something is funny because there are users that, under certain conditions and in the appropriate context, find it such. The context of use and the awareness of the object to which we address our emotions are basic elements: we can find the features of an object funny only if we believe

[16] In this sense Ferraris (2007: 195-196) speaks of works of art as automatic sweethearts, works that pretend to be people: "Thus we account for the specific form of illusion that is common to all forms of art. [...] Artworks are things that pretend to be people, i.e. automatic sweethearts. What do I mean by this? [...] In works, as well as in the Automatic Sweetheart, we are dealing with physical objects that are also social objects, and yet [...] arouse feelings, just as people do when we consider them as such and not as simple functions – except that, unlike people, they do not expect nor offer any kind of reciprocity."

that the object is fictitious[17] and, likewise, we can find something funny only if we experience it in an appropriate context.[18]

But what, exactly, makes us laugh at a given object or event? Philosophy has basically given three possible answers: one referring to the absurdity that characterizes some objects and events, one referring to the superiority that the viewer feels towards what makes her laugh, and one insisting on a sense of relief that the object arouses in the viewers. These responses have been formulated in many different ways by philosophers,[19] but we will address them very generally by referring to the *theory of absurdity*, the *theory of superiority* and the *theory of relief*.[20]

Theory of absurdity. According to this theory, absurdity can be perceived both within the comic element itself and between the world of fiction and the real world. It is a position that has noble origins and that can be traced back to Kant: "In everything that is to excite a lively convulsive laugh there must be something absurd (in which the understanding, therefore, can find no satisfaction). Laughter is an affection arising from the sudden transformation of a strained expectation into nothing."[21] This theory finds the essence of what makes us laugh in the lack of compliance with certain laws (logical, moral, etc.) or even with our expectations. This absurdity, however, must be de-

[17] Just think of the following joke, which makes us laugh only if we believe that it is just a joke and not a news story on yet another tragic plane accident: "An Italian, a Frenchman and a German are on a plane that is plummeting. There are only two parachutes. The Italian begins to cry saying that he has seven children, a wife, elderly parents without pension and if he dies it is as if they all died, then prays Santa Rosalia throwing himself on the ground, writhing and crying like a baby. Then he gets up, takes a parachute and jumps off, leaving the other two with the simple phrase 'forgive me, but I have to save myself'. Then the Frenchman asks the German: 'what do we do now?'. The guy calmly opens a bottle of beer and replies: 'No worries, the Italian jumped off with a rucksack'. "

[18] For example, if we watched a comedy in the dentist's waiting room we would certainly enjoy it less than if we watched it at home with a couple of friends, comfortably sat in our armchair.

[19] For a critical presentation of the main theories that, from Aristotle to the present, have tried to explain the phenomenon of the comic, see Morreal (2009a).

[20] See Levinson (2006: 390-394). Obviously the theories classified here are the result of the simplification of different philosophical positions. It is also evident that, with deeper explanations, some philosophers might be seen as defenders of a theory different from that which I here attribute to them. For example, Kant could be seen both as a supporter of the theory of relief and as an advocate of the theory of absurdity. In fact, he insists on both the sense of pleasure that invades the viewer when he understands that what he thinks is going to happen will not happen, and on the perception of something absurd in the object that causes us to laugh. Similarly to Kant, many authors mentioned herein may be brought under one or the other theory. Therefore this classification does not claim to be exhaustive and is presented with the sole purpose of broadly exposing the main theories on the essence of the comic.

[21] Kant (1790), First Part, sec. 54.

tected in the object by a subject whose rationality is firm: this explains the link between laughter and intelligence, which clearly highlights the reason why only humans (the rational beings *par excellence*) are able to laugh in the proper sense.

Actually, that funny objects or events should have something absurd about them seems to be neither a necessary (there may be funny jokes that have nothing absurd about them) nor a sufficient condition (many absurd situations do not elicit laughter). One could even argue that, in any case, the funniest jokes are the ones that do have some element of absurdity, but even if it were so, the absurdity would be at best a necessary condition (and therefore fall within the definitions of humor and comical without exhausting them).[22]

Then what else do we need to laugh? The absurdity should not simply be detected, but it would also be important that it was appreciated by itself, giving rise to no negative emotions (as in the above-mentioned case in which the joke was about our poor father being fooled by the parrot-seller) and having no potentially harmful practical implications.[23] In any case, the characteristic of absurdity does not seem to cover all the cases[24] that we would be willing to place under the category of "comical",[25] and therefore we should perhaps look elsewhere.

Theory of superiority. It is a theory that goes back to Hobbes who, in his *Treatise on Human Nature*[26] and in *Leviathan*,[27] explains the reasons for laughter by

[22] Martin (1983), pp. 74–84.

[23] A. KOESTLER (1964): 27-63. Here we also find a possible formulation of the paradox of laughter, although a substantially different one from that proposed here in § 3.

[24] Also, it does not explain the interesting circumstance for which we also laugh the second time: if it were only a matter of perceiving the absurdity, then in theory we should just laugh the first time, when we detect the absurdity and are surprised by it. Instead, it often happens that we laugh several times (when, presumably, the element of surprise is gone) for the same joke or comedy.

[25] Also, even when we find the absurdity, it is not always clear that this is the main reason why we laugh (for example, this does not explain why there are some people elected to become the protagonists of some jokes), and in fact much also depends on the attitude of the user. One way to save this feature from the many objections is to argue that the absurdity must be a characteristic not so much of the content, but of the structure of what we find funny. See Lipitt (1992).

[26] "The passion of laughter is nothing else but a sudden glory arising from sudden conception of some eminency in ourselves, by comparison with the infirmities of others, or with our own formerly" Hobbes (1650), Ch. 9.

[27] In *Leviathan*, laughter is seen as a typical manifestation of the weak, who constantly need to be compared to people below them so as to be reassured about their value: "Sudden glory, is the passion which makes those grimaces called laughter [...] And it is incident most to them, that are conscious of the fewest abilities in themselves; who are forced to keep themselves in their own favor by observing the imperfections of other men. And therefore much laughter at the defects of

referring to the sudden awareness of the user's own superiority, which puts him in a position of strength. In this definition, Hobbes assimilates the positions of Plato[28] and Aristotle,[29] for whom laughter was just a strange combination of pleasure and malice aroused in the viewer by someone who believes to be better than he actually is and, in any case, is worse than the viewer.

Laughter is thus an emotion of pleasure mixed with pain because in fact the user laughs at the ignorance, flaws or misfortunes of others, thereby proving to be petty and mean. To this perspective also belongs the view proposed by Bergson,[30] who dwells much on the great social power of comedy: far from being a mere expression of pettiness, comedy is used by people to criticize society or deviant behaviors, with the aim of stigmatizing and/or defending the conduct of society's members. This also explains why we better enjoy comedy together: it is a social phenomenon that can be fully understood, accepted and enjoyed only in a social context. Laughter is aroused by the perception of certain characteristics that turn any object into a caricature: a sort of inauthentic object that can elicit laughter and derision in us. The viewer feels superior to such an imperfect object, and this acknowledgment of others' imperfections (it is not by chance that we always laugh at people or humanized objects) provokes a kind of pleasure that is naturally manifested in laughter (the comic, thus, often has a *Schadenfreude* victim).

Although this theory is also very convincing, it seems clear that the feeling of superiority cannot be identified neither as a sufficient condition (not all feelings of superiority can be found in our emotional responses to the comic), nor as a necessary condition of the comic (because we might find a joke funny by itself or we may experience feelings other than superiority). Also one could – thus rejecting the theory of superiority – not share the basic assumption of this position, which is that the essence of the comic does not reside in the object judged comical or funny, but in the person considering it such. It appears that the foundation of this theory is some sort of confusion between the genesis and the structure of the comic: it is one thing to speak of the mechanism that is activated in the users causing them to find a particular object funny, but to identify the characteristics that make an object funny is another thing (and between the two levels there must not necessarily be a relationship of depen-

others, is a sign of pusillanimity. For of great minds, one of the proper works is, to help and free others from scorn; and to compare themselves only with the most able." T. Hobbes (1651), part 1, ch.6.

[28] Plato, *Republic*, Book III, 389, and *Philebus*, 48-50.

[29] Aristotle, *Poetics*, 1449a, 33-38, *Nicomachean Ethics* , 1127b-1128b.

[30] Bergson (1900).

dence or emanation, as the theory of superiority seems to take for granted)[31].

Furthermore, one could object that, even if there is a feeling of superiority in the fruition of the comic, it is very different to laugh at someone ridiculing and almost despising them, and to merely poke fun at someone. Last but not least, this theory makes it difficult to explain the widespread phenomenon of self-irony: my present I does not always laugh at an earlier I (and therefore it does not always laugh at *someone else*), but it often happens that we laugh at what we are *now* (and smile, perhaps, at what we used to be). So let's look at the third and final theory on the comic and its essential characteristics.

Theory of relief. This is the theory[32] according to which the comic relieves tension (like a safety valve) by breaking the rules (social, moral, logical, and even plain common sense – such as the rule to be serious) and momentarily releasing the users from their grip. The main proponents of this theory are Spencer[33] and Freud[34] who see the essence of the comic in the ability to *free* people from constraints allowing them to vent (for a short period of time) their pent-up energy.

Freud interpreted reactions to comedy in the light of his theory of consciousness and the unconscious: the fruition of the comic is important because it allows for the fulfillment of the drives linked to aggression and sexuality, which are usually repressed. Freud, like Bergson, also notes the social dimension of the comic, claiming that jokes and witticisms require the presence of at least two people to have the desired effect (at least, in fact, one tells the joke and the other laughs). According to Freud, there are two main types of jokes: the innocent – the typical serene laugh after a good joke – and the interested one, i.e. the laughter produced by the pleasure derived from having vented aggressive or sexual energies.

In general, the theory of relief detects the essence of the comic in its effects: the comic is what frees us from the constraints of life, taking away inhibitions and allowing us to unleash our pent-up energy. However, it seems that this position does not work either: referring to the unleashing of repressed energy helps us understand what happens when we have fun, but still it does not tell

[31]Or at least it is taken for granted by the classic presentations of the theory (see Morreall (1998), pp. 401-405; Levinson (1998), pp. 562-567, for which in fact the superiority felt by the public is what properly constitutes the essence of the comic.

[32]For a good presentation of the theory of relief, see Morreall (2009b).

[33]Spencer (1860). In addition to identifying the essence of the comic, Spencer is also interested in understanding why it provokes the outward manifestation of laughter, venturing in the search of a physiological explanation.

[34]Freud (1905).

us anything about why this happens. In fact, the position that seems to be more suited to play the role of the general theory of the comic is the theory of absurdity, since neither the theory of superiority nor that of relief seem to have sufficiently broad a scope to play such a role, and also seem to be more focused on users and on the mechanisms that trigger the enjoyment of the comic than on the object as such.[35]

3. Humor and horror

Let us now move on from the theories that try to explain the essence of humor and focus on a matter concerning fruition *parte subjecti*. For example, think about what happens when we watch a comedy: we normally laugh out loud, sometimes even to tears. How come? A possible answer is that we laugh because what we are watching is very funny. But is it true that *what* we see is really *funny*? Take for example *County Hospital*,[36] the film in which Laurel and Hardy go through an odyssey (as they always do): Hardy, poor fellow, was hospitalized with a broken leg. He's visited by Laurel who, because of a stupid accident, nearly kills the doctor getting Hardy early discharged. Then Laurel, to make it up to Hardy, decides to give his friend a lift home but, without knowing it, he is under the influence of a sleep-inducing medicine that he has inadvertently taken in the hospital by sitting on a syringe. So, barely able to keep his eyes open, he causes a new serious accident in which also his friend is involved. In the last scene, when they crash, the audience usually laughs like crazy.

Perhaps it would be worth asking *what* it is we laugh at. Is it funny to see a friend go to the hospital? Is it funny to risk killing him while driving him home because you are falling asleep? No, in fact, put it this way the story does not seem funny at all, and yet when we watch the film we just cannot help but cry with laughter. This is the *paradox of comedy*.

Why do we laugh at things that, if they occurred in real life, would make us sad or at least worried? How is it that we feel a pleasure so great that it turns into laughter when watching or reading about the misfortunes of others? Terrible accidents happen, people risk dying, floors collapse, cars crash, and we laugh. In order to bring out the real paradox, the questions to be considered are the following: 1) why do we seek in comedies what in everyday

[35] Levinson (2006: 393).

[36] J. PARROTT, *County Hospital*, with S. Laurel and O. Hardy (USA, 1932).

life we strive to avoid (and which, if it happened, would arouse anything but laughter)? 2) how can we laugh at the misfortunes of others?

First, let us ask whether it is contradictory that, in comical works of fiction, we look for what in real life we try to avoid (vases on the head, pianos on the feet, destroyed houses, etc.). As much as this is a strange behavior, we cannot really call it contradictory, since it is not contradictory to search in fiction for objects and events that we would rather avoid in real life. From this point of view, the conflict is only apparent.

The issue raised by the second question is more interesting: how can we find the misfortunes of others funny? How is it possible to be aware of the seriousness of what is happening to our characters and still have fun seeing their misfortunes? There seems to be a real conceptual impossibility: if it is true that we are aware of their misfortunes, then it is unclear how we could laugh at them. This is the paradox of comedy, which it consists of three theses that are individually plausible but, if taken together, contradict one another:

> 1) Laughter is the manifestation of a positive emotion experienced by the user;
>
> 2) The characters of comedies often undergo misfortunes of which the user is fully aware;
>
> 3) The user of comedies laughs and enjoys herself.

The paradox dissolves when we recognize that when we enjoy comedies, our entertainment is not addressed directly (or mainly) to the characters undergoing all those disasters and catastrophic events, but to the narrative *structure* and *style* of composition of the work. Not surprisingly, if the style of the play is poor we do not laugh out loud, but we die of boredom, or worse, begin to suffer along with our hapless characters.[37] What is crucial is then how objects and events are presented, what role they play within the broader narrative structure and how the misfortunes described are part of the whole. That is why it is substantially misleading to ask what is the reason why we laugh at all those disasters: the fun we have, in fact, is simply a *function* of the way in which the object is presented within the work as a whole.

Resume the initial question: what is it that makes us laugh in the comedy of Laurel and Hardy? They destroy everything, everything always goes wrong, and yet they make us laugh out loud. It is not so much a question, as Aristotle claimed, of laughing at those who are worse than us (because in

[37] Exactly the same mechanism well described by Hume (1757).

this case the paradox of comedy would still stand), but above all, as Hume observed, of appreciating the way in which their stories are constructed and articulated, of recognizing the appropriateness of the chosen style and language. This perfection is what awakens laughter within us, not the misfortunes the protagonists are involved in.

That's why the vision of *County Hospital* keeps us entertained so much: because it expresses a surreal comedy where, according to the classical scheme of comedy, an insignificant episode (Laurel tries to crack a nut with a counterweight that keeps Hardy's leg raised) is the cause of a series of disasters (Hardy finds himself upside down with the broken leg in the air and the doctor is thrown out the window threatening to fall out). Then, according to the classical scheme, Hardy is the one to pay for the consequences of his friend's actions (and in fact is thrown out of the hospital), and Laurel, who is the naïve character *par excellence*, is (as always) amazed by what happened.

Therefore, it is not at misfortunes that we laugh, but at the way in which they are presented. In fact, if the same misfortunes were presented differently or were real, then we would be likely not to laugh at all. Likewise, if the person telling a joke is very good, hearing him talk in a certain way and seeing him make certain gestures will probably suffice to make us laugh, at least initially, regardless of the content of the joke itself. This is the reason why we usually laugh more at jokes than at life: not so much because life is sad, but because jokes are built and told better (on the other hand, when would we ever get an Italian, a German and a Frenchman having to jump off a plane with a parachute?).

A film genre that well illustrates how the structure and style of the narrative are what causes us to laugh *almost independently* from the content are parodies: what about a work that has substantially the same content as another but, through a completely different narrative style and language, has a diametrically opposite effect to that elicited by the original? Think of *Young Frankenstein*[38] or *Repossessed*,[39] which are respectively the parodies[40] of *Franken-*

[38] M. *Brooks*, *Young Frankenstein*, with G. Wilder, M. Feldman, P. Boyle (USA, 1974).

[39] *B. Logan, Repossessed*, with L. Blair, N. Beatty, L. Nielsen (USA, 1990).

[40] I am only reporting here examples of film parodies. There are some interesting parodies of literary works too, but I will not take them into consideration since often, behind the parody, they express opinions critical of culture and society, so that they tend to be much more complex than the simple parodies of movies. This is the reason why in literary works it is often very difficult to distinguish clearly the genre of parody from that of satire. Suffice it to say that the literary parody *par excellence* is *Animal Farm* (G. Orwell, Secker and Warburg, London, 1945), which in fact is a satire in which, behind the history of the revolt of the animals in an English farm, lies an allegory of Soviet communism.

stein[41] and *The Exorcist*,[42] where the content is in many ways equal to that of the originals. And yet, they elicit an opposite emotional response in viewers:[43] not fear but fun and laughter. The parody is based on the idea of using elements of an existing model and taking them to the absurd, thus inducing the viewer to laugh at things and events that otherwise would make him scream with fear.

This passage from fear to laughter that the style (of parodies, in this case) is able to operate is made possible by two different orders of factors concerning respectively the *pars objecti* and the *pars subjecti* of the fruition. On the one hand, there is an intimate relationship between horror and humor because the fictional objects and events presented by both genres are characterized by similar properties[44] (think of the classical ones: being chased, being the victim of a disaster, being misunderstood, being unfortunate, etc.). Only, in the former case they terrorize us and make us laugh in the latter. The objects and events that both horror and humor are based on might be in principle indistinguishable,[45] and yet the emotion resulting in either case would be different, precisely because the user's emotional response is not directed at individual objects and events as such, but at the work as a whole.

Not only is Hume's answer[46] effective for the comic, but it also allows us to explain some aspects of the relationship between comedy and horror: in fact, if the emotional reaction of the users is not so much caused by fictitious events or objects as such but by their representation in a particular rhetorical frame, then we can understand why the representation of the same object can terrorize us or make us laugh depending on the style or narrative structure adopted. We do not act foolishly if, seeing the actress Linda Blair spinning her head and goggling her eyes, we are terrified in one case (*The Exorcist*) and laugh out loud in the other (in *Repossessed*). In fact, what triggers our reaction is not the event itself (which is the same in both films), but the narrative style and the general structure in which such event is inserted. These are the reasons why we are afraid in one case and we laugh in the other.

On the other side, the one related to the *pars subjecti*, the transition from horror to comedy seems to be favored by a certain similarity between the two

[41] J. *Whale, Frankenstein*, con C. Clive, M. Clarke, J. Boles, B. Karloff (USA, 1931).

[42] W. *Friedkin, The Exorcist*, con L. Blair, J. Miller, E. Burstyn (USA, 1973).

[43] *Carroll* (1999), pp. 145-160.

[44] *Ibid*: 147.

[45] Sometimes it is the same actor that plays the role of the same character in the parody. Think of Linda Blair, who is possessed by the devil both in *The Exorcist* and in *Repossessed*.

[46] *D. Hume* (1757).

types of reactions that in both cases contain elements such as stupor and anxiety mixed to pleasure (well summarized in the concept of the "uncanny").[47] This may explain why the boundary between the two genres is perceived as thin: we easily move from one emotional reaction to the other by the mere change of narrative register. I will not dwell further on this last point, the implications of which would lead me too far, and I will conclude this essay by presenting one of the rare cinematographic works where actually horror, tragedy and comedy coexist (and the effect on users is understandably explosive): *The Meaning of life*.[48]

The Meaning of Life is meant to represent human life from the moment of birth until death, and does it by changing the style of the narrative so often, and so constantly violating the most basic narrative rules, that it is simultaneously hilarious and horribly tragic. The events narrated are the most varied, ranging from a couple in financial difficulty selling their children for experiments, to a sex education class where the students are forced to watch the teacher have sex with his wife; then two men dressed as tiger cut a soldier's leg for a joke, followed by two nurses who go to the house of a gentleman and take his liver. Finally we move towards the end by seeing a scene in which a man eats to the point of exploding and then one in which a person sentenced to death personally chooses the type of execution as if he were choosing a pair of socks.

Not only do the objects and events described in this work have all the features that are typical both of comedy films and of horror movies, but the interplay between a change of register and the other highlights how the style and narrative determine a certain kind of emotional response instead of another. From the Humean theory it can be concluded that the user's emotional response is always directed at the work as a whole (be it comical, tragic or both), which is characterized as an object of higher order that can never be reduced to its constituent objects (and in fact, as we have seen, the same scene with the same actors can make us laugh *and* cry).

As candidly put by the announcer at the end of Monty Python's film, now that "What [viewers] want is filth: people doing things to each other with chainsaws during tupperware parties, babysitters being stabbed with knitting needles by gay presidential candidates, vigilante groups strangling chickens,

[47] For an analysis of the concept of the uncanny – the feeling that develops when the same object or event is perceived as familiar and strange at the same time – which is at the center of the link between horror and humor, the classic texts of reference are *Jentsch* (1906) and *Freud* (1919).

[48] *T. Gilliam, T. Jones, Monty Python's - The meaning of life*, with G. Chapman, J. Cleese, T. Gilliam, E. Idle, T. Jones, M. Palin (UK, 1983).

armed bands of theatre critics exterminating mutant goats. Where's the fun in pictures?" It is hard to answer, but with Pascal Engel's help we will surely keep trying.

4. References

C. Barbero, Chi ha paura di Mr. Hyde?, Genova, Il Melangolo, 2010.

C. Barbero, "Pleurer à chaudes larmes de crocodile," Philosophiques, 40/1 (2013), pp. 45-58.

H. Bergson (1900), Laughter: An Essay on the Meaning of the Comic, London, Macmillan and co. limited, 1913.

N. Carroll, "Horror and Humor," The Journal of Aesthetics and Art Criticism, 57 (1999), pp. 145-160.

T. Cathcart, D. Klein, Plato and a Platypus Walk into a Bar, New York, Penguin Books, 2007, p. 44.

M. Ferraris (La fidanzata automatica, Milan, Bompiani, 2007: 195-196)

M. Ferraris, Piangere e ridere davvero, Genova, Il Melangolo, 2009.

S. Freud (1905), The Joke and Its Relation to the Unconscious, New York, W. W. Norton & Company, 1990.

S. Freud (1919), "The 'Uncanny'". The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XVII (1917-1919): An Infantile Neurosis and Other Works, London, Hogarth Press, pp. 217-256.

T. Hobbes (1650), Human Nature in The Elements of Law, Natural and Politic, Oxford: Oxford University Press, 1990.

T. Hobbes (1651), Leviathan, New York: Penguin, 1982.

D. Hume (1757), "Of Tragedy" in The Philosophical Works of David Hume, T. H. Green and T. H. Grose (eds.) Vol 3, London: Longman, Green, 1874–75.

E. Jentsch (1906), "On the Psychology of the Uncanny," Angelaki 2.1 (1995);

I. Kant (1790), Critique of Judgment, James Creed Meredith (tr.), Oxford: Clarendon Press, 1911.

A. Koestler (1964), Act of Creation, London, Arkana, (quote from the 1989 reprinted version).

J. Levinson, Humour, in E. Craig (ed.), Routledge Encyclopedia of Philosophy, London, Routledge, 1998.

J. Levinson, Contemplating Art, Oxford, Oxford University Press, 2006: 390-394.

J. Lipitt, "Humour," in D. COOPER (ed.), A Companion to Aesthetics, Oxford, Blackwell, 1992, pp. 199-203.

M. W. Martin, "Humour and the Aesthetic Enjoyment of Incongruities," The British Journal of Aesthetics, 23 (1983), pp. 74–84.

A. Meinong (1904), "The Theory of Objects" Isaac Levi, D. B. Terrell, and Roderick Chisholm (trans.) in Roderick Chisholm (ed.) Realism and the Background of Phenomenology, Atascadero, CA: Ridgeview, 1981, pp.76-117.

G. Minois, Histoire du rire et de la derision, Paris, Fayard, 2000.

J. Morreall, Comedy, in M. Kelly (ed.), Encyclopedia of Aesthetics, Oxford, Oxford University Press, 1998.

J. Morreall, Comic Relief. A Comprehensive Philosophy of Humor, New York, Wiley-Blackwell, 2009a.

J. Morreall, Comic Relief. A Comprehensive Philosophy of Humor, New York, Wiley-Blackwell, 2009b.

C. Radford, "How Can We Be Moved by the Fate of Anna Karenina?," Proceedings of the Aristotelian Society, 49 (1975), pp. 67-80.

H. Spencer (1860), "The Physiology of Laughter," in Essays on Education and Kindred Subjects, London, Dent, 1911, pp. 298-309.

K.L. Walton, "Fearing Fictions," The Journal of Philosophy, 75 (1978), pp. 5-27;

K.L. Walton, Mimesis as Make-Believe: On the Foundations of the Representational Arts, Cambridge (Mass.), Harvard University Press, 1990;

K.L. Walton, "Spelunking, Simulation and Slime. On Being Moved by Fiction," in M. Hjort & S. Laver (eds.), Emotion and the Arts, Oxford, Oxford University Press, 1997, pp. 37-49.

9

Fictions, émotions et araignées au plafond *

FABRICE TERONI

> Il faut remarquer que la fiction, quand elle a de l'efficace, est comme une hallucination naissante: elle peut contrecarrer le jugement et le raisonnement, qui sont les facultés proprement intellectuelles.
>
> H. Bergson, *Les deux sources de la morale et de la religion*

*Mes remerciements à Florian Cova, Julien Deonna, Amanda Garcia, Federico Lauria et Hichem Naar pour leurs précieuses remarques.

On compte au nombre des intérêts philosophiques de Pascal Engel les difficiles questions entourant la nature de la rationalité ainsi que certains des problèmes fondamentaux liés à nos interactions avec les œuvres de fiction. Aussi m'a-t-il paru approprié de le fêter à travers une contribution tentant de faire le pont entre ces deux domaines : celle-ci portera donc sur la question de savoir si et dans quelle mesure les émotions que suscitent les œuvres de fiction peuvent être rationnelles.

Le sans doute trop fameux et à vrai dire un peu barbant paradoxe de la fiction (Radford 1975) restera pour l'essentiel à l'arrière-plan de ma discussion ; la raison principale en est que les débats qui font rage depuis la parution de cet article fondateur ont donné lieu à des interprétations très diverses de la nature de ce paradoxe et que je ne souhaite pas particulièrement, en cette occasion festive, ennuyer mon lecteur avec des questions d'exégèse post-Radfordienne. Ceci étant dit, une distinction importante qui affleure bien souvent au sein de cette littérature, pour se voir presque aussitôt négligée, est celle entre les deux questions suivantes : « comment les émotions peuvent-elles être suscitées par des œuvres de fiction ? » et « les émotions suscitées par de telles œuvres peuvent-elles être rationnelles ? »[1] Dans ce qui suit, je me concentrerai exclusivement sur la seconde de ces questions et chercherai à montrer, à l'encontre de ce que suggère Bergson, que la plupart de nos réponse affectives à la fiction ne sont pas plus irrationnelles qu'hallucinatoires.

Je souhaite par ailleurs délimiter le champ de mes investigations comme suit. Je m'intéresserai spécifiquement à ce que Stacie Friend (2010) a décrit comme le « contexte d'engagement » avec les œuvres de fiction, à savoir le contexte dans lequel le sujet se trouve en rapport direct avec ces œuvres. Il se distingue des différents contextes où le sujet n'est pas dans un tel rapport bien que s'y dévoile une influence plus ou moins directe de son interaction avec des œuvres de fiction – comme par exemple le cas où sa récente lecture d'un roman à l'eau de rose lui fait prendre sur le chemin du travail un simple clignement de paupières pour une œillade. Mon attention se dirigera en outre vers la rationalité dite épistémique et je ne me mêlerai pas de savoir si la fréquentation des œuvres et personnages de fiction peut être approuvée pour des raisons prudentielles ou morales – cela ne fait de toute façon aucun doute dans la mesure où elles nous sont indispensables afin d'affiner nos sensibilités affectives.[2] Enfin, je laisserai en suspens pendant la majeure partie de ma dis-

[1] Pour une confirmation particulièrement frappante de ce constat, voir Tullmann et Buckwalter (2013). Livingston et Mele (1997) soulignent quant à eux avec insistance combien ces deux questions diffèrent.

[2] Pour une discussion de ces différents types de rationalité affective, voir en particulier Rabi-

cussion le questionnement strictement ontologique concernant la possibilité pour des situations fictionnelles d'instancier les propriétés évaluatives auxquelles nous allons rapidement constater que les émotions sont intimement liées ; je présuppose donc que la question de la rationalité des émotions suscitées par la fiction ne s'y réduit pas.[3] J'y reviendrai vers la fin de ma discussion.

Ainsi délimitée, je suis d'avis qu'une investigation de la rationalité des émotions suscitées par la fiction offre un point de vue privilégié à qui souhaite mieux cerner la nature de notre implication dans la fiction ainsi que celle de l'intelligence affective. C'est en tout cas ce que je m'efforcerai de montrer au cours de ma discussion, qui se structure de la manière suivante. La première section esquisse la nature du lien entre émotions et valeurs et introduit une conception selon laquelle les émotions sont des attitudes évaluatives. Au cours de la deuxième section, nous nous tournerons vers le problème de la rationalité des émotions pour mettre l'accent sur deux de ses aspects – celui qui regarde le rapport entre une émotion et sa base cognitive, d'un côté, et, de l'autre, celui entre l'émotion elle-même et les jugements et comportements auxquels elle peut donner lieu. La troisième section applique ces considérations à propos de la rationalité affective aux émotions suscitées par la fiction, dont je distingue trois types principaux : les émotions esthétiques, les émotions-blob et les émotions-pour. Une quatrième et dernière section explore quelques conséquences des conclusions auxquelles m'auront amené ce qui précède.

1. Le lien aux valeurs

Un certain nombre de théories des émotions se détachent clairement des autres en ce qu'elles partagent l'idée selon laquelle les différents types d'émotions se distinguent des autres états mentaux et les uns par rapport aux autres de par leurs rapports aux valeurs – ce qu'illustrent les liens entre la peur et le danger, la tristesse et la perte, la joie et le succès, la honte et le dégradant, l'admiration et l'admirable, et ainsi de suite. Parmi ces théories hétérogènes, certaines analysent les émotions en termes de jugements évaluatifs, d'autres en termes de perceptions évaluatives, d'autres encore en termes d'attitudes évaluatives.

nowicz et Ronnow-Rasmussen (2004).

[3] Cette présupposition me paraît appropriée dans la mesure où les protagonistes des débats quant à la rationalité des réponses affectives à la fiction ne semblent pas mus par un souci de cette nature.

Que l'on incline pour l'une ou l'autre de ces théories, il est aujourd'hui assez usuel de décrire ce rapport aux valeurs en affirmant que ces dernières sont les objets 'formels' des émotions – un terme d'origine scholastique qui a pour fonction de démarquer ce lien, d'une nature nous allons le voir assez singulière, de celui qui unit les émotions à leurs objets 'particuliers'. Les objets particuliers d'un certain type d'émotion sont ainsi susceptibles de varier considérablement – Jean a peur du loup, Marie des araignées, Jacques de perdre son pécule et Michelle de voir la Suisse renoncer à son indépendance. Tel n'est pas le cas de l'objet formel d'un certain type d'émotion, qui reste quant à lui constant – l'un de ses rôles consiste en effet à l'individuer. On affirmera ainsi que la peur est ce type d'émotion qui a pour objet formel le danger, la honte cet autre type d'émotion pour lequel le dégradant joue ce rôle. Cette individuation des types d'émotion n'est que l'un des rôles joués par les objets formels ; un autre consiste à entrer dans la spécification des conditions de correction des émotions. La peur est en effet correcte ou incorrecte en fonction de la dangerosité avérée ou non de son objet particulier (le loup, la perte de son pécule).[4]

Souligner le fait que les objets formels jouent ces deux rôles laisse cependant complètement ouverte la question de la nature du lien entre un certain type d'émotion et l'objet formel qui lui est propre. Afin de mieux la comprendre, il vaut la peine de prendre un peu de recul et de considérer différents types d'états mentaux et leurs objets formels. Voici trois exemples : le lien entre croyance et vérité, celui entre conjecture et probabilité et celui entre supposition et possibilité. Dans les trois cas, nous faisons référence à une propriété qui permet de dissocier les contributions respectives de deux aspects fondamentaux des états mentaux, à savoir l'attitude et le contenu. En premier lieu, on peut spécifier à son aide ce que les différentes instances d'un même type d'état mental ont en commun en dépit du fait qu'elles possèdent différents contenus. Ainsi, si les croyances que la Terre est ronde, que l'herbe est verte et que les chats sont gris la nuit possèdent des contenus bien distincts, un certain rapport à la vérité en fait trois instances de l'attitude de croire. En second lieu, le rapport entre types d'états mentaux et objets formels permet d'expliquer comment des états possédant le même contenu peuvent néanmoins avoir des conditions de correction distinctes. Michel peut par exemple supposer ce que Marie croit : l'état mental du premier sera correct si son contenu est possible,

[4] Les différentes facettes des rapports entre émotions et objets formels sont discutées dans Teroni (2007). Les développements qui suivent s'inspirent de la discussion plus détaillée dans Deonna et Teroni (2012 : chap. 7).

alors que celui de la seconde ne le sera que s'il est vrai. Ce qui explique cette différence, c'est bien sûr la présence de deux attitudes distinctes par rapport à un seul et même contenu.

Ce qui conduit naturellement à se demander dans quelle mesure les objets formels en question permettent d'élucider la nature des types d'états mentaux. Je suis tenté de répondre : au moins dans la mesure où nous pouvons dire, ce qui n'est après tout pas rien, que l'on est en présence d'une croyance lorsque l'attitude est correcte si et seulement si p est vraie, en présence d'une conjecture lorsque celle-ci est correcte si et seulement si p est probable, et ainsi de suite. Prendre ceci pour argent comptant revient à souligner que les attitudes contribuent de manière significative aux conditions de correction des états mentaux sans pour autant figurer dans leur contenu – après tout, on pécherait sans doute par excès si l'on exigeait d'un sujet qu'il représente l'attitude qui est la sienne afin de l'avoir. Il n'est pas plus requis de croire qu'on croit pour croire que de conjecturer qu'on conjecture pour conjecturer.[5]

Revenons-en aux rapports entre types d'émotions et différentes valeurs : ils paraissent analogues à ceux que l'on rencontre entre un type d'état mental comme la croyance et une propriété comme la vérité. Craindre un loup ou de perdre son pécule, c'est avoir une certaine attitude envers un contenu donné. Ce contenu fournit à l'émotion son objet particulier, le loup ou la perte de son pécule, vers lequel les instances de peur en question sont intentionnellement dirigées. Ce contenu pourrait être celui d'une croyance et devoir donc être vrai afin que l'état mental soit correct. Or, il est l'objet d'une autre attitude : la peur. Dire que cette dernière a une certaine valeur, le danger, pour objet formel revient à dire que son contenu représente quelque chose qui, afin que cette attitude de peur à son encontre soit correcte, doit constituer un danger pour le sujet. De même que dans le cas de la croyance, c'est donc bien en vertu du fait que les différents types d'émotions sont autant d'attitudes distinctes qu'ils ont des valeurs pour objets formels, et non en vertu du fait que ces valeurs sont représentées par ces émotions. Il n'est pas plus requis de craindre qu'on craint pour craindre que de croire qu'on croit pour croire. Cela me paraît somme toute assez évident : une saine dose de bon sens devrait nous conduite à affirmer que la différence entre peur, joie, tristesse et colère n'est pas – ou du moins pas en premier lieu – une différence au niveau de ce qui est représenté, mais bien plutôt au niveau de l'attitude prise par rapport à ce qui l'est.

Les considérations qui précèdent autorisent donc à donner sa préférence

[5] Au vu du son passé davidsonien, on m'excusera peut-être de laisser ici à Pascal Engel le soin d'établir de mémoire les passages exacts où cette affirmation est remise en question.

à une théorie qui comprend les émotions comme des attitudes évaluatives plutôt que comme des jugements ou des perceptions de valeur. Je n'ai pas pour intention de défendre ici une conception particulière de la nature des attitudes émotionnelles. La discussion qui va suivre présuppose néanmoins que l'on souligne certains de leurs traits fondamentaux. Premièrement, les attitudes émotionnelles contribuent de la manière que je viens d'esquisser aux conditions de correction évaluatives – c'est en cela qu'elles constituent des attitudes évaluatives. Ensuite, elles sont basées sur d'autres états mentaux. On ne prend pas peur 'tout court' : la peur porte toujours sur quelque chose que l'on voit ou entend, dont on se souvient ou encore à propos duquel on a une pensée. En forme de slogan : le contenu des émotions est hérité de celui de leurs *bases cognitives*. La peur que l'on prend en percevant ou en se souvenant d'un objet ou d'une situation est dirigée de façon intentionnelle vers cet objet ou cette situation perçue ou remémorée. Et, ainsi que je le soulignais précédemment, les bases cognitives sont susceptibles de varier considérablement, des épisodes émotionnels les plus simples fondés sur des états perceptifs à ceux qui reposent sur des inférences complexes de la part du sujet qui les ressent. Enfin, le fait que les émotions possèdent de telles bases explique au moins en partie pourquoi nous sommes enclins à demander des raisons en leur faveur. Si nous constatons que la peur de Jean est basée sur la perception d'une boule de poils rabougrie et édentée, nous ne nous en formerons ainsi pas une très bonne opinion.

Ceci étant clarifié, je vais maintenant tourner mon attention vers les conséquences épistémiques du fait que les émotions sont des attitudes évaluatives possédant des bases cognitives.

2. Deux facettes de la rationalité émotionnelle

Il convient de distinguer ici deux grandes questions concernant la rationalité des émotions. La première porte sur ce qui se passe en amont d'une émotion et plus particulièrement sur la relation qu'elle entretient avec sa base cognitive. La seconde porte sur ce qui se passe en aval d'une émotion et donc sur les rapports entre celle-ci et les jugements et comportements auxquels elle donne typiquement mais, bien sûr, pas nécessairement lieu. Considérons-les tour à tour.

Que la rationalité d'un sujet doive se mesurer en partie à l'aune de ce qui se passe *en amont* des émotions qu'il ressent ne fait pas l'ombre d'un doute : il est somme toute évident que nous la mesurons souvent de la sorte. C'est ainsi

que nous considérons la crainte d'un défaut de paiement des banques européennes d'un investisseur au fait de leur situation financière comme rationnelle, et comme irrationnelle la colère suscitée par l'innocente remarque d'un proche connu pour son ingénuité. Naturellement, dans la mesure où les émotions possèdent des conditions de correction évaluatives, les bases cognitives d'une émotion d'un certain type doivent, d'une manière ou d'une autre, fournir des raisons étroitement liées à l'instanciation d'une valeur donnée. Chercher à dépasser ce constat, c'est faire face aux problèmes que l'on rencontre dès lors que l'on s'évertue à spécifier les conditions devant être remplies par des considérations afin qu'elles puissent constituer les raisons pour un sujet d'adopter une certaine attitude. Doit-il se rendre compte que ces considérations sont des raisons pour cette attitude ? Si tel est le cas, de quelle manière précise doit-il en prendre conscience ? Plus spécifiquement, la base cognitive d'une émotion doit-elle permettre au sujet d'accéder (ou lui donner l'impression d'accéder) à la valeur pertinente ? Ou alors un lien d'indication plus lâche suffit-il ? Laissons ces questions importantes de côté dans la mesure où seule l'existence d'une rationalité en amont des émotions importera pour ce qui suit.[6]

Il est tout aussi évident que la rationalité d'un sujet se manifeste également *en aval* des émotions qu'il ressent, et plus particulièrement au niveau de leurs conséquences sur les jugements qu'il forme ainsi que sur les comportements qu'il adopte. On se demande ainsi couramment si telle ou telle personne a eu raison de juger la situation dangereuse ou la remarque offensante, étant entendu que ces interrogations portent sur des situations dans lesquelles ces jugements sont formés parce que de la peur ou de la colère est ressentie. Et, pour ce qui regarde le comportement, on peut se demander si cette instance de peur justifie de prendre ses jambes à son cou, ou si cette colère justifie une certaine forme de punition.

Nous faisons donc face à une séquence typique de la forme suivante : une certaine base cognitive suscite une émotion qui, à son tour, engendre certains jugements et comportements. C'est pourquoi l'on peut interroger la rationalité du sujet en deux points, les questions pertinentes étant : « ces bases cognitives donnent-elles de bonnes raisons pour cette réaction affective ? » et « cette réaction affective fournit-elle de bonnes raisons de juger ou de se comporter de telle ou telle façon ? » Comme on le constate aisément, la rationalité des émotions est semblable à celle des croyances (« quelles sont les raisons de croire que tel est le cas ? », « cette croyance donne-t-elle des raisons pour cette autre

[6] Pour une tentative de réponse, voir Deonna et Teroni (2012 : chap. 8).

croyance ou de se comporter de cette manière ? ») et se distingue de celle des états perceptifs, plus circonscrite car ne concernant que ce qui se passe en leur aval. On ne se demande en effet jamais si un sujet possède de bonnes ou de mauvaises raisons de percevoir, mais seulement si ses états perceptifs lui donnent de bonne raisons de croire ou de se comporter.

A la lumière des deux aspects de la rationalité propre aux émotions que je viens de distinguer, la question qui va maintenant m'intéresser est celle de savoir si l'on peut conclure que l'interaction avec les œuvres de fiction est un centre de formation d'émotions irrationnelles.

3. Application à la fiction

Afin d'y répondre, il faut garder à l'esprit la grande variété des émotions qui peuvent être suscitées par les œuvres de fiction. Pour ce faire, je vais m'intéresser à trois grands types d'émotions qui me paraissent symptomatiques de notre interaction avec ces œuvres – sans pour autant présupposer, bien sûr, que la distinction entre ces types soit toujours aisée à opérer en pratique ou qu'il n'existe aucune relation intéressante entre les phénomènes qu'ils regroupent. Ces trois types d'émotions peuvent être désignés comme suit : les émotions esthétiques, les émotions-blob et les émotions-pour. Je vais écarter de ma discussion les émotions qui, bien qu'occasionnées par des œuvres de fiction, sont dirigées vers des objets non-fictionnels – comme lorsque l'on s'indigne d'une remarque que l'on avait crue innocente mais dont la lecture d'un roman nous fait prendre la véritable mesure. Non que je les considère insignifiantes, mais elles ne me paraissent pas soulever de questions ressortant spécifiquement de la rationalité de nos réponses affectives aux œuvres de fiction.

Considérons en premier lieu les *émotions esthétiques*, comme l'admiration que l'on peut ressentir à la contemplation d'un tableau de Chardin, ou le dégoût que peut susciter un long-métrage sursaturé de références idéologiques, à l'exemple de ces indigestes *Neiges du Kilimandjaro* (Robert Guédiguian, 2011). L'admiration est rationnelle pour autant que les conditions suivantes soient remplies : le sujet perçoit certaines des propriétés de l'œuvre (la délicatesse du coup de pinceau, l'originale intimité du sujet, la subtile harmonie des tons) qui donnent des raisons de l'admirer et réagit à l'occurrence de son émotion par un comportement (le tableau est l'objet d'une attention soutenue de sa part, il est exploré du regard pour mieux en cerner les symétries, etc.) et un jugement (« Quelle petite merveille ! ») justifiés par l'émotion. Ce diagnostic

me paraît parfaitement similaire à celui que l'on souhaite poser pour tout type d'émotion non-esthétique, et il ne me semble pas plus problématique ici. Remarquez en particulier qu'il ne présuppose nulle forme de réalisme à propos de l'admirable, mais tout au plus que certains contenus perceptifs donnent des raisons d'admirer.

Tournons-nous maintenant vers un type d'émotion au pedigree plus incertain et qui a joué un rôle non négligeable dans les réflexions philosophiques récentes sur l'irrationalité des émotions engendrées par les œuvres de fiction.[7] Ces *émotions-blob*, ce sont bien sûr ces instances de peur ou d'effroi dont certains metteurs en scène à la Chuck Russell ont fait un juteux fond de commerce et vers lesquelles l'on ne s'est que trop systématiquement tourné pour comprendre le rapport entre réactions affectives à la fiction et irrationalité. Trop, parce que les réactions de ce type ne manifestent aucune forme substantielle d'irrationalité de la part du sujet, pas plus qu'elles ne possèdent un lien privilégié aux œuvres de fiction. Voyons pourquoi.

Comment caractériser les émotions-blob ? Il s'agit de réponses affectives à certaines classes restreintes de stimuli que l'on pourrait qualifier avec John Deigh (1994) de naturelles et qui sont totalement imperméables à l'évidence qui pourrait se trouver à la disposition du sujet. On lorgne ici du côté des frissons suscités par les serpents, de l'arachnophobie ou encore du dégoût pour les matières organiques en putréfaction, réactions dont l'existence se prête sans doute à des explications évolutionnaires plus ou moins édifiantes. Les long-métrages du type *Le blob* (Chuck Russell, 1988) se greffent sur et parasitent pour ainsi dire un lien stimulus-réponse affective de ce type. En outre, le peu d'entrain des émotions-blob à se laisser pénétrer cognitivement les rapproche des illusions perceptives du type Müller-Lyer et des jugements qu'un sujet naïf pourrait être amené à former lorsqu'il y fait face. La plupart, si ce n'est toutes les émotions-blob suscitées par les œuvres de fiction possèdent en effet des bases cognitives visuelles ou auditives qui ne donnent au sujet aucune bonne raison de réagir affectivement par de la peur ou du dégoût – et c'est bien sûr un diagnostic sur lequel les observateurs et le sujet qui ressent l'émotion se rejoignent, du moins si ce dernier n'est pas d'une effarante naïveté. Pour autant, il ne faudrait pas en conclure que le sujet y révèle son irrationalité. C'est ce que l'application de notre distinction entre rationalité en amont et en aval d'une émotion va nous permettre de constater.

A propos de ce qui se déroule en amont, j'ai déjà souligné le fait que le sujet

[7] Pour une grand part, cet état de fait s'explique bien sûr par l'immense influence exercée par la discussion de Kendall Walton (1978).

ne peut tout simplement pas contrôler sa réponse affective ou encore faire en sorte qu'elle réponde à l'évidence qui se trouve à sa disposition – sa rationalité ne peut dans cette mesure pas se manifester d'une quelconque manière à ce niveau. Autant reprocher à un sujet de voir les lignes de Müller-Lyer comme étant de tailles différentes ou de lever son tibia lorsque le médecin joue de son marteau. C'est une conclusion qu'il me semble difficile d'éviter pour autant que l'on entende par « irrationnel » quelque chose de plus qu' « incorrect ».[8] Dans cette mesure, la rationalité d'un sujet en proie à une émotion-blob ne saurait se manifester qu'en aval de sa réponse affective, à savoir au niveau du contrôle qu'il exerce sur les conséquences comportementales et judicatives habituelles de l'émotion qu'il ressent. Et si certains des premiers soubresauts de la peur et du dégoût peuvent se montrer quelque peu rétifs, une absence de contrôle de ce type est bien plutôt l'exception que la règle – le fait si souvent souligné que seuls de rares spectateurs quittent les séances de cinéma d'horreur dénote l'existence d'une capacité répandue à exercer un contrôle rationnel sur les émotions-blob. Lorsqu'elles sont déclenchées par des œuvres de fiction, celles-ci ne sont donc pas plus systématiquement irrationnelles en raison de ce qui se passe en leur aval qu'en leur amont.

En outre, lorsque ce qui a lieu en leur aval dénote une coupable absence de contrôle, celle-ci n'est en aucun cas propre à la fiction ; au mieux joue-t-elle un rôle de miroir grossissant pour une forme d'irrationalité qui se manifeste aussi bien lorsque, par exemple, le mouvement des herbes provoqué par ce que le sujet sait être un inoffensif orvet lui fait néanmoins prendre ses jambes à son cou. Ceci rejoint un constat plus général : il serait erroné de considérer les émotions-blob comme un phénomène privilégié pour la compréhension de notre implication affective dans les œuvres de fiction et, plus généralement, de la forme de rationalité à laquelle elles sont sujettes. Au mieux peut-on conclure que des émotions de ce type sont suscitées par quelques genres de fiction bien circonscrits qui parasitent des liens rigides entre stimuli et réponses affectives. Négliger ce constat a eu et continue d'avoir un impact regrettable sur la philosophie contemporaine des émotions en permettant d'entériner l'idée selon laquelle la rationalité d'un sujet s'exerce sur ses émotions de la même façon qu'elle s'exerce sur ses états perceptifs, à savoir exclusivement en leur aval.

[8] Davies (2009) est conscient de la possibilité d'opter pour une telle stratégie, qu'il attribue à Morreal (1993). Il considère cependant que l'absence de réaction des sujets habitués à une œuvre de fiction donnée interdit une analyse en termes de ce qu'il décrit comme des « émotions réflexe ». Par ailleurs, il me semble présupposer qu'une émotion doit nécessairement être accompagnée des jugements existentiels pertinents. Ces deux raisons de rejeter cette stratégie ne me paraissent pas résister à l'examen.

C'est en tout cas ce qui se laisse aisément déduire du rôle joué par les illusions perceptives du type Müller-Lyer dans les discussions contemporaines à propos de la proximité des émotions et des états perceptifs. Pourtant, le fait que certaines émotions résistent bel et bien à l'évidence à la disposition du sujet ne parle pas plus en faveur d'une assimilation de l'ensemble des émotions aux états perceptifs qu'aux états doxastiques dont on sait, après tout, que la sensibilité à l'évidence est plus souvent que de raison *de jure* plutôt que *de facto*. En tout état de cause, ces rares cas de résistance têtue à l'évidence ont fait prendre à beaucoup des vessies pour des lanternes : dans une immense majorité de cas, la rationalité d'un sujet se mesure prioritairement à l'aune des raisons pour lesquelles il ressent. A ce niveau, les émotions s'assimilent plus aisément aux croyances et aux conjectures qu'aux états perceptifs. Notre exploration des émotions-blob aboutit donc à la conclusion suivante : celles-ci ne sont pas plus symptomatiques de notre implication émotionnelle dans des œuvres de fiction que de la rationalité affective.

Il nous reste à considérer le troisième type d'émotions suscitées par la fiction, à savoir les *émotions-pour*. A ce stade de notre discussion, un exemple qui vient immédiatement à l'esprit est celui de la peur ressentie par le spectateur pour Meg Penny alors qu'elle se trouve à portée des attaques visqueuses du blob. Un autre, plus propret, est la tristesse d'un lecteur de Flaubert pour Emma. Il s'agit peut-être là des émotions les plus révélatrices de nos interactions avec les œuvres de fiction. Et, bien sûr, la séquence complète, tronquée en son amont dans le cas des émotions-blob, est ici intégralement restaurée. Elle s'amorce cependant par une croyance ou, plus vraisemblablement, par un ensemble de croyances possédant un contenu distinctif : le sujet croit que, *dans la fiction*, ceci ou cela est le cas. C'est là un trait essentiel qui les distingue des émotions pour des entités non-fictionnelles.

Les émotions-pour suscitées par les œuvres de fiction sont-elles systématiquement irrationnelles ? En raison de leur sensibilité à l'évidence fournie par la fiction, leur procès ne saurait se dérouler de la façon dont nous avons instruit celui des émotions-blob. Ce trait ne les immunise naturellement pas contre plusieurs types d'irrationalité. Ainsi, un spectateur peut être suffisamment inattentif pour s'emmêler les pinceaux : Meg Penny est en sécurité, c'est Brian Flagg qui se retrouve, si l'on ose dire, nez à nez avec le blob. Je profite de l'occasion pour souligner également que certaines émotions-pour peuvent facilement se transformer en émotions-blob. Confortablement installé dans son salon, un spectateur ressent en premier lieu de la peur pour Meg, mais celle-ci laisse place à une pulsion de plus en plus irrépressible de vérifier ce qui se

trouve derrière le sofa …on ne sait jamais.[9] Mais laissons ces cas particuliers de côté pour nous intéresser plus généralement à cette catégorie des émotions-pour des entités fictionnelles et à leur éventuelle irrationalité.

Une ligne argumentative qui a séduit plus d'un philosophe porte sur ce qui a lieu en amont de ces émotions : elle revient à souligner l'irrationalité d'un sujet qui maintient sa peur pour Meg Penny ou sa tristesse pour Emma une fois qu'il réalise que ces femmes n'existent pas. Mais, au fond, pourquoi ? Rappelez-vous que j'ai décidé d'omettre pour l'instant la question ontologique de savoir si les entités fictionnelles peuvent faire face à des situations instanciant les valeurs pertinentes. Il me semble en tout état de cause que la raison invoquée par les partisans de l'irrationalité n'est pas de cette nature : on pointe plutôt du doigt le fait que le spectateur traite ses émotions-pour des entités fictionnelles différemment de ses émotions-pour des entités concrètes. Si l'on a peur pour la personne qui est en train de réparer la chaudière dans la cave – c'est une étuve et le risque d'explosion est accru – et que l'on se rend compte que nul ne s'y trouve, notre émotion va bien sûr se dissiper. Mais alors, comment la peur de notre spectateur pour Meg Penny peut-elle raisonnablement persister alors qu'on lui assène à l'envi que cette femme n'existe pas ? Ne révèle-t-il pas par cette obstination affective une forme profonde d'irrationalité ? Nullement, puisqu'il faut au contraire reconnaître dans cette obstination une pré-condition de la rationalité de son émotion-pour.

En effet, nous avons constaté plus haut que les émotions sont des attitudes qui peuvent posséder une grande variété de bases cognitives et qui doivent en conséquence être modulées par l'évidence pertinente. Si l'émotion de notre spectateur est dirigée vers les événements qui affligent Meg dans la fiction, alors il doit, bien sûr, afin que sa réaction affective soit rationnelle, moduler sa réponse en fonction du flux d'informations que la fiction lui fournit à propos de ces événements : sa peur doit faire place à du soulagement si le blob se retrouve coincé et laisse à la protagoniste le temps de prendre le large, et ainsi de suite. Un verdict semble donc devoir s'imposer : si la peur de notre spectateur pour Meg disparaît dès lors qu'il en vient à apprendre qu'elle n'est qu'une création hollywoodienne peu inspirée, cela nous donnerait de bonnes raisons d'affirmer que quelque chose cloche dans la base cognitive de son émotion. On conclura, par exemple, qu'il a formé des croyances « tout court » au lieu de croyances comprenant un opérateur de fiction, convaincu peut-être, à l'instar de certains participants à la première projection de *L'arrivée d'un train en gare de La Ciotat*, que l'écran qu'il contemple est une fenêtre. Au contraire, le

[9] Ce genre de situation correspond aux cas que Davies (2009) décrit comme « leaky fear ».

fait que sa peur persiste lorsqu'il apprend que « ce n'est qu'un film » ne nous donne aucune raison de l'affirmer : sa réaction affective est de la peur pour une créature fictionnelle qu'il sait parfaitement être telle. Il n'a dès lors aucune raison de chercher à se dépêtrer d'une telle combinaison d'états mentaux : il ne doit pas se décider entre modifier sa croyance que l'entité est fictionnelle et renoncer à sa réponse affective.[10]

Rappelons-nous maintenant combien émotions et croyances sont similaires pour ce qui regarde la rationalité en amont : le philosophe qui insiste sur l'irrationalité d'une émotion-pour une fois que le sujet est convaincu que l'entité pour laquelle il ressent son émotion est fictionnelle se trouve dans une situation inconfortablement proche de celle de quelqu'un qui insisterait, à l'encontre du bon sens, pour que l'on cesse de croire que, dans la fiction, un groupe d'Anglais a rasé la moustache d'Hitler (*Hitler, Dead or Alive*, Nick Grinde, 1942) puisqu'il est de notoriété publique qu'il est resté moustachu pendant toute la durée du conflit. Bien sûr, nous serions irrationnels dans ce cas si nous omettions de prendre en compte le contexte fictionnel pour, par exemple, reprocher à un historien d'avoir fait l'impasse sur cet événement.

Cette dernière remarque débouche naturellement du côté de la rationalité en aval des émotions-pour suscitées par la fiction : un sujet ressentant une telle émotion est-il en proie à une forme d'irrationalité systématique quant à son activité judicative et comportementale ? Une réponse positive reviendrait à affirmer que nous sommes souvent disposés à agir et juger d'une façon qui ne serait appropriée et rationnelle que si les personnages de fiction pour lesquels nous ressentons des émotions existaient vraiment. Que de tels cas existent est indéniable – de nombreux acteurs et actrices sont par exemple regardés par leurs admirateurs d'une manière qui paraît amalgamer les informations fournies par la fiction à celles des journaux people. Pour parler le langage de John Perry (voir par exemple Perry 1980), ces fans n'ouvrent qu'un dossier mental alors qu'au moins deux seraient requis. Mais de telles formes de douce folie demeurent tout de même assez peu fréquentes.[11] Elles manifestent à un

[10] En ce sens, de telles combinaisons d'états mentaux ne font pas l'objet de « réquisits à portée large » chers aux amateurs de raisons. Sur cette question, voir la riche discussion de Kolodny (2005).

[11] Un cas plus complexe auquel le lecteur peut être amené à penser est celui où un spectateur se projette dans un personnage de fiction pour se donner l'illusion qu'il possède un sentiment dépeint dans l'œuvre, puis se comporte dans la vie réelle comme s'il le possédait bel et bien. A son niveau paroxystique, cette forme d'irrationalité peut conduire aux formes de suicides mimétiques qui ont suivi la parution des *Souffrances du jeune Werther*. Pour fascinante qu'elle soit, cette forme d'irrationalité ne me concerne cependant pas dans la mesure où je n'ai pas souhaité m'intéresser aux émotions occasionnées par l'interaction avec des œuvres de fiction mais dirigées vers des

haut degré l'incapacité dans laquelle se trouvent certains sujets à « faire la part du réel et de la fiction », alors que la plupart des émotions-pour ne donnent lieu qu'à des jugements ou à des comportements modulés par l'influence des bases cognitives faisant référence à un contexte fictionnel.[12]

Ce rapide survol de trois types d'émotion suscitées par la fiction – les émotions esthétiques, les émotions-blob et les émotions-pour – me conduit à conclure qu'elles ne révèlent à l'examen aucune irrationalité systématique de la part du sujet qui les ressent. Ce qui se passe en amont et en aval de ces émotions encourage à poser un tel diagnostic.

4. Quelques conséquences

Je souhaite terminer ma discussion en examinant brièvement où mènent nos conclusions concernant les émotions-pour. Partant de l'hypothèse selon laquelle les émotions héritent du contenu de leurs bases cognitives, j'ai souligné que les émotions-pour des entités fictionnelles ont un contenu de la forme : dans la fiction, p. Il est tentant de se servir de ce constat afin de mettre à jour le dilemme suivant : soit nous acceptons une théorie de l'erreur radicale et peu satisfaisante à propos de la fiction, soit nous reconnaissons que certaines émotions-pour des entités fictionnelles sont rationnelles. Examinons cela de plus près.

Il semble en premier lieu difficile de nier que certaines croyances à propos de la fiction soient vraies. Affirmer de but en blanc que, dans la fiction, Meg n'est pas poursuivie par un blob semble suffisamment absurde pour qu'on fasse autant que faire se peut pour éviter une telle théorie de l'erreur. Ce qui a pour suite qu'il existe un vaste ensemble de vérités du type : dans la fiction, p. On introduit alors un principe qui me paraît assez séduisant. Selon ce principe, si le fait que a est F est une situation instanciant telle ou telle valeur pour a, alors le fait que, dans une fiction, une entité fictionnelle a' soit F est une situation instanciant telle ou telle valeur pour a'. Si Jean est en danger lorsqu'il se retrouve dans le bois de Boulogne face à un sanglier écumant de rage, alors un personnage de fiction dans une situation analogue sera également en

objets non-fictionnels.

[12] Les fictions interactives, dont le jeu vidéo est l'exemple le plus saillant, compliquent quelque peu la relation entre émotion-pour et comportement dans la mesure où le sujet peut directement influer sur les situations vécues par le personnage fictionnel (voir par exemple Davies 2009). Rien de ceci ne va à l'encontre de ce qui figure dans le texte, mais force à ne pas conclure que toute forme de comportement déclenché par une émotion-pour une entité fictionnelle est *eo ipso* irrationnelle.

danger.[13] Ce principe me paraît intimement lié à l'idée chère à de nombreux philosophes selon laquelle l'évaluatif survient sur le non-évaluatif, et il serait possible de le défendre par ce biais.[14] Quoi qu'il en soit, c'est une question que je souhaite laisser ouverte ici, ce qui ne me paraît pas constituer un problème sérieux dans la mesure où le principe a une composante intuitive assez robuste.[15] Et, si l'on y souscrit, il semble difficile de ne pas conclure que certaines émotions-pour des entités fictionnelles sont rationnelles. En effet, j'ai souligné plus haut que les émotions avaient pour effet d'ajouter une dimension évaluative aux conditions de correction du contenu fourni par leurs bases cognitives. Or, nous avons admis que les bases cognitives du type 'dans la fiction, p' pouvaient être vraies et que les entités auxquelles ces bases se réfèrent pouvaient faire face à des situations instanciant des valeurs. Dès lors, il n'y a plus aucun obstacle à admettre que des émotions correctes lorsque de telles valeurs sont instanciées puissent s'y rapporter de manière correcte et rationnelle.

Cette conclusion peut enfin être mise en rapport avec un souci souvent exprimé à l'encontre des émotions suscitées par la fiction. La meilleure façon de l'exprimer est à mon avis la suivante : pour autant qu'elle soit rationnelle, la force d'une émotion dirigée vers une entité fictionnelle ne peut pas être identique à celle d'une émotion du même type dirigée vers une entité non-fictionnelle. C'est à mon avis la meilleure manière de comprendre la séduction exercée par l'appel à des états similaires aux émotions mais distincts d'elles comme les tristement célèbres quasi-émotions invoquées par Kendall Walton (1978, 1990). A la lumière de ce qui précède, on peut conclure que ce souci est engendré par une conception erronée du domaine. Les émotions sont des attitudes qui adjoignent une dimension évaluative aux conditions de correction fournies par leurs bases cognitives. Afin de jouer un tel rôle, elles doivent conserver leur nature d'attitude à travers une immense variété de contenus – on ne voit pas bien comment elles pourraient contribuer de la même manière aux conditions de correction si, ainsi que le suggèrent les partisans des quasi-émotions, leur nature se transformait de manière significative lorsqu'elles sont

[13] Beaucoup de paramètres doivent naturellement être ajustés afin de disposer d'un principe satisfaisant – il faut par exemple que a et a' se ressemblent sous certains aspects. Ainsi, il est possible que les talents de Thésée le mettent hors de danger même lorsqu'il fait face au sanglier de Calydon. Je présuppose que les ajustements nécessaires sont possibles.

[14] Sur ces questions, voir en particulier la discussion de Jackson (1998).

[15] Ce qui ne signifie pas, cela va sans dire, que tous les philosophes y aient adhéré. Ainsi, Currie (1990) le rejette dans la mesure où il considère que les événements fictionnels jouissent d'une certaine autonomie par rapport à leurs contreparties concrètes. Pour une discussion, voir Livingston et Mele (1997 : 166f).

suscitées par des œuvres de fiction.[16] En fait, la différence pertinente ne doit pas être située au niveau des émotions elles-mêmes, mais en amont, au niveau de leurs bases cognitives : certaines émotions possèdent des bases 'sérieuses', d'autres des bases fictionnelles. Pour autant que le sujet soit rationnel, la nature de ces contenus ainsi que des attitudes qui les accompagnent doit avoir un impact tout à la fois sur l'émotion ressentie (c'est une émotion-pour plutôt qu'une émotion tout court, par exemple), sur la nature de l'évidence qui compte en sa faveur et en sa défaveur, ainsi que sur les jugements et comportements qu'elle suscite.

5. Références

Davies, S. (2009). Responding Emotionally to Fictions. *The Journal Of Aesthetics and Art Criticism* 67.3, 269-284.

Debus, D. (2007). Being Emotional About the Past : On the Nature and Role of Past-directed Emotions. *Nous* 41.4, 758-779.

Deigh, J. (1994). Cognitivism in the Theory of Emotions. *Ethics* 104.4, 824-854.

Deonna, J. et Teroni, F. (2012). *The Emotions : A Philosophical Introduction*. New York : Routledge.

Friend, S. (2010). Getting Carried Away : Evaluating the Emotional Influence of Fiction Film. *Midwest Studies in Philosophy* 34.1, 77-105.

Jackson, F. (1998). *From Metaphysics to Ethics : A Defence of Conceptual Analysis*. New York : Oxford University Press.

Kolodny, N. (2005). Why Be Rational ? *Mind* 114.455, 509-563.

Livingston, P. et Mele, A. (1997). Evaluating Emotional Responses to Fiction. Dans Mette Hjort et Sue Laver (dir.), *Emotion and the Arts* (pp. 157-198). New York : Oxford University Press.

Morreal, J. (1993). Fear Without Belief. *Journal of Philosophy* 90, 359-366.

Perry, J. (1980). A Problem About Continued Belief. *Pacific Philosophical Quarterly* 61.4, 317-322.

Rabinowicz, W. et Ronnow-Rasmussen, T. (2004). The Strike of the Demon : On Fitting Pro-Attitudes and Values. *Ethics* 114.3, 391-423.

[16] Je suis ici tenté de rajouter qu'il m'est malaisé de concevoir ces quasi-émotions qui auraient aux émotions un rapport similaire à celui d'une image mentale à un état perceptif. Pour ce genre de scepticisme, voir Debus (2007).

Radford, C. (1975). How Can We Be Moved by the Fate of Anna Karenina ? *Procedings of the Aristotelian Society Supplemental* 49, 67-80.

Teroni, F. (2007). Emotions and Formal Objects. *Dialectica* 61.3, 395-415.

Tullmann, K. et Buckwalter, W. (2013). Does the Paradox of Fiction Exist ? *Erkenntnis*.

Walton, K. (1978). Fearing Fictions. *Journal of Philosophy* 75, 5-27.

Walton, K. (1990). *Mimesis as Make-believe*. Cambridge, MA : Harvard University Press.

DEUXIÈME PARTIE

Knowledge and Belief

10

Common Sense and Skepticism : A Lecture

KEITH LEHRER

It is with great pleasure that I contribute the essay that follows to a volume to honor my esteemed friend and colleague, Pascal Engel. Whatever the merits of the current essay, it was provoked by Pascal when he invited me to contribute a lecture on common sense to a conference in Geneva. Having presented a lecture, I thought no more of what use to make of it until the invitation to contribute to this volume arrived. Then I thought, whatever the merits or lack thereof, it was my good friend Pascal who was responsible for the lecture coming into existence. So here are my thoughts, in the form of a lecture, on common sense for a philosopher of admirable philosophical sense that is not at all common.

1. Moore's Proof

G. E. Moore (1925) is famous for his defense of common sense. Moore says, famously, perhaps, infamously, that he can prove the existence of two external objects :

> "Can I prove, now, for instance that two human hands exist ? How ? By holding up my two hands, and say, as I make a certain gesture with the right hand, 'Here is one hand', and adding, as I make a certain gesture with the left, 'and here is another.' ... But did I prove just now that two human hands were in existence ? I do want to insist that I did ; and that it is perhaps impossible to give a better or more rigorous proof of anything whatever."

He says three conditions that must be satisfied for this to be a proof.

1. The premiss is different from the conclusion.
2. I must know the premiss was the case.
3. The conclusion must follow from the premiss.

The conditions are satisfied he says, and so it is a proof.

He adds wisely at the end, that the proof is

> "... shown only by the use of premises which are not known to be true, unless we do know of the existence of external things. I can know things, which I cannot prove : and among things which I certainly did know, even if (as I think) I could not prove them, were the premisses of my two proofs. I should say, therefore, that those if any, who are dissatisfied with these proofs merely on the ground that I did not know their premises, have no good reason for their dissatisfaction."

So we are left with the claim that Moore proved the existence of external objects, two hands, by appealing to premisses that such objects existed, which premisses he knows to be the case but cannot prove.

2. Reid and Hume's Skepticism

The argument is historically influenced by Moore's reading of Reid (1863),[1] and an obvious dissatisfaction with the argument was anticipated by Reid,

[1] All subsequent pages references to Reid are from Reid (1863).

namely, that Moore's assumption that he knew the premisses makes the proof question begging and leaves us without any account of the evidence and justification of those premisses.

Moore also formulated his argument as a reply to Hume. We will find clarification of the issues concerning evidence and justification by returning to the dispute between Hume and Reid. Hume (1739) began with the premiss that all the perceptions of our mind resolve themselves into impressions and ideas by which he meant sensations and feelings. I will not trace the well know skeptical consequences of his starting point. Reid noted the skeptical consequences of Mr. Hume's theory and remarked that we need not despair of a better theory.

Reid set out his theory in reply to Hume. Reid assumed that he and Hume agreed in a common project, to give a theory of the operations of the human based on experience. This commitment to theory separates Reid from Moore. Moore seems to be engaged in a simple *modus tollens* rejection of Hume's skeptical consequences, while Reid sees the need of providing an alternative theory that does not have the skeptical consequences. Reid's objections to Hume's theory were diverse. However, primary among them was that Hume's theory of impressions and ideas, to wit, that all the perceptions of the mind consist of impressions or ideas, which Reid termed *the ideal theory*, was an inadequate theory of the following :

1. The aboutness (now intentionality) of thought and conception, which allows for the possibility of thinking and conceiving of immanent objects that do not exist.

2. Our conceptions and beliefs concerning consciousness, external objects, other minds and the past.

3. Evidence or justification of our beliefs about consciousness, the external world other minds and the past.

Hume seems, on one interpretation, to be sanguine about the third of these. However, he claims to provide an account of conception and belief, even if the conceptions and beliefs are in error. Note the important distinction between having a theory of our conceptions and beliefs concerning the external world with the consequence that such conceptions and beliefs are erroneous, and failing to provide a theory of such conceptions and beliefs at all.

Why must Hume fail to provide a theory of such conceptions, and what is the better theory that avoids the untoward consequences of the ideal theory ? One basic argument is that sensations lack intentionality. They exist but they

are not internally about anything. I may say that I feel a pain, suggesting a distinction between the sensation, the pain, and the feeling of it, but Reid argues that feeling a pain is not distinct from the sensation of the pain, though we may attribute the pain to some external cause. The pain, in itself, lacks the structure of intentionality. Whatever you think of this argument, it must be conceded to Reid that Hume appears to leave the introduction of intentionality, of aboutness, of the immanent objects of thought, unexplained. Actually, I shall in the end fabricate a reply on behalf of Hume, but let us follow the theoretical alternative developed by Reid.

3. Reid's Theory

To account for conception and belief, in the existence of conscious states as well as external objects, Reid argued, we require original capacities that are innate and that are realized as we mature. The experience of impressions would not suffice because their existence does not explain our capacity to conceive of them or believe that they exist. Reid's empirical hypothesis is that that our conception and belief in the existence of conscious states as well as external objects is the result of principles of the faculties of our mind, of consciousness, perception, memory and reason, to form some original conceptions of objects and their qualities. We are not justified in accepting the hypothesis simply because it fills a gap in Hume's theory, however tempting that would be. Hume and Reid were committed to the empiricism of Newton, which meant that experience must be the basis for accepting general principles. The criteria of evidence for the existence of the innateness of the principles of our faculties Reid employed has become standard. They must appear before they can be acquired from tutelage and they must be universal. Reid added that they must be irresistible, and thought that they were revealed in the character of all languages. The main point was that a theory of innate capacities, faculties, could not be rationally justified simply because it avoided skepticism about the external world. It could only be justified empirically.

Let me cut to the chase. Here is Reid's theory in a capsule. An adequate theory of thought and conception requires first principles of our faculties that govern the operations of our minds. The principles give rise, as the child matures, to conceptions and convictions of the existence of consciousness, the existence of external objects and their qualities, the past, and the existence of other minds. These conceptions are not innate ideas in use at birth but arise in us as we mature. What evidence do we have, if any to justify the existence

of those things our faculties lead us to believe exist? Evidence, Reid, says is the ground of belief. It is something felt. The principles of our faculties, of whose operations we are conscious, reveal themselves as the grounds and evidence of our beliefs. The result is that we are conscious of the evidence of our beliefs whatever difficulty we may have in describing the evidence. We feel the force of the operations of our faculties.

In another place, Reid speaking of the justification of beliefs arising from our original faculties denies that it is derived from reasoning. Their justification is their birthright. It is, however, an internally grounded right, for we feel the evidence of such beliefs as they arise. I think this is a deep insight of Reid's and argues against externalist views of conception and evidence as well as externalist interpretations of Reid. People have suggested that memory beliefs cannot be supported by evidence because we do not remember how we acquired them. The ground of such beliefs is not ancient history of how they arose but the feeling that arises from the operation of the faculty of memory when the memory is clear and distinct. The feelings are the evidence of an operation of the faculty and the justification of the belief to which it gives rise.

4. Reid versus Moore

Now, finally, let us return to the issue of proof of the existence of the external world. Reid, unlike Moore, says the matter does not admit of proof to a total skeptic. Suppose someone questions the existence of the external objects, the hands of Moore. Reid notes that such common sense beliefs arise from the faculty of perception. Suppose someone, Hume, as Reid and I interpret him, denies the existence of external objects by appeal to philosophical argument. What reply is available? The reply Reid gave is that the conception and conviction of the external world is the result of a first principle of the faculty of perception. Such beliefs, Reid says, do not require the justification of reasoning. Moreover, they do not admit of it either because they come into existence with all the evidence of which the matter allows.

Is this just dogmatism? Reid has an additional argument. He lists what he takes to be the first principles of our faculties which are also principles of evidence and, hence justification. Someone might ask, "Why should we trust our faculties?" Reid has a reply to anyone, who, like Hume, appeals to philosophical reasoning. It is a consistency argument. Hume, Reid avers, trusts reason, for he reasons to his conclusions with his great genius of argument. Reid asks what justification Hume has for trusting one faculty but not ano-

ther, reason or consciousness but not memory or perception? If Hume has no answer, then he proceeds in a philosophically inconsistent manner.

If someone, Hume perhaps, were to inquire what evidence Reid has that his faculties are not fallacious, his reply is simple. It is a first principle, the seventh in his list, but one I have called, Lehrer (1998), the First First Principle, that our faculties are not fallacious. He puts this in another way by saying they are trustworthy. He notes that this first principle seems to have a priority in evidence over all the rest. Returning to the First First Principle and Moore, it is clear why Reid concedes something very much like what Moore concedes about a lack of proof of the knowledge of his premisses. If proof of knowledge is pressed back by a total skeptic to the First First Principle, the prior principle of evidence and knowledge, and the total skeptic is willing to deny the trustworthiness of all our faculties, Reid says he puts his hand over his mouth in silence. The reason is that you can give no proof of the First First Principle that does not assume it as a premiss, and so you can give no proof. However, as Moore and Reid agree, though we cannot prove the total skeptic wrong, we can know that he is wrong. The First First Principle provides the evidence for us to know what we cannot prove.

This brings us to the issue of common sense. The First First Principle is the first principle of common sense. Common sense on this account is not simply the judgments on which people agree, for they may come to agree on all sorts of foolishness from some common error. Common sense consists of those judgments that arise from the first principles of our faculties. Of course, we may try to ignore them, as Mr. Hume attempted to do in philosophical treatise. But Hume conceded that he failed to do so as soon as he left his study.

Why should we trust those judgments, the judgments that result from original first principles of our faculties? There are two main answers.

One concerns self-trust as I would put it, or trust in our faculties, as Reid would put it. Any reasoning, pro or con, concerning a judgment appeals to the faculty of reason and to other faculties that give rise to the premisses. We either trust our faculties, at least some of them, or reasoning is useless, and, most critically, that includes any reasoning against trusting our faculties. We may modify and qualify the trust we place in some judgments arising from our faculties as experience instructs us about the fecundity of error. But any correction of the use of our faculties presupposes that we place our trust in our faculties. We cannot even find grounds for mistrusting our faculties without appealing to them and assuming the First First Principle affirming that they are not fallacious.

The second answer is that we cannot resist. Indeed, Reid suggests, that there are conceptions and convictions that reveal their origins by their irresistibility. Try to put aside the original judgments of your faculties, if you wish to do so, but you will find, like Mr. Hume, that you cannot succeed. The second answer fills in a limitation of first. Someone might, like Cratylus, attempt to refuse to accept any view and satisfy himself by remaining silent and only moving a finger. But he will not succeed in his attempt to proceed in this way. Whether he speaks or not, he cannot resist the force of his faculties as they give rise to conception and conviction without his bidding and, as might be the case here, contrary to his will.

This brief account of Reid's defense of common sense replaces dogmatic rejection of skepticism with a theory of common sense conception, conviction and the evidence and justification thereof. Reid admits this is no reply to the total skeptic. But, he might add, it is probable than none exists. The skeptic who reasons on his behalf assumes some faculty ; reason at least, is not fallacious.

5. Skepticism versus Self-Trust

What are we to make of this ? Someone might reply, as I once did, Lehrer (1971) that the skeptic need not suppose he is justified to proceed with his argument. He may tell you what he believes without any claim to justification or knowledge, and thereby, acknowledging his common sense beliefs, refuse to trust them, treat them as fallacious, and not at all justified. Perhaps Hume shared such an attitude. I once suggested that such a skeptic could remain consistent. I now have my doubts. The fundamental doubt concerns self-trust concerning what one accepts or refuses to accept. Any reasoning about the merits of beliefs, whether it leads to a positive evaluation, what I have called and will here call, *acceptance*, or the opposite, rejection, depends on trusting oneself as one proceeds. The contrast between acceptance and belief is examined in detail by many authors in Engel (2000). Without adopting Reid's theory of our faculties, I note that to involve oneself in any reflection about the merits or defects of what one believes, one cannot escape from self-trust. Moreover, the person who trusts himself must consider himself worthy of his trust, or he shall find that self-trust, like trust of another one considers unworthy it, will fail to support the undertaking.

By now you see the familiar schema revealing itself from my previous work, Lehrer (1997), for example. A person considers himself worthy of his

trust in what he accepts. So as he accepts whatever he does, even something appealing to premisses intended to support skepticism about evidence and justification, he will accept that he is trustworthy in what he accepts to achieve the goals of reason. If he is right, and he is trustworthy in what he accepts to satisfy the demands of reason, then he is reasonable in what he accepts. His acceptance of his trustworthiness, supported by other things he accepts, supports his acceptance of them in a loop of mutual support. Finally, his acceptance of his trustworthiness, supported by other things he accepts, supports the acceptance of itself. Of course, the person must be trustworthy in what he accepts and not deluded as the mad sometimes are. But if his acceptance of trustworthiness in what he accepts is true and leads to the conclusion that he is reasonable in what he accepts. Thus, the loop of reason ties his acceptance of his trustworthiness into the knot of reasonableness and, I have argued, Lehrer (2000) finally into justification and knowledge.

6. Truth and Justification

Justification and knowledge result once we can compare the reasonableness of what a person accepts to the objections against it. The strong arch of justification and knowledge comes into view with self-trust and the worthiness of it holding the components in place. Justification must be successfully connected with truth, however, to convert to knowledge. So how do we get truth into arch of justification? Justification built on acceptance and replies to objections may fail the test of veracity. Here I want to depart from Reid a bit and note a connection with Hume. The question is whether when we engage in self-trust in accepting how we represent our world and ourselves can we find any security against error. I suggest that it is impressions, the individual qualities of experience, and what are now called tropes, that can provide us with the security we seek. However, the security will not be found in appeal to properties or predicates connecting impressions with truth. Instead, we must note, what Hume did, that the individual quality may itself become the general vehicle of representation used as an exemplar to represent a plurality. In a process of exemplar representation, what I, Lehrer (2012), have called *exemplarization*, the exemplar, is used to represent a plurality of objects reflexively representing itself. The exemplar exhibits what it represents, and exhibits itself in the process. So, the exemplar, which is an impression, also represents impressions, it is true of what it represents, and, in a loop, represents and is true of itself.

To obtain the truth security of exemplarization, it is essential that exemplar reflexively represents itself. To say the representation is *reflexive* is to deny that it represents itself by virtue of instantiating some property. Both Hume and Reid denied the existence of properties, though the latter thought it was useful to think about properties. He noted that thinking about them and even affirming things of them did not commit you to their existence any more than thinking about fictional characters and affirming things of them commits you to their existence. Moreover, it is essential to notice that exemplarization is a cognitive process, and though reflexive when it operates successfully, has the fallibility of human representation. What we can say is that exemplarization, when it occurs, has the result that the exemplar is reflexively true of itself at the same time that it is less directly true of other things. The truth security of exemplarization is restricted to the application of the exemplar to itself.

7. Exemplar Impressions of the External World

Where do we stand with common sense? We have found a truth connection for an operation of a faculty in exemplarization. It shows us that as we engage in world making, in the conceptual construction of our world, we may find a secure connection with truth in impressions. But so far, this does not take us beyond the impressions and ideas of Mr. Hume to the common sense world of external objects and their qualities. Can we proceed from the truth security of exemplarization of impressions back onto themselves to the justification of the truth of claims about external qualities and their objects?

Here is a beginning of a development of theory beyond what I have so far advocated. The exemplar impression, on Hume's account, is used to represent a class of impressions. It is reflexively a member of the class, but the reference of the impression extends beyond the exemplar itself and stands for, represents and refers to other impressions. This Hume admits. Hume must agree that the exemplar as a sign representing other impressions extends beyond itself. That is part of the meaning of a sign used in exemplar representation. It exhibits what the things it represents are like. Note, and this is crucial, the relationship between the exemplar sign and what it represents is not conventional. The conventional sign of an object does not exhibit or in any way show us what the represented object or class of objects is like. The black word "red", for example, does not in any way show us what red things, or red impressions, for that matter, are like. The exemplar, by contrast, represents things by showing us what they are like.

This role of the exemplar to represent, stand for or refer to objects by showing us what they are like makes the exemplar part of its meaning or content. In a functional theory of meaning, the exemplar functions in the meaning of the sign as an exhibit of what the represented objects are like. Once this observation is made, we understand the importance of Reid's reply to Hume's argument that we do not perceive external objects because the appearances of the external objects change. Reid's reply is that the changing appearances of the object show us what the external object is like. Indeed, those changing appearances show us what the external object is like by serving as a sign of what the unchanging external object is like. Think about how the external object looks, perhaps how it appears smaller as it recedes into the distance as you move away from it. Those appearances become exemplars of meaning exhibiting what the external object and the qualities of it are like. The sensory impressions do not exhaust our conception of what the object is like, for part of that conception is of an object that has a constant size and shape. However, another part, and an important part, of our conception of the object and its qualities is how it looks. The exemplar impressions of the external object exhibit the exemplar part of the meaning. That is another way of saying that we conceive of objects in terms of how they look to us.

8. Phenomenalism and Realism : Meaning and Evidence of Sense

The conclusion of the preceding reflections is that our sensory impressions are exemplars of representation exhibiting part of the meaning of our representation of the external world. They do not give us the whole of the meaning. Phenomenalism was a mistake generated by the correct insight that phenomenal impressions are part of the meaning. But they are not all of it. Just as the phenomenalists were wrong to think that sensory impressions exhibit all of the meaning of our conception of the external world, so the direct realists were wrong when they thought that impressions were no part of the meaning of our conception of external objects. The mere fact that an impression can represent other impressions already entails that an impression can represent something beyond itself. Of course, the phenomenalist thought that a sensory impression could not show us anything about what a thing of another kind is like. But where is the argument that the appearances of a material object are so different in kind from the object that they cannot show us anything about what the object is like ? Where is the argument that an appearance cannot be

sign of the existence of an entity unless the entity is itself another appearance ? There is no sound argument, and the assumption creates a chasm between appearance and reality that is closed by the meaning of exemplar representation. Exemplars, like other signs, words, for example, permit multiplicity of meaning and reference. So just as the exemplar impression exhibits what the class of represented impressions is like, though distinct from them, so the exemplar impression exhibits what the external objects are like, though distinct from them.

I shall not undertake to argue further here that sensory experience serves as exemplars of meaning of our representations of external objects. This amounts to no more than that part of the meaning of our thoughts about the external world is in terms of how external objects appear to us, that, for example, that part of meaning of our thoughts about red things is how red things look to us. I do not say this is beyond controversy, only that, as Reid argued, it is a reasonable assumption that Hume and phenomenalists have not refuted.

How do we proceed from idea that that impressions of sense exhibit part of the meaning, the exemplar part of representation, to the conclusion that we have the sort of evidence and justification that we need for knowledge of the external world ? If an impression of sense is part of the meaning of our thought about external objects, then the impression provides evidence and justification for the existence of the external object. We must, of course, concede, as Reid did, that we are fallible. However, the fact the exemplar impressions of meaning of the existence of the external object is fallible is compatible with it being trustworthy and not fallacious. In short, the evidence of sense, the evidence of sense impressions, may give us evidence of the external world that is as reasonable as our faculties allow. Our justification depends on the evidence of sensory exemplars exhibiting to us a central part of the meaning of our common sense conception of the external world.

9. Back to Moore

Where does this leave us in the discussion with Moore with which we began ? We should agree with Moore that he knew what he said he knew. We now, however, are in a position to explain how he knew. He knew because the sensory experiences he produced by shaking his hands at us are part of the meaning, the exemplar meaning, of the conception of those things, and, therefore, evidence and justification for what he claimed to know. However, this is, contrary to Moore's claim of proof, not a proof against the skeptic, be-

cause it begs the question against the skeptic. Moore needed to assume that his faculties, in this case perception, were trustworthy and not fallacious. So, Moore was partly right and partly wrong. Moore knew what he said he knew, but what he argued was not, contrary to his claim, a proof of the existence of external objects. We have an explanation of what he knew and how he knew in terms of the self-trust placed in his faculties and the meaning of sensory experience in terms of exemplar representation. It is notable, and perhaps ironic, that Hume introduced the idea of an impression being used to stand for things beyond itself that explained, contrary to what he concluded, how we can know of the existence of external objects from the evidence of the changing appearances of the unchanging external object.

Of course, to get from meaning, evidence and justification to knowledge, our evidence and justification, though fallible, must not be refuted or defeated by errors in what we accept. That takes us beyond Hume and Reid to 20th century epistemology, Lehrer (2000), but the insights they offered us remain. Hume taught us the lesson of exemplar representation of impression of sensory experience, while Reid taught us the lesson of epistemological primacy of trust in our faculties. Reid said that Hume was the genius of the age, and of any age, noting, however, that it is genius and not the want of it that leads to false philosophy. The combination of genius and common sense, Hume and Reid, leads us to an explanation of our knowledge of the external world, and, I have argued elsewhere, Lehrer (2012), to knowledge of the world of theoretical entities as well. The legacy of Hume and Reid is the construction of an epistemology combining their genius and common sense.

10. Références

Engel, P. (2000) *Believing and Accepting*. Dordrecht : Springer Publishing.

Hume, D. (1739) *A Treatise of Human Nature*. London : John Noon.

Lehrer, K. (1971) "Why Not Skepticism ?" *The Philosophical Forum*, 206-98.

Lehrer, K. (1997) *Self Trust : A Study of Reason, Knowledge and Autonomy*. Oxford : Clarendon Press, Oxford.

Lehrer, K. (1998) "Reid, Hume and Common Sense," *Reid Studies* 2(1), 15-26.

Lehrer, K. (2000) *Theory of Knowledge : Second Edition*. Boulder : Westview Press.

Lehrer, K. (2012) *Art, Self and Knowledge*. New York : Oxford University Press.

Moore, G. E. (1925) "A Defense of Common Sense," in *Contemporary British Philosophy (2nd series)*, J. H. Muirhead, ed. London : Allen and Unwin.

Reid, T. (1863) *The Philosophical Works of Thomas Reid, D. D.*, 6th edition, editor Sir W. Hamilton (Edinburgh : James Thin).

11

Engel on pragmatic encroachment and epistemic value

DUNCAN PRITCHARD

Abstract. I discuss Engel's (2009) critique of pragmatic encroachment in epistemology and his related discussion of epistemic value. While I am sympathetic to Engel's remarks on the former, I think he makes a crucial misstep when he relates this discussion to the latter topic. The goal of this paper is to offer a better articulation of the relationship between these two epistemological issues, with the ultimate goal of lending further support to Engel's scepticism about pragmatic encroachment in epistemology. As we will see, key to this articulation will be the drawing of a distinction between two importantly different ways of thinking about epistemic value.

Introduction

Let me begin by saying that it is a pleasure and an honour to be able to contribute to this *festschrift* for Pascal Engel. In a long and highly distinguished career, Pascal has made distinctive contributions to many of the most important philosophical debates. He has also been an active and prominent member of the European philosophical scene. I know that I have learnt a lot by engaging with his work and with him personally over the years, and I am very pleased to be able to contribute to this volume in tribute to the man on his sixtieth birthday. Pascal is now at the peak of his intellectual powers, and long may he continue ! (As Woody Allen is reported to have quipped on his sixtieth birthday : 'I'm sixty years old—a third of my life is over already !' I have similar optimism for my friend Pascal's longevity).

It is customary in these volumes to follow one's eulogy to the person being honoured with a devastating critical broadside against his or her work. I'm afraid that I must disappoint the reader on this score, as no such broadside is in the offering here. This is because I am broadly in agreement with much of what is to be found in Pascal's work. Instead, I want to focus on a very interesting recent piece by Pascal which is concerned with the question of pragmatic encroachment in epistemology. I will argue that there is an interesting way of developing Pascal's position in this regard. As we will see, key to this development will be the introduction of a distinction regarding epistemic value which I think is both extremely important but also often overlooked.[1]

1. Engel on Pragmatic Encroachment and Epistemic Value

Pragmatic encroachment in epistemology is best understood in terms of what it rejects. In particular, it is usually understood as the rejection of the widely held view (until quite recently anyway) that whether an agent counts as having knowledge is purely a function of epistemic factors, and not determined, even in part, by non-epistemic factors (such as the practical consequences of having knowledge).[2] Jeremy Fantl and Matthew McGrath describe this view as *epistemological purism*, and express it as follows :

[1] In the interests of maintaining at least the appearance of scholarship, I will henceforth refer to Pascal as 'Engel'.

[2] I am here focusing on pragmatic encroachment about knowledge, specifically, though of course there are versions of the pragmatic encroachment thesis which apply to other epistemic standings.

> *Epistemological purism* : two subjects alike with respect to their strength of epistemic position with respect to *p* are alike with respect to whether they know that *p* (or at least with respect to whether they are in a position to know that *p*). (Fantl & McGrath 2010, 562 ; Cf. Fantl & McGrath 2007, 558)

Fantl and McGrath reject epistemological purism and argue that two subjects alike in their epistemic position might nonetheless differ in terms of whether they have knowledge in virtue of non-epistemic (e.g., purely practical) features of their situation. They are not alone in arguing for this claim.[3]

The cases marshalled in support of pragmatic encroachment in epistemology are now familiar. They characteristically involve two agents who are putatively in the same epistemic situation—they have the same overall evidence, say—but where the agents are in very different conditions from a practical point of view. So, for example, both agents have the same evidence about when a certain train will arrive, but whereas nothing much hangs on the correctness of the target belief for the one agent, a great deal hangs on its correctness for the other agent (the agent's livelihood, say). The thinking goes that we are less inclined to attribute knowledge to the second agent (the one in the 'high-stakes' context), and that this reveals that there is something amiss with epistemological purism, in that non-epistemic factors—in this case purely practical factors—are having a bearing on whether an agent counts as having knowledge.

I don't want to get into such cases in detail here. My view, which broadly accords with Engel's (2009), is that we should not take our intuitions about these cases at face-value. More precisely, while I would grant that we do feel a *prima facie* pull to treat these two agents differently *vis-à-vis* their possession of knowledge, even despite their putative sameness of epistemic standings, I think there are better explanations available of why this is so.

For example, it seems plausible to me that conversational contexts might affect the propriety of knowledge ascriptions. This idea is particularly compelling when it comes to self-ascriptions of knowledge. In a conversational context where it is made clear that a lot hangs on the correctness of *p*, an unqualified claim to know that *p* might conversationally imply that one is in a particularly strong epistemic position with regard to *p*, one that is far higher

[3] For the main defences of pragmatic encroachment, see Fantl & McGrath (2002 ; 2007 ; 2009), Hawthorne (2004), and Stanley (2005). For a helpful survey of recent work on pragmatic encroachment, see Fantl & McGrath (2010).

than what one would typically demand for knowledge that p. This would explain one's reluctance to self-ascribe knowledge in such conditions in a way that is entirely compatible with epistemological purism. And once this point is granted about self-ascriptions of knowledge, it doesn't take too much imagination to see how this detail might have a bearing on our intuitions about knowledge ascriptions more generally. In particular, if we explicitly set to one side the question of whether it would be appropriate for our agent in the high-stakes to make an unqualified knowledge claim, and focus instead on whether this agent counts as having knowledge (bearing in mind too that it has already been granted that the counterpart agent has knowledge), then what is left of the intuition that we should issue a negative verdict to this question? My guess is : 'not much'.[4]

And note that we have only considered one defensive response to the cases in support of pragmatic encroachment in epistemology. Properly developed, I think that a range of responses can be made to this proposal. In particular—and here Engel (2009) is especially clear—we also need to keep in mind that pragmatic factors can have a bearing on such matters as whether one forms a view at all about a certain proposition without this thereby having any negative implications for epistemological purism.[5]

In any case, let us not try to settle the issues about pragmatic encroachment in epistemology here. It suffices to say that the position is controversial, and that there are at least points to be made against this proposal. Engel and myself stand with the epistemological purists on these questions. What interests me for the purposes of this paper is a conclusion which Engel draws from this claim, and which I think should be resisted. The conclusion in question is that Engel argues—see especially Engel (2009, §5)—that once we grant that there is no such phenomenon as pragmatic encroachment on knowledge, then it follows that the kind of pragmatic factors appealed to by proponents of this view cannot confer any value on knowledge. I want to suggest that this is a mistake. Indeed, as we will see, my claim is that we can strengthen Engel's rejection of pragmatic encroachment by allowing the kind of pragmatic factors appealed to by proponents of this view as having a role to play in determining the value of knowledge. Essentially, my point will be that provided we are clear about the manner in which pragmatic factors can confer value on knowledge, then one can accept this claim without it having any bearing at all

[4]I discuss such conversational effects on knowledge ascriptions in Pritchard (2012*b*, part 3).

[5]See also Pritchard (2007*a*) for a different kind of response to the lottery-style cases that Hawthorne (2004) employs to motivate a version of pragmatic encroachment.

on whether epistemological purism is true.

2. Knowledge, Action, and Epistemic Value

In order to sharpen up our discussion in this regard, let's focus on the claim that knowledge has a kind of practical value in virtue of its role in action. In particular, this is the claim that, at least in some suitably restricted sense, it is knowledge, as opposed to true belief, which guides action and which therefore plays a pivotal role in practical reasoning. Versions of this kind of thesis have been defended by a number of prominent philosophers, and Engel is happy to endorse a version of this thesis too.[6] One might see in a thesis of this sort a direct argument for pragmatic encroachment, but given the foregoing it should be clear that this conclusion is at least resistible. As Engel himself puts the point, this conception of the relationship between knowledge and action :

> "[...] does not in any way show that there is pragmatic encroachment on knowledge, for it is quite open to someone to hold that knowledge is relevant to the explanation of action while denying that whether one knows that *p* turns on practical matters." (Engel 2009, 201)

I think that this is absolutely right.

Suppose, however, that we set aside the further claim about pragmatic encroachment and instead focus on the point about knowledge playing a fundamental role in action. Indeed, let us grant this point for the sake of argument. Ought it not to have axiological consequences for one's thinking about knowledge ? That is, shouldn't it follow from this thesis that knowledge is more valuable than mere true belief on account of the fact that only the former plays a fundamental role in action ? Engel is, however, quite explicit that one can't derive a claim about the greater value of knowledge over mere true belief by appeal to these factors. He writes :

> "[*I*]s knowledge more valuable than any of its subparts ? We would have the beginning of such an answer if it could be shown, for instance in the reliabilist way, that knowledge is apt to produce more true beliefs than sheer luck or absence of method, or if the way

[6] See, for example, Engel (2009, 199). For some of the main defences of this general view about the relationship between knowledge and action, see Williamson (2000), Fantl & McGrath (2002), Hawthorne (2004), Stanley (2005), and Hawthorne & Stanley (2008).

> in which knowledge matters could be associated to some specific dispositions of knowers, as virtue epistemology proposes. But the fact that our judgements about knowledge are relevant to our evaluation of actions, or that they are relevant for practical reasons, to repeat, shows nothing." (Engel 2009, 201)

I find this rather mysterious. Why do such 'facts' about the relationship between knowledge and action "show nothing" about the value of knowledge? Indeed, to put this point into sharper relief, why is that the kind of epistemological 'facts' that are attributed to reliabilists and virtue epistemologists here *can* confer value on knowledge but these other facts about the practical import of knowledge in action cannot?

I think the answer lies in a failure to recognise a crucial ambiguity in the very notion of epistemic value. With this ambiguity made clear, we can allow that there is a perfectly legitimate sense in which pragmatic factors, such as concerning the relationship between knowledge and action, can contribute to the value of knowledge. Moreover, the way in which they make this contribution offers no basis at all for endorsing pragmatic encroachment about knowledge.

3. Epistemic Value and the Value of the Epistemic

The ambiguity I have in mind can be brought out by considering a distinction that Peter Geach (1956) draws between 'predicative' and 'attributive' expressions. Consider the following two expressions :

(1) X is a red fly.

(2) X is a big fly.

According to Geach, (1) is a predicative expression while (2) is an attributive expression. What he means by this is that while we can re-phrase (1) as the claim that X is both red and a fly, it would be a mistake to rephrase (2) as the claim that X is both big and a fly. After all, the claim at issue in (2) is precisely that X is big *for a fly*.

In the same way, we can distinguish between a predictive and an attributive version of claims about epistemic value. On a predicative reading, this means that we are dealing with something which is both epistemic and of value. On an attributive reading, in contrast, this means that we are dealing with something which is valuable in a specific way—*viz.*, that it is of specifically

epistemic value. These are clearly distinct claims, as we will see. Henceforth, when we talk of 'epistemic value' we will mean a particular kind of value (i.e., we will presuppose the attributive reading), and we will refer to the predicative reading of 'epistemic value' by talking instead about 'the value of the epistemic'. With this in mind, let us now see how epistemic value comes apart from the value of the epistemic.

That something is epistemically valuable does not in itself mean that it is valuable *simpliciter*, any more than a big fly is thereby big *simpliciter*. Of course, it may be that there are bridging claims that one can bring to bear in this regard that make the necessary connection. Perhaps the epistemic axiological realm is such that it generates a kind of value which would sustain the predicative reading. There are precedents for this after all. For example, it is plausible that ethical value is both a kind of value and also value *simpliciter*—i.e., from the fact that something is ethically good one can plausibly infer that it is good *simpliciter*. Equally, however, there are also domains where this inference would be illegitimate. For example, that something is practically good does not mean that it is good *simpliciter*. In any case, absent a case being made for the relevant bridging claims, one cannot derive the value of the epistemic from epistemic value.

There is a similar distinction to be drawn in the opposite direction, from the value of the epistemic to epistemic value. Indeed, arguably the point here is even more straightforward : that an epistemic standing is valuable does not entail that it is of specifically *epistemic* value, since the value in question could be wholly non-epistemic (such as practical value, ethical value, aesthetic value, and so on). As before, some sort of bridging claim would be required to make the relevant transition, though here it is not particularly obvious how such a claim would be motivated. Why should the value of an epistemic standing entail epistemic value specifically ?

Once this distinction between epistemic value and the value of the epistemic is made clear, then I think we are in a position to understand why Engel's response to pragmatic factors having a bearing on the value of knowledge is too strong. In particular, there is nothing to prevent us from admitting that pragmatic factors, such as the relationship between knowledge and action, can add value to knowledge just so long as we are clear that when we talk of 'epistemic value' here we have in mind the predicative reading of this expression (i.e., the value of the epistemic, as we have characterised it above). That is, all we are saying is that knowledge has a value in virtue of these pragmatic factors that lesser epistemic standings, such as mere true belief, lack. But this value is not a specifically epistemic kind of value ; indeed, it is, presumably,

just the practical kind of value that it appears to be.

Engel is, however, quite right to resist the thought that these pragmatic factors generate specifically *epistemic value*. This would indeed be highly controversial and would imply that pragmatic encroachment about knowledge is true.[7] But in saying that knowledge has value in virtue of practical factors we are not making a claim about epistemic value at all.

Moreover, we can now explain why Engel maintains that reliabilism and virtue epistemology are able to offer accounts of knowledge which can explain (in contrast to appeals to the practical value of knowledge) the greater value of knowledge over its subparts. Since these are accounts of the value of knowledge which appeal to the nature of knowledge (i.e., its essential epistemic properties), I take it that Engel is quite naturally understanding them as making a claim which is specifically about the epistemic value of knowledge. In contrast, since appeals to the practical value of knowledge are not appealing to the nature of knowledge—particularly once it is granted that pragmatic encroachment about knowledge is false—such a proposal will not have a bearing on the value of knowledge in this sense.

Let us grant that this is the correct way to unpack Engel's reasoning in this regard. We might now ask : is Engel *right* to reason in this way ? I think not. With our distinction between epistemic value and the value of the epistemic in hand, it ought to be clear that in offering an account of knowledge which explains its value one is not thereby committing oneself to making a claim about the epistemic value of knowledge. In particular, it is at least an option that one's theory of knowledge explains the greater value of knowledge over its sub-parts by arguing that this value is exclusively non-epistemic. Indeed, I think that recognising this point is crucial to charting a way through the debate about the value of knowledge.

Consider the so-called 'swamping problem', for example, which is often alleged to show that knowledge cannot be more valuable than mere true belief.[8] Very roughly, this problem asks how knowledge can be more valuable

[7] Actually, I think that rather than lending support for pragmatic encroachment about knowledge, this claim would simply be incoherent. For pragmatic encroachment to even make sense we need a fairly clear sense of the distinction between epistemic and non-epistemic (e.g., practical) factors. If practical factors are now allowed to generate a specifically epistemic kind of value, then in what sense is this still pragmatic encroachment at all ? Haven't we instead just extended the realm of epistemic to take in factors hitherto considered non-epistemic ? This is not to say that such a view is unavailable, only that it is not best thought of in terms of pragmatic encroachment but as a different claim entirely.

[8] For more on the swamping problem, see Jones (1997), Swinburne (1999), Kvanvig (2003), and Zagzebski (2003). See also Pritchard, Millar & Haddock (2010, ch. 1) and Pritchard (2011).

than mere true belief given that we evaluate epistemic standings instrumentally in terms of their propensity to promote true belief. Just as a cup of coffee created by a 'good' (from a coffee-making point of view) coffee-making machine is no more valuable than an identical cup of coffee produced by a 'bad' (from a coffee-making point of view) coffee-making machine, why should we care whether a true belief is accompanied by an epistemic standing which is the mark of it being acquired via an epistemically good process?[9]

In fact, properly understood, this problem at most only demonstrates that on a particular veritistic conception of epistemic value—whereby the fundamental epistemic good is true belief—knowledge is not of greater *epistemic* value than mere true belief.[10] But that conclusion is compatible with the idea that knowledge is more valuable than mere true belief (i.e., where the additional value is of a non-epistemic variety). Accordingly, even if, for example, the reliabilist is committed to the relevant veritistic claim about epistemic value, they can still potentially tell a story about how the nature of knowledge is such that its epistemic properties ensure that knowledge is more valuable than mere true belief.[11] Perhaps knowledge is of greater value than mere true belief because of the greater practical value of reliably formed belief, for example?[12] It follows that one can explain the value of knowledge by appeal to the nature of knowledge without thereby making any claim about the greater epistemic value of knowledge over lesser epistemic standings.[13]

[9] The coffee cup analogy is due to Zagzebski (2003).

[10] The chief exponent of veritism is Goldman (1999; 2002), though a view of this sort is implicit in the work of a lot of key contemporary epistemologists. For further discussion of veritism, see Pritchard (*forthcominga*; *forthcomingb*).

[11] For more on this point, see Pritchard (2011; *forthcomingb*).

[12] Indeed, I think that the best responses that reliabilists offer to the question of the value of knowledge are essentially of this form (though to my knowledge they do not register the distinction between epistemic value and the value of the epistemic that I mark here). See Olsson (2007; 2009) and Goldman & Olsson (2009). For further discussion of reliabilism in this regard, see Pritchard (*forthcominga*; *forthcomingb*).

Note that the possibility that one's theory of knowledge can explain the value of knowledge by appealing to non-epistemic value is even clearer in the case of virtue epistemology. This is because of the general plausibility of the idea that intellectual virtues have broadly ethical value. Thus it could follow from the nature of knowledge that knowledge is of greater value than its sub-parts in virtue of its greater ethical value, even though it is conceded that knowledge is not of greater epistemic value than its sub-parts. For more on virtue epistemology and the value of knowledge, see Pritchard (2009*a*; 2009*b*) and Pritchard, Millar & Haddock (2010, chs. 1-4). See also Pritchard (2012*a*).

[13] I think that understanding this point also helps us to see why the claim that truth is the fundamental epistemic good is not nearly as problematic as it is (these days anyway) typically supposed to be. For further discussion of this claim, see Pritchard (*forthcominga*).

There could be another thought underlying Engel's reasoning here though. For one might think that there is something essentially contingent about explaining the value of knowledge by appeal to practical value, in contrast to explaining the value of knowledge by appeal to its essential epistemic properties. One can see the attraction of this idea. Whether or not knowledge has practical value will very much depend on the particular conditions in which it is possessed. In contrast, if one is appealing to the essential epistemic properties of knowledge in order to explain its value, then one is showing that it has this value regardless of the particular conditions under which this knowledge possessed. Despite the attraction of this idea, however, I think it should be resisted.

To begin with, we need to think a bit more about what it is we are trying to show when we say that knowledge is valuable. There are stronger and weaker theses that we might have in mind, along at least three axes. One axis, which we've just noted, concerns epistemic value *versus* the value of the epistemic. The claim that knowledge is valuable in both these senses (i.e., both epistemically valuable and valuable *simpliciter*) is on the face of it stronger than the claim that it is valuable in just one of these senses (e.g., just epistemically valuable). A second axis concerns the relevant contrast. Is the claim that knowledge is more valuable than mere true belief, or more valuable than its sub-parts, or more valuable in comparison to something else entirely ?[14] A third axis concerns the strength of the claim that knowledge is valuable. On a very strong reading this could mean that it necessarily always of value. But weaker readings seem available too. Suppose it were true that knowledge is generally the kind of thing that is of value to creatures like us (i.e., creatures in the sort of conditions that we tend to find ourselves in). Wouldn't that suffice to show that knowledge is valuable ?[15]

This third axis is particularly relevant to our current purposes. If one thinks that the intuition that knowledge is valuable is to be understood as the claim

[14] Elsewhere—see Pritchard (2007*b*), Pritchard, Millar & Haddock (2010, ch. 1), and Pritchard & Turri (2011)—I've referred to the value problem in terms of these first two contrasts as the "primary" and "secondary" value problems, respectively. See also endnote 15.

[15] There are other axes along which to cast the question of the value of knowledge. For example, one issue we haven't touched on here is whether knowledge has a distinctive kind of value that its sub-parts lack, such that the difference in value in play is not merely a difference of degree but of kind. (This is a problem that I've elsewhere called the "tertiary" value problem—see Pritchard (2007*b*), Pritchard, Millar & Haddock (2010, ch. 1), and Pritchard & Turri (2011)). Relatedly, one gets different versions of the value problem for knowledge by combining different axes : why is knowledge epistemically more valuable than mere true belief ? ; why is knowledge more valuable than its sub-parts ? ; and so on.

that knowledge is generally the kind of thing that is of valuable to us, then there need be no particular bar to supposing that contingent facts about knowledge—such that it generally has a certain practical utility—could underwrite its value.

Moreover, notice that even where one is appealing to essential features of knowledge to explain its value, it still doesn't follow that one is thereby undertaking the project of showing that knowledge is necessarily always of value. Reliabilism is a case in point in this regard. We noted earlier that it is open to the reliabilist to maintain that the explanation for why knowledge is more valuable than its sub-parts is that an essential epistemic property of knowledge—that it is true belief reliably gained—has practical value. But that's entirely consistent with the thought that such practical value is contingent on the nature of the circumstances that one has the knowledge in question.

The upshot is that theories of knowledge like reliabilism or virtue epistemology are not better placed to account for the value of knowledge than pragmatic accounts of the value of knowledge. The only difference in play here is that the former can explain the value of knowledge in terms of the essential epistemic properties of knowledge (something which is not available to the latter since it is not an account of knowledge). As we have seen, however, even that point is consistent with their explanation of the value of knowledge being in terms of non-epistemic value.

We are thus back to our original contention, which is that there is nothing inherently dubious about the idea that the value of knowledge might be attributable to purely pragmatic factors. As we have seen, one can accept this claim without conceding anything at all to pragmatic encroachment about knowledge.

4. Concluding remarks

Although I have here been critiquing something that Engel has argued, I hope it is also clear that this line of critique is one which is very sympathetic to Engel's general approach in this regard. What I have been arguing, after all, is that we can reject pragmatic encroachment about knowledge while nonetheless accepting that the kind of practical considerations which the proponents of pragmatic encroachment appeal to can have a role to play in explaining the value of knowledge. If anything, this is yet another count against pragmatic encroachment about knowledge, since in denying this thesis we are not led

into making claims about the value of knowledge that are otherwise contentious. In this sense, then, these critical remarks are in the spirit of Engel and myself being comrades against pragmatic encroachment in epistemology.

5. Références

Brady, M. S., *&* Pritchard, D. H. (eds.) (2003). *Moral and Epistemic Virtues*, Oxford : Blackwell.

Engel, P. (2009). 'Pragmatic Encroachment and Epistemic Value', *Epistemic Value*, (eds.) A. Haddock, A. Millar *&* D. H. Pritchard, 183-203, Oxford : Oxford University Press.

Fantl, J., *&* McGrath, M. (2002). 'Evidence, Pragmatics and Justification', *Philosophical Review* 111, 67-94.

— (2007). 'On Pragmatic Encroachment in Epistemology', *Philosophy and Phenomenological Research* 75, 558-89.

— (2009). *Knowledge in an Uncertain World*, Oxford : Oxford University Press.

— (2010). 'Pragmatic Encroachment', *Routledge Companion to Epistemology*, (eds.) S. Bernecker *&* D. H. Pritchard, 558-68, London : Routledge.

Geach, P. T. (1956). 'Good and Evil', *Analysis* 17, 32-42.

Goldman, A. (1999). *Knowledge in a Social World*, Oxford : Oxford University Press.

— (2002). 'The Unity of the Epistemic Virtues', in his *Pathways to Knowledge : Private and Public*, 51-72, Oxford : Oxford University Press.

Goldman, A., *&* Olsson, E. J. (2009). 'Reliabilism and the Value of Knowledge', *Epistemic Value*, (eds.) A. Haddock, A. Millar *&* D. H. Pritchard, 19-41, Oxford : Oxford University Press.

Hawthorne, J. (2004). *Knowledge and Lotteries*, Oxford : Oxford University Press.

Hawthorne, J., *&* Stanley, J. (2008). 'Knowledge and Action', *Journal of Philosophy* 105, 571-90.

Jones, W. (1997). 'Why Do We Value Knowledge ?', *American Philosophical Quarterly* 34, 423-40.

Kvanvig, J. (2003). *The Value of Knowledge and the Pursuit of Understanding*, Cambridge : Cambridge University Press.

Olsson, E. J. (2007). 'Reliabilism, Stability, and the Value of Knowledge', *American Philosophical Quarterly* 44, 343-55.

— (2009). 'In Defence of the Conditional Probability Solution to the Swamping Problem', *Grazer Philosophische Studien* 79, 93-114.

Pritchard, D. H. (2007*a*). 'Knowledge, Luck, and Lotteries', *New Waves in Epistemology*, (eds.) V. F. Hendricks & D. H. Pritchard, 28-51, London : Palgrave Macmillan.

— (2007*b*). 'Recent Work on Epistemic Value', *American Philosophical Quarterly*, 44, 85-110.

— (2009*a*). 'Knowledge, Understanding and Epistemic Value', *Epistemology (Royal Institute of Philosophy Lectures)*, (ed.) A. O'Hear, 19-43, Cambridge : Cambridge University Press.

— (2009*b*). 'The Value of Knowledge', *Harvard Review of Philosophy* 16, 2-19.

— (2011). 'What is the Swamping Problem ?', *Reasons for Belief*, (eds.) A. Reisner & A. Steglich-Petersen, 244-59, Cambridge : Cambridge University Press.

— (2012*a*). 'Anti-Luck Virtue Epistemology', *Journal of Philosophy* 109, 247-79.

— (2012*b*). *Epistemological Disjunctivism*, Oxford : Oxford University Press.

— (*Forthcominga*). 'Truth as the Fundamental Epistemic Good', *The Ethics of Belief : Individual and Social*, (eds.) J. Matheson & R. Vitz, Oxford : Oxford University Press.

— (*Forthcomingb*). 'Veritism and Epistemic Value', *Alvin Goldman and His Critics*, (eds.) H. Kornblith & B. McLaughlin, Oxford : Blackwell.

Pritchard, D. H., Millar, A., & Haddock, A. (2010). *The Nature and Value of Knowledge : Three Investigations*, Oxford : Oxford University Press.

Pritchard, D. H., & Turri, J. (2011). 'Knowledge, the Value of', *Stanford Encyclopaedia of Philosophy*, (ed.) E. Zalta, http://plato.stanford.edu/entries/knowledge-value/.

Stanley, J. (2005). *Knowledge and Practical Interests*, Oxford : Oxford University Press.

Swinburne, R. (1999). *Providence and the Problem of Evil*, Oxford : Oxford University Press.

Williamson, T. (2000). *Knowledge and its Limits*, Oxford : Oxford University Press.

Zagzebski, L. (2003). 'The Search for the Source of the Epistemic Good', *Metaphilosophy* 34, 12-28 ; and reprinted in Brady & Pritchard (2003), 13-28.

12

Engel on Knowledge and Assertion *

J. ADAM CARTER

Abstract Pascal Engel has insisted that a number of notable strategies for rejecting the knowledge norm of assertion are misguided ; the limited defence of the knowledge norm he offers does not go so far as to insist that the knowledge norm is *correct*. In a similarly qualified spirit, and without insisting the knowledge norm is *false*, I shall argue that a prevailing rationale for accepting the knowledge norm is misguided.

*Thanks to Pascal Engel for originally peaking my interest in norms of assertion during my time as a post-doctoral research fellow in Geneva. Thanks also to Mikkel Gerken and Emma C. Gordon for helpful conversation.

1. Engel on the Knowledge Account of Assertion

It is a great pleasure to contribute to Pascal Engel's *Festschrift* on the occasion of his 60th birthday. Pascal is really a remarkable philosopher—somehow, and I'm not sure how (an eidetic memory, maybe ?), he manages to keep abreast of current debates in nearly every area of philosophy (and, moreover, can talk about most papers as though he's just recently read them—maybe he has !). One particular topic I've had the opportunity to discuss with him in some detail is assertion, and in particular, assertoric norms, and so naturally, this is a topic I thought would be most fitting to explore here.

In his excellent paper 'In What Sense is Knowledge The Norm of Assertion ?', Engel begins by noting three components that, taken together, constitute the Knowledge Account of Assertion :

> *Knowledge Account of Assertion* (KAA) :
> (i) There is a norm for the speech act of assertion
> (ii) This norm is unique and constitutive of assertion
> (iii) This norm is that one must assert that P only if one knows that P

Whilst Engel's paper does constitute a kind of defence of the KAA, no where does Engel insist that the KAA is correct ; rather, he treads a more cautious line, which is to identify a variety of lines of criticism that have emerged against the KAA and to show how these criticisms are based on different kinds of confusions.

One such confusion has its source in the subtle differences between different articulations of the knowledge norm in (iii). Following Engel, let's call the basic formulation of the knowledge norm 'KN' :

> (KN) : One must : assert p only if one knows that p

KN is Williamson's (1996 ; 2000) presentation of the norm (and Engel's preferred formulation). But notice that KN is not equivalent to Moore's 'To assert that P is to *imply* that one knows that P' nor to Unger's 'To assert that P is to *represent* oneself as knowing that P[1]'. Consequently, commentators are missing the mark when overlooking that challenging these other articulations of the knowledge norm needn't be successful against the more basic KN.

Another problem Engel locates is the too-common failure to appreciate the *sense* in which assertion is subject to a norm. In particular, it is sometimes

[1] *Cf.* Pagin (2006) for a criticism.

overlooked that it is perfectly compatible with KAA that some *bona fide* assertions violate the norm, and that other norms or dimensions of evaluation of assertions come into play[2]. Engel at this point makes a number of analogies between assertion and belief. We can believe on the basis of bad reasons[3], but this point seems orthogonal to the matter of whether belief is subject to a particular norm of correctness (e.g. truth). As Engel notes : 'The fact that the norm can be in many cases violated, overridden by other norms, or be applied in a very loose and relaxed way in many conversational circumstances does not show that the norm is not in place.'

Although I do not think the KN or KAA are correct, I agree with much of what Engel has to say, and on reflection, this should not be surprising. This is because a major theme in Engel's work over the years has been normativism about belief, and in defending this position, Engel has taken care to show what does and does not count as a legitimate objection to this view[4]. Engel's rationale for defending normativist accounts of belief strikes me as broadly right, and so I am accordingly sympathetic to his drawing attention to ways in which KN and KAA have been resisted for the wrong reasons.

In this contribution my aim will be to register some ways that I think KN and KAA have been *accepted* for the wrong reasons. As Engel is not outright saying KN and KAA are *true*, I won't here be saying the position is false. Rather, I'll take a similar approach and explain why I think that some of the philosophical motivations for accepting these accounts rest on mistakes.

2. Uniqueness

A starting point to this end will be to examine more carefully a presupposition of condition (ii) in the KAA. Condition (ii), recall, states that the knowledge

[2] In particular, Engel thinks that such oversights explain why we some are taken to thinking Jennifer Lackey's (2007) selfless assertion cases—cases where assertions are claimed appropriate in the absence of the satisfaction of the belief condition—are a datum from which it should be concluded that KN and KAA are false.

[3] As Engel puts it, 'There is a clear sense in which a belief which is held for reasons which fall short of being epistemic – for instance a self deceptive belief or one which we aim to have to secure a form of comfort- still counts as a belief, so why could not assertions which are made for reasons which fall short of being epistemic, or which happen to be epistemically weak fail to count as assertions ?'

[4] See here, for instance, Engel (2007 ; 2013). In doing so, he's taken particular care to make it apparent how being subject to a norm needn't involve any positive avowal to conform to it (which is why desiring to hold false beliefs for pragmatic reasons, for instance, is not a datum that should lead us to think truth is not the standard of correctness for belief.) *Cf.* Shah & Velleman (2005).

norm is unique and constitutive of assertion. A presupposition of this condition is that there is one unique epistemic rule for assertion, and that such a rule will govern assertions uniformly. This is obviously the assumption in play in mainstream debates about the norm of assertion ; the rules of the game seem to be : *some* epistemic rule governs assertion ; it's plausibly either the truth norm (Weiner 2005), the justification norm (e.g. Douven (2006), Lackey (2007), Kvanvig (2009)), or the knowledge norm. The philosophical objective is to work out *which* one is right.One would think that the 'uniqueness' assumption is on safe ground, and that at the very least there is some (reasonably compelling) positive rationale for proceeding this way[5]. And even more, given Williamson's classic defence of this norm in his 1996 paper *Knowing and Asserting*, one would have expected that just such an argument for uniqueness would have come straight from him. Interestingly, as Jessica Brown notes, though "Williamson provides *no argument* for the assumption of uniqueness when he introduces it[6]" (Brown 2008 : 97) Williamson's reasons for accepting uniqueness are rather indirect.

> There might be several rules of assertion. There might be one ... Nevertheless, a simple account of assertion would be theoretically satisfying, if it worked." (Williamson 2000 : 242).

And because Williamson of course finds the knowledge norm to 'work', he thus takes himself to have reason to prefer a unique norm.

It is important to be clear here that the kind of support we find from Williamson for uniqueness does *not* rationalize the presumption we actually find in play in the literature—*viz.,* that if one norm is shown to (for instance) deal with problem cases *better* than the other two main contenders, then we should endorse that norm as 'the' norm of assertion. This 'last norm standing' approach simply bypasses the matter of whether we should be looking for just one norm in the first place. For reasons of simplicity, Williamson is probably right that a unique norm would be *ceteris paribus* more theoretically satisfying. But even if the KN can be argued to do *better* than weaker norms, such as Weiner's truth norm and the Lackey/Douven/Kvanvig justification norm, this falls short of a compelling reason to accept the uniqueness assumption.

[5] For discussion on this point, see also Carter & Gordon (2011).

[6] My italics.

3. Sufficiency

Engel and Williamson might point out here that if the KN is *not* the unique norm of assertion, it is not because one of the weaker norms is 'also' a norm of assertion. Engel quotes Williamson here, in response to cases of lying and selfless assertion (where assertion would appear proper despite a lack of knowledge), as remarking that :

> Such cases do not show that the knowledge rule is not the rule of assertion. They merely show that it can be overridden by other norms not specific to assertion. The other norms do not give me warrant to assert p, *for to have such warrant is to satisfy the rule of assertion*" (Williamson 2000 : 256), cited also in Engel (2004), my italics.

This reply is telling in two ways. Firstly, it suggests a kind of cook-book recipe for explaining away challenges to KN on the basis of being 'too strong' an epistemic norm. The recipe is to locate a non-epistemic norm satisfied in such cases (e.g. perhaps a Gricean or 'institutional norm',) and then to insist that the apparent propriety of the assertion is a matter of satisfying *that* norm, all whilst maintaining that 'the rule' would have been satisfied only were one to have knowledge. Setting aside questionbegging-worries (*vis-à-vis* the uniqueness assumption), notice that this reply reveals also the thought that whatever *epistemic* norm it is that uniquely governs assertion, it is one that would be satisfied were one to satisfy 'the' rule, which is knowledge. This is tantamount to an endorsement of knowledge as a *sufficient* epistemic credential for 'whatever the assertion rule is'.

As I've argued elsewhere[7], the idea that knowledge is *sufficient* as an epistemic credential[8] is something that operates in the background not only in the replies of proponents of KN in response to charges that the norm is too strong, but *also* in the thinking of those who endorse epistemically *weaker* unique assertion norms. This is revealed in *modus operandi* of asking 'how *much* epistemic strength' is needed to warrant assertion ? Given that knowledge is about as *much* as one could hope for, both critics and proponents of the KN and KAA implicitly maintain that 'the epistemic rule' is satisfied *if* one knows what one asserts.

[7] See Carter & Gordon (2011).

[8] The term 'epistemic credential' is Lackey's (2012).

4. Uniqueness, revisited

Reason to resist the sufficiency thesis *vis-à-vis* knowledge is at the same time reason to doubt that the KN 'works'. After all, if sufficiency fails, then sometimes there is a different epistemic rule in play, one *not* satisfied just by having knowledge. Recall that (*a la* Williamson) it was because KN 'worked' that we were entitled to think that a unique rule governs assertion. If the sufficiency thesis fails, then, there is a kind of undercutting defeater for the initial reason for ever having accepted uniqueness. And moreover—as I've stressed—it is precisely *because* uniqueness is presumed that writers are often lured into accepting the KNA because they claim it does *better* than certain weaker norms. In (1.) I presented my aim as to diagnose what I thought were some mistaken reasons for thinking that KN and KAA are correct, and now it should be clear both that (i) if uniqueness is wrong, then so is a popular rationale for accepting KN ; and (ii) if the sufficiency thesis is wrong, then all the worse for 'uniqueness'.

Why is the sufficiency thesis mistaken ? I think there are two arguments on this score, one that draws from Lackey's recent cases of what she calls 'Isolated Second-Hand' knowledge, and another that draws from cases that feature what I'll call 'epistemic hypocrisy.'

5. Against Sufficiency : Isolated Second-hand Knowledge

The very suggestion that knowledge might not be 'sufficient' for assertion might sound bizarre on first blush. After all, *what more could be expected* ? Consider, though, the following case Lackey offers, in her paper 'Assertion and Isolated Secondhand Knowledge' :

> DOCTOR : Matilda is an oncologist at a teaching hospital who has been diagnosing and treating various kinds of cancers for the past fifteen years. One of her patients, Derek, was recently referred to her office because he has been experiencing intense abdominal pain for a couple of weeks. After requesting an ultrasound and MRI, the results of the tests arrived on Matilda's day off ; consequently, all of the relevant data were reviewed by Nancy, a competent medical student in oncology training at her hospital. Being able to confer for only a very brief period of time prior to Derek's appointment today, Nancy communicated to Matilda simply that her diagnosis is pancreatic cancer, without offering any of

> the details of the test results or the reasons underlying her conclusion. Shortly thereafter, Matilda had her appointment with Derek, where she truly asserts to him purely on the basis of Nancy's reliable testimony, "I am very sorry to tell you this, but you have pancreatic cancer. (Lackey 2008 : 3-4)

Whilst Lackey offers other cases, I think this example is the most convincing[9]. As she puts it :

> The question we must now consider is whether, under these conditions, Matilda is properly epistemically positioned to flat out assert to Derek that he has pancreatic cancer. And here the answer is clearly no. (Lackey 2008 : 6).

One reason to balk at Lackey's negative answer here is to think that we're entitled to fall back on Williamson's 'cook-book' recipe for explaining such examples away, and in particular, that we might explain away the apparent *impropriety* of Matilda's assertion simply with reference to her failing to satisfy some other, non-epistemic norm (all while the unique epistemic rule, knowledge, is satisfied). Such a norm might be, for instance, an institutional norm according to which first-hand testimony is expected, in hospital settings, when the verdict is grave. But this tack won't work here, at least, not in this case.

To see why, let's distinguish between two kinds of institutional norms. Call the first kind *non-epistemic* institutional norms : for instance, in the ancient Roman republic, the institutional role of being an 'augur' of the republic carries the expectation that augurs declare the signs propitious on the occasion of the election of a new consul. This institutional role of the augur is of course is not that they say the signs are propitious *on the basis* of anything in particular ; *qua* augurs, they are just supposed to *say* it (and then the incoming consul will be pleased). But not all institutional roles that carry with them speech-act expectations are roles wherein the expectations are just that certain relevant speech acts be *made.* Some institutional roles have more refined epistemic expectations—*viz*., some institutional roles require that assertions be made on the basis of certain *kinds* of epistemic support. Call these *epistemic* institutional norms ; Matilda's assertion seems defective here because her

[9] Lackey (2012) takes the DOCTOR case to be an example of a wider phenomenon she calls isolated second hand knowledge ; on this she says 'There are two central components to this phenomenon : first, the subject in question knows that p solely on the basis of another speaker's testimony that p—hence the knowledge is secondhand ; and, second, the subject knows nothing (or very little) relevant about the matter other than that p—hence the knowledge is isolated.'

assertion to Derek should be made with a particular *kind* of epistemic support which she lacked, even though she possessed second-hand testimonial *knowledge*.

Now, is it fair to brush this case aside (as a proponent of the KN and KAA) as just another case that shows that some 'other norm' overrides (without counting against the truth of) the knowledge rule in this case ? I don't think so. But that's because I think we should reject the idea that *so long as* the impropriety of an assertion is in some way dependent on an institutional norm, that therefore, the impropriety isn't a datum that could ever count against KN. With such thinking in place, it would simply be *too easy* to explain away any potential counterexample to KN (inviting the charge of irrefutability) ; institutional norms after all govern all manner of assertions—it's not as though the practice of assertion is a practice out with the institutional norms that pervade it.

At any rate, to take seriously that DOCTOR does not count against the KN sufficiency thesis, an argument is needed—one which *doesn't reduce KN to an irrefutable thesis*—and yet can make sense of how KN 'works' in cases like DOCTOR, where the impropriety is *epistemic*. (Indeed, Matilda's assertion was improprietous because she failed to have a certain kind of epistemic support).

If no such explanation is forthcoming, we should just conclude that (alas) knowledge isn't always enough to warrant assertion ; sometimes there is another rule in play. I've suggested elsewhere that cases like DOCTOR are ones where a certain degree of explanatory understanding is the epistemic credential needed for assertion, where explanatory understanding doesn't reduce to propositional knowledge[10]. But the point that the sufficiency thesis is in trouble, however, doesn't rely on the argument that understanding in particular is what is lacking. The point here has been to show how it's hard (without already assuming KN, or endorsing it in a way that leaves it irrefutable) to explain away certain challenges that seem to pose a genuine problem for the supposition that the uniqueness thesis relies on—that KN 'works.'

6. Against sufficiency : epistemic hypocrisy

Like cases of isolated second-hand knowledge (like DOCTOR), cases of what I'll call 'epistemic hypocrisy' are cases where assertions satisfy the knowledge

[10] Carter & Gordon (2011). See also Carter & Gordon (2013) for some reasons to doubt that understanding reduces to knowledge. On this point, see also Pritchard (2009).

norm but are nonetheless *epistemically* criticisable. In cases of epistemic hypocrisy, the target assertion would (like in cases of isolated second-hand knowledge) be appropriate were one just to have had better epistemic support than one does when one asserts. Here's the idea :

> *Epistemic hypocrisy :* An assertor A's assertion p exhibits what I'll call *epistemic hypocrisy* if (i) A asserts p ; (ii) A wouldn't use p as a premise in A's own practical deliberation ; but (iii) would do so if only were A's epistemic support for p better.

Epistemic hypocrisy is I think a widespread phenomenon ; consider a plausible view about the function of assertion ; as Jessica Brown (2012) puts it, that one of the characteristic functions of assertion is to entitle hearers to rely on the asserted proposition in their practical reasoning (e.g. Brandom 1983 ; Milne 2009). If this 'licensing' view of the function of assertion is plausible, then epistemic hypocrisy is as widespread as the phenomenon of me giving you a premise to use in your practical reasoning, which I wouldn't act on unless my grounds were better.

Importantly for the present purposes, cases of epistemic hypocrisy can include cases in which individuals satisfy knowledge conditions *and* where the impropriety (as in cases like DOCTOR) is epistemic ; in DOCTOR and in cases of epistemic hypocrisy, the target assertions would not be criticisable so long as the epistemic support possessed by the asserter were stronger. And so I take it that if the argument in the previous section that (as Lackey had originally suggested) DOCTOR poses a problem for the sufficiency claim is right, then this also gives us reason to think that epistemic hypocrisy cases are trouble for the sufficiency thesis.

Obviously, the suggestion that epistemic hypocrisy of an assertion is relevant *vis-à-vis* whether one satisfies 'the' or 'a' epistemic rule of assertion commits me to thinking that something like the epistemic 'integrity' of an assertion bears on whether an assertion passes epistemic scrutiny. No objections here. We can think of the epistemic integrity of an assertion is a function not only of the epistemic grounds one has *simpliciter*, but also of the grounds one has relative to one's disposition to act on these same grounds. If those grounds are such that—by asserting on those grounds one's assertion is criticisable, but would not be were the grounds better—then the kind of criticism here is epistemic. But once the assertion is epistemically criticisable on such a basis, then if one has (say) second-hand testimonial knowledge of what one asserts, we have cases structurally akin to DOCTOR in so far as they are a problem for the

sufficiency thesis—*viz*., cases where the epistemic impropriety of an assertion is epistemic even though the knowledge norm is satisfied.

7. Concluding remarks

I've not argued that the KN or KAA is false any more than Engel has argued them correct. My goal has been, like, his diagnostic ; it is a mistake to search for an epistemic assertion rule by first supposing there is one unique norm, and that knowledge is the strongest of them. *If* the knowledge norm 'worked' then as Williamson noted, will have (of course) ourselves good reason to accept the uniqueness thesis. But as I've suggested here, challenges to the sufficiency thesis should lead us to worry that the knowledge norm does not work, *even if* it were to fare all-things-considered better than the leading weaker-norm candidates (e.g. the TN and the JBN). Given then that we shouldn't start out by taking for granted the uniqueness assumption, where does that leave us ? I think that where that leaves us is a spot where we should be more prepared to think that the epistemic rule governing assertion will not always be *uniform*. (This is tantamount to thinking outside the box of the uniqueness assumption). Recent work by Gerken (2013) has moved in this direction, and in particular, in the direction of supposing different epistemic rules will be in play in different contexts. This is messier than Williamson wanted—uniqueness is much cleaner—and it would be preferable if one unique norm 'worked.' But it's hard to see just how one would.

8. Références

Brandom, R. (1983). 'Asserting', *Noûs* 17.4 (1983) : 637-650.

Brown, J. (2008*a*) 'The Knowledge Norm for Assertion', *Philosophical Issues*, 18, Interdisciplinary Core Philosophy, 2008.

– (2008*b*). 'Knowledge and Practical Reason', *Philosophy Compass*, 3.6 1135–1152.

Carter, J. A., *&* Gordon, E. C. (2011). 'Norms of Assertion : The Quantity and Quality of Epistemic Support', *Philosophia*, *39*(4), 615-635.

– (2013). 'Objectual Understanding as an Epistemic State', (manuscript).

Douven, I. (2008) 'Assertion, Knowledge and Rational Credibility', *Philosophical Review* 2006 115(4) :449-48.

Engel, P. (2008). 'In What Sense Is Knowledge the Norm of Assertion ?', *Grazer Philosophische Studien*, *77*(1), 45-59.

– (2007). 'Belief and Normativity', *Disputatio*, *2*(23), 1-25.

– (2013). 'Doxastic Correctness'. *Aristotelian Society Supplementary Volume*, 87 : 199–216. doi : 10.1111/j.1467-8349.2013.00226.x.

Gerken, M. (2013). 'Same, Same but Different : the Epistemic Norms of Assertion, Action and Practical Reasoning', *Philosophical Studies*, 1-20.

Kvanvig, J. (2009) 'Assertion, Knowledge, and Lotteries', *Williamson on Knowledge*, eds Pritchard, D. & Greenough, P. (Oxford : Oxford University Press, 2009) pp. 140-160.

Lackey, J. (2008). *Learning From Words : Testimony as a Source of Knowledge*. Oxford University Press.

– (2007). 'Norms of Assertion', *Noûs*, *41*(4), 594-626.

– (2012). 'Assertion and Isolated Secondhand Knowledge', forthcoming in Jessica Brown and Herman Cappelen (eds.), *Assertion* (Oxford : Oxford University Press).

Milne, P. (2009). 'What is the Normative Role of Logic ?', In *Aristotelian Society Supplementary Volume* (Vol. 83, No. 1, pp. 269-298). Blackwell Publishing Ltd.

Pagin, P. (2006). 'Against Normative Accounts of Assertion' (draft).

Pritchard, D. (2009). 'Knowledge, Understanding and Epistemic Value', *Royal Institute of Philosophy Supplement 64*, 19.

Shah, N., & Velleman, J. D. (2005). 'Doxastic deliberation' *The Philosophical Review*,*114*(4), 497-534.

Williamson, T. (1996). 'Knowing and asserting', *The Philosophical Review* 105, 489-523.

– (2000). *Knowledge and its Limits*, Oxford : Oxford University Press.

13

Epistemic Justication, Normative Guidance, and Knowledge *

ARTURS LOGINS

Abstract. Recently, Pascal Engel has defended a version of a compatibilist view in epistemology that combines both an element of externalism and an element of internalism (Engel 2007, 2012). According to this position externalism has to be adopted about knowledge, whereas internalism has to be endorsed concerning epistemic justification. In this paper I argue that considerations that, allegedly, motivates Engel's internalism about epistemic justification, can be explained equally well, or, indeed, even better by a knowledge based externalist account of epistemic justification.

*This paper is dedicated to Pascal Engel. I would like to express my gratefulness to Pascal for his teaching and support. His work has made an important impact on me. In particular his Engel 2000, and his Engel 2007 contributed largely to my initiation to analytic philosophy and contemporary epistemology, respectively. The research work that lead to this article was supported by the Swiss National Science Foundation (SNSF) grant number 100015_131794 (project *Knowledge, Evidence, and Practice*).

1. Introduction

Recently, Pascal Engel has defended a version of a compatibilist view in epistemology that combines both an element of externalism and an element of internalism (Engel 2007, 2012). In short, according to this view, *knowledge* has to be characterized in externalist terms, whereas *epistemic justification and rationality* has to be characterized in internalist terms.

The externalist view about knowledge that Engel favours integrates a version of safety account of knowledge that requires that knowledge is safe belief and does not require that a subject has a reflective access to p in order for the subject to know that p (see Engel 2012 : 8). Where safety requirement (which is not to be conflated with a necessary and sufficient conditions for knowledge) that Engel has in mind is one defended by Timothy Williamson. According to Williamson's account of safety "a belief P is safe if the subject S could not easily been wrong in similar cases" (Williamson 2000 : 124, see Engel 2012 : 4).

The internalist element that Engel aims to accommodate in his account is the view that possession of epistemic reasons has to be understood ultimately in an internalist sense. According to a paradigmatic, or traditional internalism, when a subject is justified in believing that p, she has to have some kind of availability or access to reasons *for* p. Engel accords to the traditional internalist understanding that there is an intuitive force in supposing that it is in a sense essential to (epistemic) reasons for (believing in) a proposition that these reasons are available to the subject who possesses them (cf. Engel 2012 : 1, 9). Nevertheless, Engel demonstrates that internalist access requirement is implausible, because it implies a kind of vicious regress of epistemic support. In short the objection goes as follows : first, we acknowledge the following legitimate question - once, you have an access to a reason r for a proposition p, why shouldn't you also be required to have an access to the support relation that obtains between r and p in order to be justified in believing that p ? Then, we observe that the same question can be iterated, and so on *ad infinitum*. But such an access requirement is too demanding, for it seems it cannot stop the vicious regress in a non *ad hoc* way. Hence, the conclusion follows - internalist access requirement is implausible.

Despite the problems of access internalism, Engel thinks that a version of internalism is true. According to Engel, a sort of *sensitivity* to epistemic reasons counts also as possession of epistemic reasons. In Engel's view, this sensitivity to reasons is best understood in a specific internalist (quasi-externalist) way. Engel accepts a broad sense of epistemic reasons. According to this sense, epistemic reasons include epistemic norms (such as the normative principle of

correctness for belief : "A belief that p is correct if and only if p is true" (Engel 2012 : 8), for instance). In order to have a reason for a belief, agent has to possess that reason. If epistemic norms are epistemic reasons, then they also have to be possessed by subjects. It seems reasonable to think that in order for an agent to possess N as a norm, she has to be guided by N. The internalist element in Engel's account comes from his commitment to a sort of internalism about normative guidance. For Engel seems to assume that the requirement of normative guidance can be understood only as an internalism-compatible requirement (see for instance Engel 2012 : 9).

The aim of the present paper is to argue that while Engel is right to give justice to our intuitions about possession of reasons and normative guidance, he is mistaken in endorsing a quasi-externalist rather than a full blown externalist account of epistemic justification. For, I will argue, *contra* Engel, one can be externalist about knowledge, externalist about justification or rationality, *and* still accept that one has to be guided by a norm in order to possess it as reason. Notably, I will argue that the crucial intuition according to which one has to have some kind of sensitivity to epistemic norms in order to have justified belief can and, indeed should, be accounted in terms of knowledge.

In what follows I will, first, present in more details Engel's view and arguments that he proposes for his compatibilist position. I will, then, argue that a purely externalist account can also explain all the data - the intuitions that Engel puts forwards as main reason for accepting a compatibilist position. More specifically, I will argue that knowledge based account of normative guidance can deal with all the relevant intuitions. Moreover, I will claim that knowledge based account of normative guidance is even more plausible than other accounts. This, in turn, will authorize us to endorse a purely externalist position in epistemology. Third, I will claim, that although this resulting position is not faithful to the letter of Engel's account, it is still faithful to the spirit of Engel's approach, so to say. It is faithful to Engel's approach, for it does not conflict with the kind of rationalism that Engel seems to be favourable to.

2. Engel on internalist requirements

In his recent paper on knowledge and reasons (Engel 2012), Pascal Engel has advocated a compatibilist view in epistemology. Engel characterizes the view in the following way :

> "The view suggested here is a form of epistemic compatibilism about knowledge. It combines externalist elements - since it allows

> a definition of knowledge as ungettierized safe belief, and does not require access - with internalist elements - since beliefs have to be sensitive to reasons and to epistemic norms." (Engel 2012 : 8).

In short, the compatabilism that Engel endorses is a conjunction of (i) a version of safety account of knowledge and (ii) a version of internalism about reasons and epistemic norms. In what follows I will be concerned with (ii). Ultimately, I will suggest that there is a plausible externalist account of justification that can deal with the data that Engel takes to support (ii). If I am right, then a full blown externalism is preferable, since it is an unified position in epistemology. Other things being equal, a unified theory should always be preferred, since it is theoretically more simple.

Before considering my argument for a full blown externalism, let's consider, first, Engel's account and motivation for (ii). A crucial element in his internalism is assumption about sensitivity to reasons and to epistemic norms of believers. In short, there is a requirement of sensitivity to reasons and epistemic norms that a subject has to satisfy in order to be justified in her belief, according to this assumption. This sensitivity, according to Engel is to be understood in some kind of internalist terms. Hence, a version of internalism about epistemic justification, namely, what he call "quasi externalism", is true, according to Engel. In the remainder of this section we will specifiy in more details what is this sensitivity to reasons and how view about justification that is based on it differs from other forms of internalism.

Engel on access, epistemic reasons, and norms Traditionally, internalist requirements for epistemic justification have been understood as requirements of a certain kind of access to *that* which justifies one's belief. Namely, an access to one's epistemic reasons or evidence, or justificatory basis. Pacal Engel distinguishes, very usefully, various kinds of traditional internalist understandings of this access requirement. Going from the weakest to the strongest, Engel, distinguishes : (first level) the requirement of "an awareness of our reasons and an access to them", where "the access can be only potential and need not be conscious", and even "mere sensitivity to reasons" would count as access ; (second level) the requirement of actual access to reasons where in order for a subject to have a reason she has to have an actual access to them "through reflective second-order beliefs" ; (third level) the requirement of ability to treat reason as reason, where internalists who endorse this understanding of the access requirement "require not only that the agent has reasons and has access to them, but also that he can be capable of treating them *as* reasons, by being

able to argue in favour of them, to deliberate about them, and to defend them against opposing view." (Engel 2012 : 6).

Engel observes that the main motivation for internalist views comes from the observation that we have to base our beliefs on relevant reasons in order for them to be epistemically justified :

> "The main motivation for the internalist requirements comes from the fact that the basing relation is naturally construed as a requirement upon the availability of a reason to the person who holds the belief : in order for one to have a reason or a justification in virtue of believing P *on the basis of* a reason R, one must believe that R supports P - because otherwise, one wouldn't count as basing one's belief that P upon R." (Engel 2012 : 5).

At the end of the day, however, neither of the traditional characterizations of access requirement will be accepted by Engel. To the contrary, Engel observes that traditional internalist accounts are all vulnerable to the objection from vicious regress.

The main argument that Engel considers against the views that require accessibility of reasons (of any of the three levels that he has distinguished), is the argument from regress. In short, according to this objection, if we accept the view that one's justificatory basis need to be accessible, then we are engaging in a vicious regress, since we also have to accept that we have to have access to the support relation that obtains between the basis for p and p itself. And so on *ad infinitum*. Such regress requirement is highly implausible. Therefore, it seems very implausible that we have to have access to the justificatory basis (reasons/evidence) in order to be justified.

Some internalists themselves tend to take this objection seriously and adapt their views in accordance. See for instance Smithies, forthcoming, who restricts his version of access internalism to propositional justification, on pain of implausible consequences of infinite regress for access internalism of doxastic justification[1]. Assessment of whether strategy that is used by Smithies is a plausible is not the aim of the present discussion, though.

[1]Where a propositional justification determines what a subject is justified in believing, independently of whether she actually believes it or not. Whereas doxastic justification concerns her actual beliefs, namely, whether a subject is doxastically justified in believing that p depends on whether the subject has proposition justification for p and whether she has actually based her belief in p on the right grounds. See for more details on this distinction Swain 1979, Korcz 1997 and many others. For the same distinction in a different terminology see the distinction between *ex ante* and *ex post* justification, in Goldman 1979.

Despite the failure of traditional internalist accounts, however, Engel still holds that a version of internalism has to be accepted. He accepts a kind of the internalist requirement without endorsing the internalist understanding of access requirement :

> "It is not the place to settle the dispute between internalism and externalism about epistemic reasons and justification. I shall only grant that the internalist requirements on reasons are well motivated, and that an externalist theory of knowledge has to take them into account anyway." (Engel 2012 : 5)

According to Engel's view, there is no requirement of having a reflective or even only conscious access to r in order to have r as one's epistemic reason for p. One has only to be sensitive to *epistemic norms*, such as truth norm of belief formation, for instance[2]. This is how what Engel labels "internalist requirement on reasons" has to be understood - it is not about (internalist) access to some propositional content, it is rather about subject's sensitivity to norms that govern belief formation. Hence, Engel states :

> "Such normative principles [e.g. as "A belief that p is correct if and only if p is true"] need not be explicitly before the mind of believers, nor do they need to figure in their doxastic deliberations as explicit prescriptions which they would have to follow consciously. Their cognitive status can remain largely implicit. They can nevertheless figure among our reasons to believe in a broad sense." (Engel 2012 : 8).

To resume then, according to Engel's broad sense of "epistemic reasons", epistemic norms also count as epistemic reasons. Epistemic norms, however, can remain implicit. That is, it is not required that one has an (internalist) access to them (of first, second or third level sort) in order to possess them. However, their possession is to be understood in internalist terms. For Engel seems to assume that sensitivity to norms is something that only internalism can account for. Hence, a version of internalism has to be accepted, according to Engel.

It is natural, however, to ask what exactly does internalism about sensitivity to epistemic norms means.

[2]Engel presents some of epistemic norms, discussed by a number of philosophers, such as Pollock and Cruz 1999, Boghossian 2008. These norms include, among others the following ones for instance : (Truth norm) "A belief that p is correct if and only if p is true", (Evidence norm) "A belief that p is correct if and only if it is based on sufficient evidence" (Engel 2012 : 7).

The sensitivity to epistemic norms, as Engel understands it, seems to be characterized by an "implicit guidance" by a norm that a subject has :

> "Even the most general norm for belief, the truth-norm (i) [the truth norm of the footnote 1] need not imply more than an implicit guidance. A familiar feature of belief is that it is transparent to truth – if one tries to figure out whether to believe that P, the best way to answer this question is to ask oneself whether P. This feature is enough to explain why we are sensitive to the truth norm (Shah 2003, Engel 2010). Although these epistemic norms have been most of the time invoked by internalist, we can understand them in a quasi-externalist sense." (Engel 2012 : 8).

The last sentence of the quote may lead to a confusion, if one takes the contrast between "internalist" and "quasi-externalist" to denote mutually exclusive positions. The underlying idea is that we should understand the sensitivity to epistemic norms not in traditional internalist accessibilist terms. It is not required that we have reflective or conscious access to these norms. However, Engel, maintains that implicit guidance is a sort of internalist requirement.

3. Internalism, guidance, and knowledge

In the previous section we have seen in some details what Engel's "internalist elements" of his compatibilist view are supposed to be. We have seen that Engel rejects traditional internalist requirement of reflective or conscious access to one's epistemic reasons. We have also seen that Engel advocates a view according to which in order to be justified one has to be sensitive to epistemic norms. The sensitivity in question has not to be understood in reflective or conscious access terms. However, the mere fact that it implies a kind of implicit guidance, makes it, according to Engel, an internalist requirement on epistemic justification.

In this section, I aim to challenge the assumption that implicit normative guidance constitutes an internalist requirement for epistemic justifiction. My view is that, if there is anything that is genuine implicit guidance by an epistemic norm, then it is fully compatible with the view that all the epistemic reasons (or evidence) that a subject has supervenes the subject's knowledge. Moreover, it seems that there are some reasons to think that a knowledge based account of normative guidance explains best some of the features of normative guidance.

Before we discuss my view, however, we should first consider why one would think that implicit guidance by an epistemic norm should be understood in internalist terms. In order to succeed in this task, it might be useful to ask ourselves what is implicit guidance. But before considering what is implicit guidance, we have to say something more about one central distinction that we have used only in an unspecified way until now. Namely, we have to specify in more detail what exactly internalism and externalism about epistemic justification amounts to.

A common way to distinguish internalism from externalism about epistemic justification in contemporary epistemology is to appeal to *non-factive mental states*. In short, any position that states that epistemic justification that a subject has of her beliefs, supervenes on her non factive mental states, is an internalist theory of epistemic justification. Where by non-factive mental states we understand states that do not entail the truth of their content (see for instance Wedgwood 2002a for this canonical understanding of non-factive mental states). Whereas an externalist theory of epistemic justification is any theory that deny internalism about epistemic justification. We can formulate this distinction more precisely in the following way :

Internalism about epistemic justification Necessarily, if two subjects, S_1 and S_2 are internally alike, then S_1 and and S_2 are equally alike with respect to what epistemic justification they have for their beliefs. (See, for instance, Bonjour 1999, Audi 2001, 2007, Wedgwood 2002a, Huemer 2001, Conee and Feldman 2004, 2008, Silins 2005).

Non-factive mental states Non-factive mental states include beliefs, seemings, apparent experiences, appearings, feelings, imaginings, desires, hopes, wishes, etc. These states have in common that they do not require the truth of their content (see Wedgwood 2002a).

Externalism about epistemic justification It is false that necessarily, if two subjects, S_1 and S_2 are internally alike, then S_1 and and S_2 are equally alike with respect to what epistemic justification they have for their beliefs.

One sort of externalism about epistemic justification is a view that endorses evidentialism about justification (in short, the view that justification is deter-

mined by one's evidence) and epistemicism about evidence (the view that evidence that one possesses supervenes on one's knowledge)[3].

Evidential Epistemicism Necessarily, if S_1 and S_2 are alike with respect to what they know, then they are alike with respect to what evidence they possess. (Cf. Williamson 2000).[4]

With these precisions in mind we can now turn to the question of normative implicit guidance and it's alleged implication of internalism about epistemic justification.

A good place for gaining a deeper insight about Engel's view about implicite normative guidance of epistemic norm is his views about the norm of belief. In short, according to Engel and other so called normativists (e.g. Wedgwood 2002b, Shah 2003, Shah and Velleman 2005), the correctness condition of belief (i.e. the condition that states : "For any P, a belief that P is correct iff P is true" Engel 2013 : 2) constitutes main and unique norm of belief. In a sense the correctness condition is constitutive of belief (see Engel 2013 : 3). This normativist understanding of the correctness condition is expressed by Engel in the following principle :

> "(NT) It is the *norm* of belief that one ought to believe that P if and only if P is true." (Engel 2013 : 3)

Normativist accounts of correctness of belief have met various objections. In response, normativists have defended their approach in subsequent work. Pascal Engel has largely contributed to this debate. Our aim, however, is not to enter into this debate here. Such task would take us much further than what we can discuss in the present work. We present the debate about normativism about correctness of belief only as long as it can help us to understand better normative guidance. Which in turn is indispensable for assessing properly Engel's compatibilism.

[3]In general, in this paper, we treat the question of possession of reasons (which we take to be equivalent to the possession of evidence or justificatory basis) as the question of epistemic justification. This, however, is not precise enough. One could coherently endorse internalism or externalism about justification without endorsing the corresponding view about possession of reasons. For one could think that reasons are not necessary or sufficient for justification. Such position, of course, is incompatible with evidentialism. For our purposes, however, this distinction is not crucial.

[4]Where knowledge does not supervene on one's non-factive mental states. Thanks to Julien Dutant for pointing to me this possibility.

The debate concerning normativism about correctness condition of belief has been partially a debate about normative guidance of epistemic norms. For, a prominent objection against normativism has been relying on the assumption that correctness condition cannot constitute a norm of belief because it cannot guide belief formation (see notably Glüer and Wikforss 2009). In a sense (NT) is, the objection goes, impotent and, hence, cannot be the norm of belief. The argument presupposes that in order for a principle to be a norm for someone, it should be able to guide the subject. Norm has to have, as it was famously put by Peter Railton, a normative force and a normative freedom (Railton 1999, see also Engel 2013 : 8).

Now, as it happens in philosophy, it comes out that it is notoriously difficult to say something uncontroversial and at the same time more substantial than that there is this necessary condition of normative guidance as normative force and freedom for any norm.

Recently, Peter Railton has proposed an insightful analysis of normative guidance (see Railton 2006). In particular, he has distinguished two substantial accounts of normative guidance. Where a substantial account has to identify "mental acts", "states of mind" or "attitudes" that underwrite normative guidance by a norm for a subject (see Railton 2006 : 13). According to one of the two views, the relevant mental *relata* underwriting normative guidance by a norm N of a subject S is *acceptance* of N by S, whereas according to the second it is *endorsement* of N by S. Where accepting is not the same as believing, even though it is ultimately depending on some beliefs (see Ralton 2006 : 20), and endorsement has to do more with subject's judgemental rather than psychological part of agency (see Railton 2006 : 23). At the end of the day, however, Railton does not endorse any of these two views as universal characterization of normative guidance. He judges that describing normative guidance as acceptance without identifying which mental relata underwrites acceptance is not sufficient for a substantial account of normative guidance (see Railton 2006 : 16). It seems reasonable to Railton that the relevant kind of mental state is not belief (Railton 2006 : 20). Nevertheless, he acknowledges that the relevant states that underwrites norm-acceptance (i.e. normative equivalent to doxastic acceptance), has to be belief-like. A natural candidate according to Railton for such state is endorsement (Railton 2006 : 20). However, Railton also argues that in certain cases it makes sense to describe subjects as not being normatively guided by their judgement, but rather by psychological aspects of their agency, such as their (moral) character for instance (see Railton 2006 : 31). Hence, it seems that endorsement is not the mental state that underwrites normative guidance neither. For, it does not account for all

cases of normative guidance. Instead, Railton proposes to accept a pluralism of mental *relata* that can underwrite normative guidance. In Railton's view, we should abstain from proposing a universal characterization of normative guidance. We should rather accept that what normative guidance is can be best explained "from inside-out", that is, by considering every particular case and every particular agent and her perspective. Hence, according to Railton :

> "No privileged attitude—of endorsement, acceptance, or identification—accounts for the role of norms in shaping our lived world and contributing to the reasons for which we act. Humble *internalization* of norms without the self's permission, approval, or identification, like humble acquisition of beliefs without the benefit of judgement or reflection, provides much of our substance as agents." (Railton 2006 : 31-32)

Independently of whether Railton is right in rejecting acceptance, endorsement any other unification and universal account of normative guidance, we can observe here one crucial point that seems to be accepted by many within that debate. Namely, in order for a subject to be guided by a norm, she has to internalize it in some way or another. That is, if a special connection between a given norm and central parts of one's agenthood has not been established, it is not the case that the norm guides the subject. It seems that the majority in the debate about normative guidance will accept this point. In difference to others, Railton only thinks that in terms of universal characterizations nothing more can be said about normative guidance. The rest of the picture about normative guidance has to be filled "from inside-out", according to Railton.

Crucially, however, from the fact that normative guidance via internalization of norms supervenes on some mental relata it does not follow that internalism about epistemic justification is true. There is no reason to think that the requirement of internalization favours internalism about epistemic justification over externalism. Why should we think that internalization of norms supervenes on one's non-factive mental states ? Indeed, if one is willing to grant, as is Railton and Engel with respect to epistemic norms, that normative guidance does not have to be explicit, that is, that an agent in order to be guided by a norm does not have to have reflective or conscious access to the norm, then it seems that there is no other independent reason to think that internalization of norms supervenes on one's non-factive mental states[5]. For

[5]Of course, I also think that there is no good reason for holding that internalization of norms supervenes on reflective or conscious access to norms. For as Williamson has shown, there is good

it is usually accepted that reflective and conscious access has to be understood in terms of non-factive mental states. And I don't see any other reason that could motivate the view that internalization of norms *has* to supervene on one's non-factive mental states.

It seems that common understanding of the term "internalization of norm" is that of acquisition by a subject of a deeply agenthood-impacting and strong connection between her and a norm. This also seems to be the core of the usage that Railton makes of this term. But acquisition of a deeply agenthood impacting and strong connection between a subject and a norm need not necessarily be underwritten by a non factive mental state. Hence, it seems that internalization of norms does not imply internalism about epistemic justification or about epistemic reasons.

Moreover, there is a reason to think that normative guidance supervenes on agent's knowledge. For a natural way of explaining what internalization of norms is, is to claim that it is a kind of learning. Surely, it is a special kind of learning, it is learning of norms, but it is learning nevertheless. But a common way to describe what learning is, is to characterize it as a kind of acquisition of knowledge (it is important for what will follow to notice that we don't claim that all acquisition of knowledge is learning). That is, when a subject learns that p, then the subject comes to know that p. If we are right about these two last assumptions, then it follows that when one internalizes a norm, one comes to know a norm. Therefore, internalization of norms is acquisition of knowledge.

Interestingly, in describing one particular case of an agent acquiring normative guidance by a norm, Railton himself refers to it explicitly as learning. He describes a case of a subject, Felicity, who comes from modest rural region, and has got a scholarship for attending an expensive college in a different region. She believes that her success depends on her being able to overcome her rural manners, and, as states Railton, on her ability to "generally learn to comport herself in accord with the Upper Middle Class Professional norms". Felecity learns the relevant norms and takes them to guide her everyday actions (Railton 2006 : 19). Hence, according to Railton, acquiring normative guidance is, at least in this case, underwritten by learning.

Now, it is important to notice that we allow to classify as learning not only acquisition of beliefs, but also acquisition of certain moral (and other) traits

reason to think that for any non trivial mental condition C, it is never the case that we can always know that we are in C. This applies also to reflective and conscious access. See Williamson 2000, 2007.

of character, for instance. This in turn commits us to the view that there may be different kinds of knowledge, or at least different kinds of acquisition of knowledge. But this assumption does not seems to be theoretically costly. Indeed, it seems plausible, independently of our discussion, to suppose that we can have theoretical knowledge, as well as practical knowledge (often referred to by the term "know-how"), and knowledge of norms)[6]. The existence of different sorts of learning and knowledge in turn seems to fit well with Railton's observation about internalization of norms. Namely, it fits well the observation that internalization of norms doesn't involve only the judgemental part of agenthood, but it depend also on psychological part. Knowledge based account of internalization of norms then may provide grounds for an unificatory account of normative guidance.

Furthermore, another reason that speaks in favour of knowledge based account of internalization of norms is that knowledge guarantees the stability aspect that is necessary for normative guidance. It is reasonable to think that normative guidance is stable. That is, when someone is guided by a norm, then she will be guided by it in various contexts. In particular, it is not easy for someone to lose a norm that she has acquired. Hence, a plausible account of internalization of norms has to pay sufficient attention to the stability feature of normative guidance. Knowledge, contrary to many non-factive mental states, possesses the desired stability aspect. In the sense that once someone has a bit of knowledge, she cannot easily lose it, all other things being equal. Hence, it seems that knowledge is the best candidate for guaranteeing the stability feature of normative guidance.

To conclude, internalization of norms does not imply internalism about epistemic justification, at least, as long as one is willing to abandon strong internalist accessibilism. Furthermore, there are also reasons to think that internalization of norms is underwritten by acquisition of knowledge. Hence, there are reasons to think that normative guidance supervenes on knowledge. But if we are right about this point, then it seems that there is no more motivation to endorse a compatibilism of the sort that Engel has proposed. For a more simple and hence theoretically preferable view is a full blown externalism (i.e. externalism about knowledge *and* externalism about justification). Moreover, there is no other plausible competing view (in particular a full blown internalism is not an option). Therefore, we should prefer a full blown externalism.

[6]Which is not to say that knowledge can be non-propositional. See Williamson and Stanley 2001, Stanley 2011.

4. Conclusion : Knowledge and Rationalism

In this paper we have presented and opposed Engel's comptabilism - the view that externalism has to be adopted about knowledge, whereas internalism has to be endorsed concerning epistemic justification. We have argued, that the main considerations that, allegedly, motivates Engel's internalism about justification, can be explained equally well, or, indeed, even better by a knowledge based externalist account of epistemic justification. The considerations that have motivated Engel to adopt internalism about epistemic justification concern normative guidance aspect of epistemic norms. Engel claims that our belief formation is subjected to epistemic norms. These norms have to be understood as part of subject's epistemic reasons. To have a justified belief, one need to possess epistemic reasons in favour of that belief, in particular, it is not enough that some reasons merely exist in favour of that belief. Possession of epistemic norms by a subject has to be understood in internalist terms, because of the nature of normative guidance, according to Engel. For one has to be guided by the norm in order for her to have that norm. And normative guidance, according to Engel has to be understood in internalist terms. We have shown, however, that there is no conclusive reason for thinking that normative guidance supervenes on one's non-factive mental states (according to paradigmatic statement of internalism about epistemic justification, possession of epistemic reasons supervenes on subject's non factive mental states). Moreover, there are good reasons for thinking that normative guidance supervenes on knowledge. If we are right then there is no independent motivation for Engel's internalism about epistemic justification. This authorise us to conclude that we have better to accept a full blown externalism, rather than compatibilism.

There rests, however, one last possible worry for our argument. A worry that can be also found in Engel 2012. According to this line of worry, a full blown externalism rules out a plausible version of traditional rationalism. In these closing remarks we consider briefly this objection and respond to it. In short, we think that a kind of rationalism is compatible with the knowledge based externalism.

According to a version of rationalism, norms are *a priori* relations. According to some views, having norms is even a prerequisite of agenthood (see Railton's discussion of kantian positions in Railton 2006). What is important for our purposes is that, according to this understanding of norms, norms are not learned. Our central argument against internalist view of normative guidance supposed that norms are learned. We claimed that internali-

zation of norms is learning of norms. But learning is acquisition of knowledge. Hence, to acquire norms is to acquire knowledge. Normative guidance does not supervene on non factive mental states, according to our conclusion, since, knowledge is not a non factive mental state. Now, if norms are not learned, but are rather accessed *a priori*, without any learning, or at least some norms are not learned, then our conclusion doesn't follow, it seems. For learning has nothing to do with having norms according to this picture. One may think that this is the kind of objection that Engel has in mind when he states that : "It is inconsistent with the notion of normativity to suppose that normative relations are ultimately purely factual. It is at this point that the classical concerns of the philosophers whom the philosophical tradition has called "rationalists" come back into the picture" (Engel 2012 : 9).

This objection fails to undermine our argument, however. For our argument can be restated in purely rationalist terms. It suffice to replace learning of norms, by *a priori* knowledge of norms. *A priori* knowledge is not a non factive mental state. Hence, even if norms are prerequisites and are accessed in *a priori* way, there is no conclusive reason to think that normative guidance supervenes on non factive mental states.

Engel states that :

> "One feature, however, of the traditional notion of reason, is resistant to a strong externalist conception : the epistemology of the relation of *being a reason for* and the kind of knowledge that we can have of epistemic reasons and norms seem to be purely *a priori*. Entitlement itself is also, on most views, an *a priori* status. It is inconsistent with the notion of normativity to suppose that normative relations are ultimately factual." (Engel 2012 : 9)

Indeed, Engel is right, norms contrast with facts. To say that something ought to be the case, is fundamentally different from saying that something is the case. However, it is not inconsistent with the notion of normativity to suppose that normative relations are known. Only a strongly empiricist externalism would be inconsistent with the notion of normativity, as Engel understands it. Only, an externalism that states that all knowledge comes from learning would be resistant to what Engel calls "the traditional notion of reason". We haven't based our argument on this kind of externalism, however.

Crucially, Engel himself states that the possession of epistemic norms is knowledge of norms, when he claims that "(..) the kind of *knowledge* that we can have of epistemic reasons and norms seem to be purely *a priori*" (*ibid*, my

italics). Engel talks about *a priori* knowledge, but *a priori* knowledge is a kind of knowledge. Hence, there is no reason to suppose that normative guidance supervenes on non factive mental states and not on knowledge, even if a kind of traditional rationalism is true.

We can therefore conclude that we have shown that our argument for a full blown externalism holds. And knowledge based account of normative guidance is compatible with and, indeed, friendly to the rationalist spirit of Engel's approach.

5. Références

Audi, R. (2001), 'An Internalist Theory of Normative Grounds', *Philosophical Topics* 29(1/2), 19–46.

Boghossian, P. A. (2008), *Content and Justification : Philosophical Papers*, OUP Oxford.

BonJour, L. (1999), 'Foundationalism and the External World', *Philosophical Perspectives* 13(s13), 229–249.

Conee, E. & Feldman, R. (2008), Evidence, *in* Quentin Smith, ed., 'Epistemology : New Essays', Oxford University Press, .

Conee, E. & Feldman, R. (2004), *Evidentialism*, Oxford University Press.

Engel, P. (2012), Knowledge and Reason, *in* Maria Cristina Amoretti & Nicla Vassallo, ed., 'Reason and Rationality', Ontos verlag, .

Engel, P. (2010), Epistemic Norms, *in* Sven Bernecker & Duncan Pritchard, eds., *The Routledge Companion to Epistemology*, London Routledge.

Engel, P. (2013), In Defence of Normativism About The Aim of Belief, *in* Timothy Chan, ed., 'The Aim of Belief', Oxford University Press, .

Engel, P. (2007), *Va Savoir : De la Connaissance En Général*, Hermann.

Engel, P. (dir.) (2000) *Précis de philosophie analytique*, Presses Universitaires de France.

Glüer, K. & sa Wikforss (2009), 'Against Content Normativity', *Mind* 118(469), 31–70.

Goldman, A. (1979), What is Justified Belief ?, *in* George S. Pappas, ed., 'Justification and Knowledge', Boston : D. Reidel, , pp. 1–25.

Huemer, M. (2001), *Skepticism and the Veil of Perception*, Lanham : Rowman & Littlefield.

Korcz, K. A. (1997), 'Recent Work on the Basing Relation', *American Philosophical Quarterly* 34(2), 171–191.

Pollock, J. L. & Cruz, J. (1999), *Contemporary theories of knowledge*, Vol. 35, Rowman & Littlefield.

Railton, P. (2006), Normative Guidance, *in* Russ Shafer-Landau, ed., 'Oxford Studies in Metaethics, Volume 1', Oxford University Press, .

Railton, P. (1999), 'Normative Force and Normative Freedom : Hume and Kant, but Not Hume Versus Kant', *Ratio* 12(4), 320–353.

Shah, N. (2003), 'How Truth Governs Belief', *Philosophical Review* 112(4), 447–482.

Shah, N. & Velleman, J. D. (2005), 'Doxastic Deliberation', *Philosophical Review* 114(4), 497–534.

Silins, N. (2005), 'Deception and Evidence', *Philosophical Perspectives* 19(1), 375–404.

Smithies, D. (forthcoming), Why Justification Matters, *in* John Greco & David Henderson, ed., *Epistemic Evaluation : Point and Purpose in Epistemology*, Oxford University Press.

Stanley, J. (2011), *Know How*, Oxford University Press.

Stanley, J. & Williamson, T. (2001), 'Knowing How', *Journal of Philosophy* 98(8), 411–444.

Swain, M. (1979), Justification and the Basis of Belief, *in* George S. Pappas, ed., 'Justification and Knowledge', Boston : D. Reidel, , pp. 25–50.

Wedgwood, R. (2002b), 'The Aim of Belief', *Philosophical Perspectives* 16(s16), 267–97.

Wedgwood, R. (2002a), 'Internalism Explained', *Philosophy and Phenomenological Research* 65(2), 349–369.

Williamson, T. (2000), *Knowledge and its Limits*, Oxford University Press.

14

Commodious Knowledge

CHRISTOPH KELP AND MONA MARICA

Abstract This paper offers a novel account of the value of knowledge. The account is novel insofar as it advocates a shift in focus from the value of individual items of knowledge to the value of the commodity of knowledge. It is argued that the commodity of knowledge is valuable in at least two ways : (i) in a wide range of areas, knowledge is our way of being in cognitive contact with the world and (ii) for us the good life is a life rich enough in knowledge.

1. Introduction

We care a lot about knowledge. As a society, we invest a lot of time and energy in the development of institutions whose aim it is to accumulate or distribute knowledge. Universities, schools, libraries and the internet are among the most prominent of these. On an individual level, we send our children to school and encourage them to go to university so that they can acquire knowledge about a wide range of topics. Some of us go to considerable financial lengths in order to make this possible. Finally, in philosophy, the study of knowledge has historically received a great deal of attention. A lot of effort has been made to get clear on what exactly is involved in knowing.

From a philosophical point of view, the fact that we seem to care so much about knowledge gives rise to a number of interesting questions : First, is our concern with knowledge warranted ? In other words, does knowledge have value that is *special* at least in the sense that it would warrant this concern ? Second, in what respect(s) exactly is knowledge valuable ? The aim of this paper is to provide novel answers to both of these questions.

2. The Value Problem

Three Challenges

What does it take to provide a satisfactory answer to the question whether knowledge has value that is special enough to warrant our concern with it? There are a number of proposed answers in the literature. They differ from one another in the strength of the demands imposed.

Let's begin with what is widely regarded as the most lenient proposal, which dates back as far as Plato's *Meno*. To begin with, notice that it is nearly universally accepted that knowledge requires true belief.[1] Now suppose it turns out that knowledge is in no respect more valuable than true belief. In that case it would seem wrong to care about knowledge rather than true belief. Our special concern with knowledge would seem misplaced. This motivates a first constraint on satisfactory accounts of the value of knowledge :

(1) Any satisfactory account of the value of knowledge must explain why knowledge is in some respect more valuable than *true belief*.

Some have claimed that simply meeting C1 won't be enough to give a satisfactory account of the value of knowledge. Jonathan Kvanvig, for one, argues that more is needed : Suppose knowledge consists of a set of constituents. Suppose, next, it turns out that knowledge is in no respect more valuable than some proper subset of its constituents. In that case it would be wrong to care specifically about knowledge rather than the proper subset of equally valuable constituents. Our special concern with knowledge would still seem misplaced. In view of these considerations, Kvanvig favours the following constraint :

(2) Any satisfactory account of the value of knowledge must explain why knowledge is in some respect more valuable than *any proper subset of its constituents* (Kvanvig 2003, xii–xiii).

Duncan Pritchard ups the stakes even further. He argues that a satisfactory account of knowledge must, in addition, show that knowledge enjoys a different kind of value that belief that falls short of knowledge. According to Pritchard, then,

[1] Notable exceptions are David Lewis (1996) and Colin Radford (1966) who have denied that knowledge requires belief and Allan Hazlett (2010) who has challenged the thesis that knowledge is factive.

(3) Any satisfactory account of the value of knowledge must explain why knowledge is in some respect more valuable than any proper subset of its constituents *not just as a matter of degree but as a matter of kind* (Pritchard *et al.* 2010, 7–8).

Riggs's Requirement

It has widely been taken for granted that the task of explaining the value of knowledge consists in showing that individual items of knowledge are more valuable than individual beliefs that fall short of knowledge. Wayne Riggs, for one, is very clear about this when he proposes that the best way to understand C2 is as follows :

> $(\forall s)(\forall p)$[Value (sKp) > Value (sRp)] (where R is some relation comprising elements of K, and $R \neq K$)
>
> (Riggs 2009, 334)

Roughly, the idea here is that in order to meet C2 we need to show that it is better for one to know *that p* than to have a belief *that p* that falls short of knowledge, for all propositions p.

While the challenge Riggs unpacks in the above quote is of course C2, it is not hard to see that C1 and C3 can be given the same treatment. In case of C1, for all propositions p, knowledge that p must be more valuable than mere true belief that p, and in case of C3 knowledge that p must have a different kind of value than belief that p that falls short of knowledge. Key to this way of fleshing out the challenges is that they require showing that *every item* of knowledge is more valuable than the corresponding belief that falls short of knowledge. (In what follows, we will refer to this requirement as 'Riggs's requirement'.)

Let's get one thing out of the way : We think that knowledge is valuable in a way that satisfies Riggs's requirement. By way of explanation, consider first the following distinction between two types of value : final and non-final.[2] For something to have *final* value is for it to be valuable for its own sake. For instance, it is widely acknowledged that happiness is valuable for its own sake. In contrast, for something to have *non-final* value is for it to derive its value from something else that is of value. There may be different species of non-final value. Our main focus will be on the most widely discussed species,

[2]For more on different types of value see e.g. (Korsgaard 1983 ; Kagan 1998, Rabinowicz & Rønnow-Rasmussen 2000) all of which are reprinted in (Rønnow-Rasmussen & Zimmerman 2005).

to wit, *instrumental value*. For something to have instrumental value is for it to have value as a means to an end. For instance, it is widely acknowledged that money is instrumentally valuable, at least in the kinds of society we live in : it allows its possessors to buy things that enable them to achieve a certain level of comfort in life. The reason we think knowledge is valuable in a way that complies with Riggs's requirement is that we believe that knowledge but not belief that falls short of knowledge is finally valuable. We take this to establish that every item of knowledge is more valuable than the corresponding belief that isn't knowledge.

That said, for the purposes of this paper, we'd like to bracket responses to the challenges in terms of the final value of knowledge and focus on responses in terms of instrumental value instead. Curiously, once we are clear that we are aiming for this kind of response, the prospects of success look somewhat dim. The reason for this is that, when focusing solely on instrumental value, some items of knowledge appear to be of no value whatsoever. (Let's call items of knowledge that have no instrumental value whatsoever items of 'useless knowledge'.) Knowing the exact number of grains of sand in the jar you brought back from last year's summer holiday is but one popular example (Sosa 2003). If some items of knowledge are useless, however, they are no more valuable than the corresponding beliefs that fall short of knowledge, which are useless as well. A satisfactory response to any of C1 – C3 appears no longer available.

On the other hand, recall that C1 – C3 are to be motivated by our concern with knowledge. That is to say, we wanted an account of the value of knowledge that reflects our concern with knowledge. If so, however, it is far less evident that a satisfactory account of the value of knowledge must satisfy Riggs's requirement. After all, our concern with knowledge does not appear to extend to all items of knowledge. In particular, we seem to have little concern for items of useless knowledge—and, we want to add, rightly so. If so, examples of useless knowledge are best understood as suggesting that a satisfactory account of the value of knowledge need not satisfy Riggs's requirement.

Once we abandon Riggs's requirement, we might think that it will be enough to rise to the challenges if we can show that *enough* items of knowledge are more valuable than the corresponding beliefs that aren't knowledge. While we do not mean to deny that this is a promising line to pursue, in what follows we'd like to explore a different approach.

3. The Value of Knowledge

Commodity Value

'Knowledge' is a mass term, like 'water'. It is widely agreed, however, that mass terms denote stuff that can be measured but not counted.[3] In the case of knowledge and water the stuff is a kind of commodity—something one can have more or less of. Now suppose that it can be shown that the commodity of knowledge has special value, that an account of the value of the commodity can be given that satisfies C1 – C3. There is reason to believe that this will also be sufficient to adequately meet these challenges. After all, an account of the value of the commodity would make good sense our concern with knowledge. What's more, if we did succeed in providing an account of the value of the commodity of knowledge we would still stand as a good chance as any of vindicating the special focus on knowledge in the history of epistemology. For that reason, it seems that everything is to be gained and nothing to be lost by exploring the prospects for an account of the value of the commodity of knowledge.

Before moving on to the value of the commodity of knowledge, we'd like to take a look at the value of another central commodity in our lives, to wit, water. Water is of course valuable in many respects. For the purposes of this paper, we'd like focus on one valuable quality of water, its power to quench our thirst. Suppose liquid hydrogen were just as well suited to quench our thirst as is water. The constituents of liquid hydrogen, H_2, are a proper subset of the constituents of water, H_2O (so that liquid hydrogen stands to water as, for instance, justified true belief stands to knowledge.) Now suppose you have before you two glasses, one containing water, the other liquid hydrogen.

[3] [**?**, 128]. Pelletier also points out that mass terms are generally regarded to have divisive and cumulative reference : one can subdivide the stuff of which a mass term is true infinitely and the mass terms will continue to true of its parts (divisive reference) and one can add as much of the stuff to an existing quantity of it as one likes and the term will continue to be true of the resulting mass (cumulative reference). One might think that this is implausible for knowledge. After all, if we take away the justification component from a bunch of beliefs that make up some quantity of knowledge, for instance, 'knowledge' won't be true of the resulting mass. But notice that this is not a special problem for 'knowledge'. After all, if you chop up some water molecules that make up a quantity of water and take away the oxygen atoms, say, 'water' won't be true of the resulting mass either. Pelletier [**?**, 129] suggests that this problem may be solved by distinguishing between metaphysical and semantic facts. The thought is that while as a matter of metaphysical fact water consists of hydrogen and oxygen atoms, this is not recognised semantically, at least not in English. Whether or not this solution works need not concern us here. After all, so long as the problem is not exclusive to 'knowledge', we can, for present purposes at least, safely ignore it.

It is plausible that the glass with water is no more valuable than the glass with liquid hydrogen, at least not with respect to its power to quench your thirst. After all, *ex hypothesi*, liquid hydrogen is as well suited to do the job as water is. Does that mean that water, the commodity, couldn't have special value, value that warrants our concern with water ? No. To see this, suppose (as happens to be the case) that liquid hydrogen is extremely rare and can exists only in very special environments. Suppose that, at the same time, water is easily available in a wide range of places and to a wide range of people. In that case water is plausibly valuable to us in a way that would warrant our concern with it. What makes water thus valuable is not just the fact that it has the power to quench our thirst. After all, we are supposing that water shares this property with liquid hydrogen. It is a combination of the fact that it has this property and the fact that it is so widely and easily available. To put a snappy label to it, water is of special value because it is our way of quenching our thirst.

Now, we want to suggest that the situation is in essence the same with knowledge on the one hand and true belief that falls short of knowledge on the other. One valuable property of knowledge is that it is a way of correctly representing the world around us. It is undeniable that the same holds for true belief that falls short of knowledge. If we compare two agents, A and B, where A knows that p and B truly believes but doesn't know that p. Here it is very plausible that A's belief that p is no more valuable than B's, at least not with respect to its correctly representing the world—just as it is very plausible, in the imagined case above, that the glass of water is no more valuable with respect to thirst-quenching than the glass of liquid hydrogen. Arguably, however, just as in the case of water and liquid hydrogen, this result is compatible with knowledge being valuable in a way that would warrant our concern with it. In fact, the very same properties that account for the special value of water in the imagined case account for the corresponding special value of knowledge : in a wide range of areas, knowledge is widely and readily available.

To see this, consider first perceptual beliefs about middle-sized dry goods. On any non-sceptical account of knowledge, given formation by suitable processes (alternatively : on suitable grounds) in sufficiently hospitable epistemic environments, these beliefs will qualify as knowledge. For instance, my belief that there is a computer on the desk before qualifies as knowledge : it is produced by a highly reliable ability to recognise tables in an epistemically hospitable environment. Now the crucial point is that, for beliefs in this range, formation by suitable processes in hospitable environments is the norm ; formation of beliefs by unsuitable process, or in inhospitable environments is the exception. If this isn't immediately clear, consider again my belief that there is

a computer on the desk before me and ask yourself what would have to be the case for my belief to remain true but fall short of knowledge. Those with some training in epistemology will find it easy to answer this question : I mistake a hologram for a computer, whilst unbeknownst to me there is a computer somewhere else on the desk, I acquire my belief by a highly unreliable process such as a coin-toss, etc. While any of this might come to pass, it is undeniable that, as a matter of fact, it only rarely does. For that reason, cases of knowledge are the norm and cases of true belief that fall short of knowledge are the exception.

Perceptual beliefs about middle-sized dry goods are not the only cases in point. Consider testimonial belief about propositions of crucial practical importance in our lives : propositions about bills that need to be payed, the nature of the sickness of your child and the medication that will cure it, what's available at the local restaurant, etc. Or consider inferentially supported beliefs that exploit a variety of natural and social regularities : that my car is still parked outside the institute, that Cameron is still the prime minister of the UK, etc. Here too, when beliefs in these ranges are formed by suitable processes in sufficiently hospitable epistemic environments, they will qualify as knowledge. Here too, cases of knowledge are norm and cases of true belief that fall short of knowledge are the exception.

These considerations suggest that in wide range of cases, knowledge is widely and readily available. All we have to do to acquire knowledge is open our eyes, listen to what other people tell us, attend to our feelings, etc. In comparison, in those areas true belief that falls short of knowledge is a rare commodity that it exists only in very special environments. In parallel with the case of water, then, what makes knowledge specially valuable is not just the fact that it involves a correct representation of the world. It is the fact that it has this property in combination with the fact that, in a wide range of areas, it is so widely and easily available. Just as water is of special value because it is our way of quenching our thirst, knowledge is of special value because, in a wide range of areas, it is our way of correctly representing the world.

Value Inheritance

Suppose that the above account of the value of the commodity of knowledge is successful. One question that we might ask ourselves at this stage is what, if anything, this entails for the value of individual items of knowledge.

One possible answer is that individual items of knowledge inherit special value from being instances of a specially valuable type and that, in conse-

quence, individual items of knowledge turn out to be more valuable than individual items of true beliefs that fall short of knowledge. (We will henceforth refer to the thesis that tokens of a valuable type inherit value from the type as 'Value Inheritance'.)

Now we think that Value Inheritance is most plausible when the value of type is *final* value. For instance, if happiness as a type of state is finally valuable, one might think that instantiations of happiness are finally valuable also. At the same time, we said that we would bracket the issue of final value for the purposes of this paper. So the crucial question is whether Value Inheritance holds for instrumental value. For instance, do individual glasses of water inherit value from the fact that they are instances of a commodity that is valuable because of its wide and easy availability ? That seems somewhat implausible. To see why, consider an individual glass of water that it is currently on an unmanned spaceship in the orbit of a faraway planet. It seems plausible that no value whatsoever need attach to this glass of water. In other words, it may be a useless glass of water. So, suppose it is. If the individual glass is useless, however, it cannot have inherited value in virtue of being an instance of a valuable commodity. The fact that there can be useless instances of an instrumentally valuable commodity suggests that Value Inheritance does not generally hold for instrumental value.

In a recent paper, Alvin Goldman and Erik Olsson (2009) argue, roughly, that Value Inheritance is possible even for instrumental value. They offer the state of possessing money as a case in point. The idea is that possessing money is an instrumentally valuable type of state as it frequently produces states that are valuable for their own sake (e.g. happiness). Crucially, Goldman and Olsson claim that "each token of this type inherits instrumental value from the type" (Goldman & Olsson 2009, 32). If they are right, Value Inheritance might still hold for instrumentally valuable types in some cases.

It may look as though, compatibly with Value Inheritance, Goldman and Olsson can allow for the existence of token states of possession of money that are useless. As they point out, each token state inherits value from the type even though they may not actually produce finally valuable states. In the case of money, this may happen when the money isn't spent at all or it is badly invested (Goldman & Olsson 2009, 32). On reflection, however, it cannot be the case that a token state of possessing money is genuinely useless (i.e. has zero instrumental value) and yet inherits positive instrumental value from the instrumentally valuable type. One of the two has to go. The question is which one and why.

Given that Goldman and Olsson accept that Value Inheritance holds for

possessing money, it has to be useless money. What appears to be going on here is that the value of token states of possessing money resides in a *dispositional* property—roughly, its power to bring about something finally valuable (though perhaps only by producing something else of instrumental value). Since a dispositional property can be had even when the disposition is not manifested, if the value of a token state of possessing money resides in the dispositional property, money can be valuable even when it does not bring about something of final value. Crucially, however, in that case the token state is not useless.

But couldn't there be useless money? We think there could be. Suppose you own a one hundred Euro bill. Unfortunately, however, the bill is on an unmanned spaceship that is now in such a remote part of the universe that it is certain to have disintegrated before it can reach the next living being. In this case, we submit, the token state of your owning this bill is useless. There is a rationale behind this verdict. Even when the instrumental value of a certain item resides in a dispositional property, the item can be useless if there is no chance at all that the disposition will be manifested. This is what is happening in (a suitably fleshed out version of) the above case. There is simply no longer any chance that the money is spent and hence that the disposition in which the money's value resides will be manifested. That's why the state of your owning the bill in this case is useless.

There is thus reason to believe, *pace* Goldman and Olsson, that Value Inheritance does not hold for money. Goldman and Olsson also offer a second example, namely that of good motives for actions. The idea is that good motives are a type of state that acquires value from its relation to good action. Over time, good motives have come to be valued independently, "in themselves" (Goldman & Olsson 2009, 33) as they put it. Now, it is not clear to us what exactly Goldman and Olsson mean here. Is it that good motives come to be valued for their own sake? Whatever the answer to this question may be, we think that it is very plausible that good motives are valued for their own sake. As we already pointed out, we are also sympathetic to idea that Value Inheritance holds when the value of the type is *final* value. So, we are happy to grant that Goldman and Olsson have succeeded in identifying a case of Value Inheritance. It's just that the case is not of the kind we were looking for. After all, at least for all Goldman and Olsson have argued, the value inherited here is final, not instrumental.

In view of these considerations, we are suspicious of the idea that Value Inheritance holds for instrumental value. Of course, we are open to be convin-

ced otherwise.[4] Suppose it can be established that Value Inheritance holds in a certain range of cases and that knowledge is within that range. All that follows from our account is that individual items of knowledge are more valuable than the corresponding beliefs that aren't knowledge after all. While this would require us to change our verdicts about the value of beliefs about the number of grains of sand in some jar and like cases, the result is not unwelcome. After all, prospects for an alternative explanation of the intuition of zero value are fairly bright : in case of knowledge, the value of the type is often overlooked and, additionally, instrumental value is often not inherited by all tokens of the type. No surprise that the inherited value could remain unrecognised in case of knowledge.

4. The Third Challenge

On the account developed in §3., knowledge is valuable because it is our way of correctly representing the world. Note that this will serve to address the first two challenges from §2. (C1, C2) as it explains why knowledge is more valuable than mere true belief and more valuable than beliefs that fall short of knowledge. At the same time, it is not hard to see that our account does precious little to address C3, according to which knowledge must have a different kind of value than beliefs that falls short of knowledge. For that reason, we'd now like to take a brief look at whether C3 can be met as well.

Superiority

First, we'd like to express a worry about the challenge. Suppose it can be shown that knowledge is finally valuable. In that case, it would seem that we have everything we could hope for. By the same token, any plausible challenge for an account of the value of knowledge should at this stage be met. Unfortunately, there is no guarantee that C3 will be met. To see this, suppose

[4] A promising place to look are normative properties. For instance, possession of legal tender is a type of state that has the instrumentally valuable normative property of entitling you meet financial obligations by using it. We think that the idea that the instrumentally valuable normative property is inherited by each token of the state type carries promise. If, in addition, knowledge has instrumentally valuable normative properties, we would have what it takes to argue that knowledge has instrumental value that is inherited by each token item of knowledge. That said, we will not pursue this line in any more detail here. Suffice it to say that the instrumental value that we argue attaches to the commodity of knowledge is not a normative property. Even if Value Inheritance holds for normative properties, there still is little reason to think that it holds for the properties that, according to our proposal, make the commodity of knowledge valuable.

that true belief turns out to be finally valuable as well. Given, additionally, that all other kinds of value are equally shared between true belief and knowledge, there will no kind of value that attaches to knowledge that does not attach to true belief. C3 will not be met. As a result, there is reason to believe that C3 is too demanding.

At the same time, we think that Pritchard may have been on to something when he introduced C3. To see this, let's take a look at how he motivates it :

> [I]f one regards knowledge as being more valuable than that which falls short of knowledge merely as a matter of degree rather than kind, then this has the effect of putting knowledge on a kind of continuum of value with regard to the epistemic, albeit further up the continuum than anything that falls short of knowledge. The problem with this 'continuum' account of the value of knowledge, however, is that it fails to explain why the long history of epistemological discussion has focused specifically on the stage in this continuum of value that knowledge marks rather than some other stage (such as a stage just before the one marked out by knowledge, or just after). Accordingly, it seems that accounting for our intuitions about the value of knowledge requires us to offer an explanation of why knowledge has not just a greater *degree* but also a different *kind* of value than whatever falls short of knowledge.
>
> (Pritchard *et al.* 2010, 7–8)

What becomes clear here is that Pritchard takes it, first, that no account of the value of knowledge on which it is on a continuum with the value of belief that isn't knowledge can be successful. He also seems to think, second, that the only way in which we can avoid placing knowledge on such a continuum is by showing that knowledge enjoys a different kind of value.

Importantly, the second claim is false. Even if a difference in kind of value is sufficient to get knowledge off the value continuum with belief that falls short of knowledge, it isn't necessary. There are other ways in which the value of one type of good, A, can be discontinuous with the value of another type of good, B. For instance, it may be (i) that any amount of A is better than any amount of B (henceforth also 'Strong Superiority') or (ii) that some amount of A is better than any amount of B (henceforth also 'Weak Superiority'). Mill famously put forth such a discontinuous account of value relations :

> It is quite compatible with the principle of utility to recognise the fact, that some kinds of pleasure are more desirable and valuable than others. – Of two pleasures, if [...] one of the two is, by those who are competently acquainted with both, placed so far above the other that they [...]

> would not resign it for any quantity of the other pleasure which their nature is capable of, we are justified in ascribing to the preferred enjoyment a superiority in quality, so far outweighing quantity as to render it, in comparison, of small account.
>
> (Mill 1963, 210)

Both Strong and Weak Superiority will take knowledge off a continuum with belief that isn't knowledge. At the same time, neither requires a difference in kind between these two. So, a more promising way of understanding C3 is that in order to account for the special value of knowledge, we have to show that knowledge is, in some respect, at least weakly superior to belief that falls short of knowledge. That is to say, we need to show, at a minimum, that some amount of knowledge is in some respect better than any amount of true belief.

Eudaimonic Value

While we aren't certain that even the modified version of C3 constitutes a reasonable demand on adequate accounts of the value of knowledge, in the remainder of this section we want to try and provide some support for the claim that knowledge is weakly superior to belief that isn't knowledge. We take this part of the paper to be rather speculative : a sketch of an argument that it may be worth pursuing in more detail elsewhere rather than a thorough defence. Roughly, the idea is that a certain amount of knowledge is required to achieve one of the highest goods in life : human flourishing or what Aristotle called 'eudaimonia'. Eudaimonia is a type of happiness. Crucially, however, it is not happiness of any old sort. As Rosalind Hursthouse points out, eudaimonia is "the sort of happiness worth seeking or having." (Hursthouse 2007, S2)

We will assume that a eudaimonic life is at least weakly superior to a life without eudaimonia. To put it in Mills's terms, no one fully acquainted with both lives would sacrifice a eudaimonic life for a life without it, no matter how good the non-eudaimonic life may be in other respects. Derek Parfit nicely illustrates the spirit of this idea in the following passage :

> I could live for another 100 years, all of an extremely high quality. Call this the Century of Ecstasy. I could instead live forever, with a life that would always be barely worth living [. . .] the only good things would be muzak and potatoes. Call this the Drab Eternity. I claim that, though each day of the Drab Eternity would be worth living, the Century of Ecstasy would give me a better life. Though each day of the Drab Eternity would have some value for me, no amount of this value could be as good for me as the

> Century of Ecstasy.
>
> (Parfit 1984, 17–18)

Now suppose it can be shown that a eudaimonic life requires a certain amount of knowledge and that no amount of belief that falls short of knowledge will do the trick. In that case, knowledge will also be weakly superior to belief that isn't knowledge. After all, there will be an amount of knowledge that cannot be sacrificed for any amount of belief that falls short of knowledge without losing something of superior value, to wit, the eudaimonic life.

What remains to be shown is that the eudaimonic life requires a certain amount of knowledge. Here is one way of venturing to achieve this. Recall that knowledge is widely and easily available to us in a wide range of areas. As a result, in these areas we often have the ability to know, knowledge is within our reach as cognitive agents. Notice, furthermore, that knowledge often features in our motivations and aims, which seems reasonable, given that it is within reach. Now, anyone who systematically failed to attain knowledge would systematically fall short of his potential as a cognitive agent and, when aiming for knowledge, would systematically fail to attain his aims as a cognitive agent. Plausibly, however, no one who systematically falls short of his potential as a cognitive agent and systematically fails to attain his aims as a cognitive agent will attain intellectual flourishing. For that reason, agents like us, for whom knowledge is within reach in a wide range of areas, won't be able to attain intellectual flourishing without attaining a wide range of knowledge. Insofar as it is plausible that, in agents with our cognitive sophistication and potential, eudaimonia also requires intellectual flourishing, knowledge is requisite for eudaimonia in agents like us. For us, the eudaimonic life is a life rich (enough) in knowledge.

5. Références

Goldman, A., & Olsson, E. 2009. Reliabilism and the value of knowledge. *In* : Haddock, A., Millar, A., & Pritchard, D. (eds), *Epistemic Value*. Oxford : Oxford University Press.

Hazlett, A. 2010. The myth of factive verbs. *Philosophy and Phenomenological Research*, 80, 497–522.

Hursthouse, R. 2007. Virtue ethics. *In* : Zalta, E.N. (ed), *The Stanford Encyclopedia of Philosophy*, july 2007 edn.

Kagan, S. 1998. Rethinking intrinsic value. *Journal of Ethics*, 2, 277–297.

Korsgaard, C. 1983. Two distinctions in value. *Philosophical Review*, 92(2), 169–195.

Kvanvig, J. 2003. *The Value of Knowledge and the Pursuit of Understand- ing. Cambridge* : Cambridge University Press.

Lewis, D. 1996. Elusive knowledge. *Australasian Journal of Philosophy*, 74, 549–567.

Mill, J.S. 1963. *Essays on Ethics, Religion, and Society*. The Collected Works of John Stuart Mill, vol. 10. Toronto : University of Toronto Press.

Parfit, D. 1984. *Reasons and Persons*. Oxford : Clarendon Press.

Pelletier, F.J. 2010. Mass Terms : A Philosophical Introduction. *In* : Pelletier, F.J. (ed), *Kinds, Things, and Stuff. Mass Terms and Generics*. Oxford : Oxford University Press.

Pritchard, D., Millar, A., & Haddock, A. 2010. *The Nature and Value of Knowledge*. Oxford : Oxford University Press.

Rabinowicz, W., & Rønnow-Rasmussen, T. 2000. A distinction in value : Intrinsic and for its own sake. *Proceedings of the Aristotelian Society*, 100, 31–53.

Radford, C. 1966. Knowledge–by examples. *Analysis*, 27, 1–11.

Riggs, W. 2009. Understanding, knowledge, and the Meno requirement. *In* : Haddock, A., Millar A. Pritchard D. (ed), *Epistemic Value*. Oxford : Oxford University Press.

Rønnow-Rasmussen, T., & Zimmermann, M. (eds). 2005. *Recent Work on Intrinsic Value*. Dordrecht : Springer.

Sosa, E. 2003. The place of truth in epistemology. In : DePaul, Michael, & Zagzebski, Linda (eds), *Intellectual Virtue : Perspectives from Ethics and Epistemology*. Oxford : Oxford University Press.

15

The Value and Normative Role of Knowledge

JULIEN DUTANT

Over a decade as my supervisor and mentor, Pascal has skilfully managed to keep me under the impression that I carried my research freely while quietly guiding my attention towards deeper and unfamiliar issues. He thus introduced me to knowledge-first epistemology, the debate over pragmatic encroachment and the role of norms, values and reasons in epistemology before I quite grasped their significance. Each time the new thoughts would slowly make their way into my own and I would eventually find myself intensely preoccupied with the issues that were central to Pascal's seminars a few years back. The present paper is another instance of this phenomenon. I dedicate it to Pascal, with all my respect and gratitude, and with apologies for my *esprit d'escalier*.

In his (2009), Pascal connects two topics that epistemologists have mostly kept apart. One is the debate over pragmatic encroachment, namely, the idea that whether one knows (or believes, or justifiably believes) partly depends on practical factors such as one's interests and the stakes one faces. Central to this debate is the claim that *knowledge is the norm of action* : that is, that in some salient sense of "ought", one ought to act in view of what one knows. The other issue is the question of the *value of knowledge*, namely, whether and why knowledge is a good thing, and in particular, a better thing than mere true belief. Most of Pascal's paper deals with pragmatic encroachment. It was what attracted all my attention then. But it also raised a second question that epistemologists seldom discuss : could the idea that knowledge is the norm of action explain the value of knowledge ? Pascal's answer was negative. At

the time I could hardly get my mind around the question. It is not *prima facie* clear how one could even try to derive one claim from the other. One may sketch some paths ; for instance, if you should act only on what you know, then when you know you have "more" to act on than when you merely have a true belief. But these hardly constitute a suitable basis for discussion. Pascal's negative answer supported the widespread attitude of keeping the two ideas apart. I went along and forgot all about it. A few years later, I feel I have finally reached a perspective from which I can take up Pascal's question. My views on the matter are far from settled, so this is more of a progress report. The option that I currently find the most appealing differs from Pascal's. Like him, it denies that we can explain the (alleged) value of knowledge by its normative role, but unlike him, it does take the normative role of knowledge to shed light on its value, by showing why it need not have value at all. Before I get to this, however, I will lay out the perspective from which I take up Pascal's question.

1. Why knowledge matters

Epistemology in the second half of the XXth century enjoyed a spectacular revival. But when it came to knowledge it focused almost exclusively on two questions : what it is and whether we have it. It also asked when a belief was "justified", which, to some at least, was the same as asking what one ought to believe. But the latter question was mostly treated as independent of, and prior to, questions about knowledge. For a variety of reasons, by the 1990s epistemologists increasingly wondered whether and why knowledge mattered. It is well to ask what knowledge is and whether we have it, but, they began to ask, why should we care ?

The question is pressing if we distinguish knowledge from justified true belief — as is common in post-Gettier times. Suppose I come to the conclusion that I do not know whether there are lions. That may be an unsettling conclusion to reach. But why, exactly ? I may start wondering whether there really are lions. And I may be unsettled at the thought that there are no lions. But all that suggests is that *whether there are lions* matter, not whether I know that they are. Similarly, I may be unsettled at the thought that I may have been mistaken on that matter and others. But all that suggests is that *whether my beliefs are true* matters. Another reaction I may have is to judge that I ought not to believe that there are lions. But all that suggests is that *whether my belief that there are lions is justified — whether I ought to believe* — matters. If whether I ought to believe it does not whether I know it, then again that does not entail

that knowledge itself matters. Putting it all together, it may matter whether p, whether I have a true belief that p, whether I am justified in believing that p; but if I can have a justified true belief that p without knowing p, it is unclear whether and why knowing itself matters.[1]

A somewhat shallow answer defers to common sense. We think about knowledge a lot. The verb *know* is currently one of the ten most used verbs in the Oxford English Corpus. It is the second propositional verb (after *say*) and the most common verb describing a mental state (just before *see*, *think*, *look* and *want*). It is much more used than *believe*, *true*, *justified* and even more used than *ought*, *should* and *must*.[2] Since we talk about knowledge a lot, we think about it a lot. Moreover, we take ourselves to know many things and we want to know many things.[3] So philosophers can rest assured that knowledge at least matters *to us*.[4] The answer is somewhat shallow, however. First, even on the assumption that knowledge is something we desire, we may still wonder whether and why it is desirable.[5] Second, similar remarks can be made about other common notions. We take ourselves to *do*, *make* and *get* many things, and we want to *do*, *make* and *get* many things. Yet few philosophers would say that *doing*, *making* or *getting* matter. That is so, I venture, because philosophers take these notions to be too crude to describe the underlying phenomena. There is little useful theory to be made about the making that is common to making a plan, making a present and making a soup. Philosophers found it more useful to theorize about the underlying phenomena in terms of *intention*, *action*, *causation*, *ownership* and so on. One may worry that knowledge is also too crude a notion for picking up something that matters and that is worth theorizing about. The worry is made more acute by the existence of epistemological traditions that do without the notion altogether, adopting instead notions such as *justification*, *evidence*, *confirmation* and probabilistic notions.[6]

[1] This line is forcefully pushed by Mark Kaplan (1985).

[2] "The OEC : Facts about language", Oxford University Press, http://www.oxforddictionaries.com/words/the-oec-facts-about-the-language, retrieved Jan 4th2014.

[3] As Aristotle famously notes in *Metaphysics* A, 1.

[4] See Williamson (2000, 31) : "For knowing matters ; the difference between knowing and not knowing is very important to us. Even unsophisticated curiosity is the desire to *know*."

[5] Even subjectivists who think that things are at bottom made valuable by our valuing them are not committed to the view that everything we desire is desirable.

[6] Peirce was an early defender of the view that the notion of knowledge is disreputable : "there will remain over no relic of the good old tenth-century infallibilism, except that of the infallible scientists, under which head I include [...] all those respectable and cultivated persons who, having acquired their notions of science from reading, and not from research, have the idea that

Two more substantial answers have been prominent in recent literature.[7] The first is that knowledge is *good*, or, as philosophers prefer to say, that it *has value*. Good things obviously matter ; so if knowledge is good, it matters. The idea famously figures in Plato's *Meno*, where Socrates approvingly reports that "knowledge is prized higher than correct opinion".[8] The answer only goes so far. We may still wonder why knowledge is good, and in particular why it is better than justified true belief. This has been the subject of much discussion in the last decade. But as long as we grant that knowledge is good, as the recent value of knowledge overwhelmingly does, we already have an answer to why it matters. Call it the *value* answer.

The second answer comes from the idea that knowledge is something we *ought* to have in order to do certain things or have certain attitudes. I will focus on two such claims. The first is that *one ought to act only on the basis of what one knows*. Call it the Knowledge–Action Principle. The second is that *one ought to believe only what one knows*. Call it the Knowledge–Belief Principle. There are alternative or additional claims in the vicinity : one ought to do what is best in view of what one knows ; one ought to believe only on the basis of what one knows ; one ought to be certain only of what one knows ; and so on. It does not matter for our purposes which ones we choose ; the two selected above will serve as concrete illustrations. Now what one ought to believe and what one ought to do obviously matters. So if what one knows partly determines what one ought to believe and to do, then knowledge matters. So principle like the two above also offer an answer to the question why knowledge matters. Call it the *normative role* answer.

What the normative role answer amounts to is not entirely clear because *ought* may mean many things. For now we just flag the issue ; it will take a central importance later on.

The normative principles have an ancient pedigree as well. Zeno (of Citium, the founder of the Stoa) claimed that the wise only assent to what they have a "grasping impression" of — by which he essentially meant, what they know.[9] Since he clearly thought that one ought to do what the wise does,

"science" means knowledge, while the truth is, it is a misnomer applied to the pursuit of those who are devoured by a desire to find things out." (Peirce, 1950, 3) (It does not seem to occur to Peirce that *finding something out* may come down to *coming to know*.)

[7] They are not the only ones. Another one (inspired by Craig, 1990) is that knowledge is property of people that is useful for us to spot while inquiring : we figure out who knows what to decide who to use as source of information.

[8] 98a, trad. G.M.A. Grube.

[9] See *e.g.* Cicero, *Academica* 2.77–8, quoted in Long and Sedley (1987) as 40D. Stoics would

that amounts to an endorsement of the second principle. Academic sceptics agreed ; but since on their view neither we nor the wise knew anything, they argued that one ought not to assent to anything. As a result they were under pressure to deny the first. Confronted with the Stoic objection that the wise would have to act, Carneades developed the idea that one could act on the basis of merely *convincing* impressions. The fact that Stoics saw that as an objection shows that they endorsed the first principle as well. Just as the claim that knowledge is better than mere true belief is often called *Meno's thesis*, we may call the two principles *Zeno's norms*.

The two answers are mutually compatible : it may be that knowledge matters both because it is a good thing *and* because it plays a certain normative role. But it is tempting to see whether one could be used to derive the other. A *Value to Norm* derivation would derive Zeno's norms from Meno's thesis and plausible. A *Norm to Value* derivation would derive Meno's thesis from Zeno's norms. In both cases we would allow the use of plausible background assumptions. Since there is little about norm or value that is uncontroversial, we may also generously allow the use of controversial claims about norms or values in general. As I understand "derivation" no order of priority is required : it may be that both Meno's thesis follows from Zeno's norms and the other way round.[10]

This paper discusses both derivations. Section 2. discusses the Value to Norm route. I am not optimist for it. Some reasons for pessimism come from Firth (1998a) and Berker (2013a, 2013b). They argue that "consequentialist" or "teleological" ways of deriving epistemic norms from epistemic values fail because they result in norms allowing for trade-offs that the correct epistemic norms for belief forbid. I do not find the objection decisive, however. It leaves some "teleological" derivations standing, as well as non-teleological ones. A more serious problem seems to me to be the impossibility of deriving anything

say that what we would now describe as paradigm situations of perceptual knowledge (seeing that an apple is on a table) involve "having a grasping impression". But they would not call it "knowledge" (*episteme*) yet, for they thought that knowledge required resistance to dialectical cross-examination. Once we set aside this inflated view of knowledge — or once we read Stoics' notion of *episteme* as denoting something like science or scientific understanding —, we can take their theory of "grasping impressions" as a theory of knowledge. (Commentators sometimes do so without further ado, *e.g.* Long and Sedley 1987 and Frede 1983/1987 ; see Frede 1999 and Hankinson 2003 for more guarded statements.)

[10]Some characterize "consequential theories" as those that explain the right in terms of the good and "deontological theories" as those that explain the good in terms of the right. Because these terms have many associations and because they are meant to be exclusive of each other, I think it would be misleading to use them to label the Value to Norm and Norm to Value derivations, respectively.

like the Knowledge–Action Principle. In a nutshell, the problem is that Meno's Thesis cannot ground a difference in value between a case where one has a good and a bad belief, but acts on the good one, and a case where one has a good and a bad belief, but acts on the bad one. Thus the problem arises because the Knowledge–Action Principle is not simply a norm about belief but about the coordination of belief and actions.

Section 3. discusses the Norm to Value route. It is *prima facie* more promising. Saying so goes against a strong trend in current epistemology : while there has been much debate over why knowledge is good, little of it has explored the idea that it is good because of its normative role. Epistemologists appear to have assumed that something like the Value to Norm derivation will in turn explain the normative role of knowledge, and that it would therefore be illicit to appeal to the normative role of knowledge to explain its value. That being said, I have doubts about it as well. As far as I can tell, the derivation would have to rely on the idea that knowledge is good because it is required to be allowed to believe or act on one's belief. But in general it is not the case that necessary conditions for being allowed to do something are good — not even that necessary conditions for being allowed to do something good are themselves good. The problem does not show that the derivation fails, but it indicates that more needs to be say.

Section 4. turns the apparent failure of the Norms to Value derivation into a virtue. For once we assume that knowledge plays a central normative role, it becomes unclear what is left of the motivation for the idea that it is a good thing. For instance, the fact that it plays a central normative role is sufficient to explain why knowledge matters. There is no need to make the additional claim that it is a good thing. So we may try to use Zeno's norms to explain *away* Meno's Thesis. I have put the suggestion forward elsewhere (Dutant 2012, forthcoming). Here I want to discuss two problems for it. The first is that the proposal has a hard time explaining why knowledge is something worth *aiming* at, for Zeno's norms themselves do not prescribe *acquiring* knowledge. In reply I argue that such prescriptions follow from Zeno's norms in conjunction with other aims and other norms of action. Another is that the proposal requires a strong primitive, namely a layer of normativity distinct from the usual "objective" and "subjective" *ought* that are commonly accepted. I will put forward a few considerations in its favour.

2. From Value to Norms

Let us assume Meno's Thesis and examine whether we can derive Zeno's norms. Meno's Thesis, expressed as the slogan "knowledge is better than mere true belief", is somewhat unspecified. It is unclear whether it is a generic or universal claim and what exactly the bearers of value are supposed to be. For the sake of concreteness we will use on a more precise claim. The claim ascribes values to states of affairs. It states that knowledge is *pro tanto* good and that belief without knowledge is *pro tanto* bad :

(MT) For every S, t, p, the state of affairs of S knowing p at t is (*pro tanto*) good, and the state of affairs of S believing p without knowing p at t is (*pro tanto)* bad.

It follows from (MT) that knowing p is better than having a true belief in p that does not constitute knowledge. For there is a disvalue in the latter that is absent in the former, namely believing without knowing. It also follows from (MT) that knowledge-constituting belief is better than lack of belief, and that lack of belief is better than belief that does not constitute knowledge. It does not follow from (MT) that knowing p is always *overall* good ; the value that it has in virtue of being knowledge can be offset by other considerations. Similarly, it does not follow from (MT) that believing without knowledge is always *overall* bad ; its disvalue may be offset by other considerations. (MT) is neutral on whether the *pro tanto* value of knowing p is the same for every p. Perhaps some things are more valuable to know than others.

(MT) is stronger than the claim that *some* or *most* state of affairs of knowing are *pro tanto* good. It is also stronger than its first conjunct alone. If the derivation fails with that strong assumption, it will fail with weaker ones. We may give the derivation its best chance.

Norms are about what we ought to do ; values about what is good or bad. How do we derive one from the other ? A common paradigm is *consequentialist* : roughly, one ought to do what has or tends to have the best consequences. As Berker (2013a, 351–7) notes, much contemporary epistemology adopts such a framework.[11] It is assumed that we have certain epistemic aims

[11] In Berker (2013a, 342) prefers the term "epistemic teleology", because he thinks that "epistemic consequentialism" will evoke the view that what one ought to, epistemically speaking, is what promotes *practical* (non-epistemic) goods. Firth (1998a) uses the term "epistemic utilitarianism" ; others "epistemic instrumentalism" (Kelly, 2003) . "Epistemic consequentialism" is used by Percival (2002), Stalnaker (2002) and Berker (2013b), among others. There are some differences in how these authors characterize the view so labelled. For instance Kelly takes it to include the

— such as having true belief and no false beliefs — and that what we ought to do, epistemically speaking, is what promotes those aims. Berker (2013a, 2013b), building on a problem due to Firth (1998a, 1998b), argues that any such derivation of epistemic norms will fail. That is, any such derivations will misclassify some justified beliefs as unjustified and conversely. While I share the view that consequentialism is unsuited to derive norms of belief and I agree that Firth and Berker's problem shows that many versions of epistemic consequentialism fail, I do not think they rule out all such versions. Be that as it may, Berker's and Firth's problem leave untouched *non-consequentialist* ways of deriving norms from value. So for our purposes, the discussion of epistemic consequentialism is mostly a side-show. Since, however, the paradigm is the most familiar one, it is worth going through it.

Berker (2013a, 344–7) characterizes consequentialist normative theories as having three components. First, a *theory of final value*, which states what things have value in themselves. Second, a *theory of overall* value, which ascribes value to things according to whether and how they promote finally valuable things. Third, a *deontic* theory, which states what one ought to do in terms of overall value. For our purposes we call a belief one ought to have (or is allowed to have) a *justified* belief and a belief one ought not to have a *unjustified* belief. In our attempted derivation, the theory of final value is given by (MT).[12] To illustrate a complete theory :

> *Theory of final value*. For every S, t, p :
> S's knowing p at t is (*pro tanto*) finally good,
> S's believing p without knowing p at t is (*pro tanto)* bad.

We call "final *epistemic* value" the value that things have in virtue of these clauses. We assume that there is some way of summing final values so that the total final epistemic value of a compound state of affairs is the sum of the final epistemic value of its components.

> *Theory of overall epistemic value* (for state of affairs).
> A state of affairs is epistemically better than another iff the total final epistemic value it brings about (or would bring about if it

idea that epistemic norms and values are contingent upon one's having certain epistemic goals — an idea he objects too. But all share the core idea that what one ought to do, epistemically speaking, is a matter of what promotes the epistemically best consequences.

[12]Rather, we treat it as final. The theory leaves open the possibility that the value of knowledge is ultimately reduced to something else, *e.g.* the value of true belief or practical value.

> obtained) is greater than the total final epistemic value the other brings about (or would bring about if it obtained).

We leave open what exactly counts as *brought about* by a state of affairs : all effects, including long-term ones ; proximate effects ; constitution ; constituents (see Berker, 2013a, 347 for some discussion).

> *Deontic theory* (for beliefs). For every S, p, t :
> S ought to believe p at t, epistemically speaking, iff S's believing p at t is overall epistemically better than S's not believing p at t.

The qualification "epistemically speaking" leaves room for one's epistemic duties to be overruled by other duties. The *ought* claim we derive here is an 'objective' one : it roughly says that one ought to have the beliefs that *in fact* have the best epistemic consequences, whether or not one is aware of them. As in consequentialist ethics, we may associate a 'subjective' *ought* to the objective one :

> S (subjectively) ought to believe p at t, epistemically speaking, iff S's believing p at t is *expectably* overall epistemically better than S's not believing p at t.

Where something is *expectably* overall better iff its expected overall value (to S at t) is higher.[13]

The crucial feature of consequentialist views, in Berker's characterization, is to ascribe overall value to what *promotes* final value. As a result, overall value typically allows for trade-offs : something may be overall good despite having bad consequences, provided it has many good consequences as well. Berker take these trade-offs to generate mistaken epistemic norms. He does not propose a general argument that it is so, however ; rather, he mainly argues by generalizing from cases.

Berker's prediction seems borne out when we consider a *direct* deontic theory. I call a *direct* deontic theory one that prescribes a belief directly as a function of its overall value. Their form is along the lines of :

> Believe p iff the (expected) overall value of doing so is above a certain threshold.

[13]The principle assumes that a notion of expected overall value is defined — *e.g.*, a sum of values of possible outcomes weighted by their probability. It leaves open what sort of expectation is relevant, *e.g.* what degrees of belief the subject has, or what degrees of belief she should have in view of her evidence, and so on.

The theory given above is an illustration. Now take a case of unjustified belief — say, a belief based on reading tea leaves, while one knows very well that tea leaves do not indicate anything. We can alter the case so that the belief has many epistemically good consequences — for pretty much any notion of consequence and any notion of epistemic good. With enough good consequences, the belief will be counted as overall good. We can even pile up the good consequences until any desired threshold of overall value. By the direct deontic theory, the belief will be counted as justified, *contra hypothesi*. So the theory is false.

Firth (1998c) has put forward cases along those lines (see also Berker 2013b, 369). A brilliant set theorist is on the verge of a ground-breaking discovery, but she is suffering from a serious illness and the doctors give us less than two months' time. Against all evidence, she clinches to the conviction that she will live one full year. The belief in fact raises the chances that she survives long enough to complete her work. Her present belief that she will live has good epistemic consequences : it is a means for her to acquire further knowledge. However, it is not a belief she ought to have, epistemically speaking ; it is unjustified. So the theory stated above misclassifies it.[14]

It is less clear that Berker's prediction holds good when we consider *indirect* deontic theories. Broadly, we may call "indirect" deontic theories those that prescribe beliefs in virtue of a *relation* to something of overall value. But more precisely, prominent indirect theories all prescribe beliefs in virtue of the overall value of the *process*, *disposition* or *rule* they result from.[15] These deontic theories are along the lines :

Believe *p on basis X* iff the (expected) overall value of *X* is above a

[14]The example targets the 'objective' ought claim, but we can adapt it to 'subjective' ones. We may suppose, for instance, that the set theorist knows that if she somehow manages to convince herself that she will live *ten more years*, that will keep her alive for the *six months* needed to complete her work and acquire much new knowledge. Hence the theorist may *expect* the belief to have good epistemic consequences ; yet it is not a belief she ought to have, epistemically speaking, since everything indicates that she will *not* survive ten years.

[15]There is a rough parallel between act- *vs.* rule-utilitarianism and direct *vs.* indirect theories. The theories require a theory of overall value for process types, dispositions or rules. Typically it is characterized in terms of effects of (actual and possible) instances of the process, manifestations of the dispositions or applications of the rule.

Berker's (2013a, 347) characterization of indirect theories is slightly different. On his account an indirect deontic theory comprises (a) a norm that directly prescribes *processes* (rules etc.) on the basis of their overall value, and (b) a norm that prescribes *beliefs* depending on whether they result from allowed processes (rules etc.). My characterization leaves (a) out and replaces "allowed" by "overall good enough" in (b). They are equivalent for present purposes.

certain threshold.

The argument sketched above does not apply to those theories. By the trade-off aspect of overall value, there will be *bases* with some bad consequences — but many good ones — that will have an above-threshold overall value. But why think that we will find unjustified beliefs with such bases ? Considering a few concrete cases will help.

If we are liberal about consequences, we will certainly find such cases. Take a case of unjustified belief resulting from a certain process *X*. Modify the case so that uses of *X* regularly but indirectly bring about good epistemic consequences. For instance, whenever one reads tea leaves, one gets in a good mood that greatly increases one's inferential abilities. In the resulting case process *X* has overall good epistemic consequences. But the belief based on it is still unjustified. As Berker (2013b, 374) notes, these cases are avoided by indirect theories that restrict the consequences relevant to overall value to *proximate* ones.

Berker (2013b, 374) puts forward a further type of case. The case targets epistemic consequentialist theories that use *true belief* as final value. It goes as follows : a man has a single process to evaluate whether a number is prime : namely, when presented with any number, he forms the belief that it is not prime. The process is quite dumb, but it produces a high ratio of true beliefs, given the relative rarity of primes among numbers. Hence it has an overall good value ; by the indirect deontic theory, the beliefs it produces are justified. But they are not. I agree, but it is not easy to generalize the example to theories that take *knowledge* as final value. For the envisaged process does not produce any knowledge : when the man forms the true belief that 8 is not prime, he is merely guessing. So knowledge-based indirect epistemic consequentialism does not have to ascribe the process any overall value. To get a parallel example with knowledge, we need a process or disposition that typically produces knowledge, but on one occasion produces an unjustified belief. It is not clear that there are such cases. *Prima facie*, if one forms one's belief in a manner that would typically yield knowledge, that one's belief would seem justified. To discuss precise examples would get us into unclear debates about what counts as "the" process by which a belief is formed. For present purposes it suffices to register the worry that Berker's case reveals a problem with the idea that reliability — in a sense relevant to justification — is merely a matter of ratio of true beliefs.

A perhaps more serious problem with knowledge-centred indirect consequentialism is the following.[16] Suppose that a process typically fails to produce knowledge, but sometimes does. An indirect consequentialist account may count the overall value of the process bad, and as a result the belief it produces as unjustified. In particular, those that constitute knowledge would nevertheless be unjustified. That goes against the common idea that knowledge entails justification. However, there may be independent reasons to reject it (Lasonen-Aarnio, 2010). Alternatively, one may as before doubt whether such cases are possible.

So while Firth and Berker's problem rule out direct epistemic consequentialist theories and truth-belief-centred ones, it is less clear that it arises for knowledge-centred indirect theories restricted to proximate effects. Such theories need not even adhere to the "separateness of propositions" (the idea that final epistemic value with respect to one proposition cannot be aggregated with final epistemic value with respect to another proposition) and the "separateness of times" (the idea that overall value at a time is only a matter of promoting final value *at that time)* that Berker takes to be necessary to avoid certain trade-offs.

Whatever we think of epistemic consequentialism, there are *non-consequentialist* ways of deriving epistemic norms from values. A simple one is :

> *S* ought to believe *p* at *t* iff *S*'s believing *p* at *t* has (would have) final epistemic value.

In conjunction with (MT) it follows that *S* ought to believe *p* if *S* knows *p*, or if *S* would know *p* were they to form the belief. So the theory gives us the Knowledge–Belief Principle.

But I fail to see how to derive the Knowledge–Action Principle from Meno's Thesis alone. Meno's thesis does imply that acting on knowledge has more value than acting on a belief that is not knowledge. For the first entails having knowledge, which is good, and the second entails having a belief that is not knowledge, which is bad. But consider the following pair of states of affairs :

(a) one knows that *p*, has a mere belief that *q*, and acts on *p*.

(b) one knows that *p*, has a mere belief that *q*, and acts on *q*.

[16] I owe the problem to John Hawthorne.

Meno's Thesis cannot count one state of affairs as better than the other. Both include a piece of knowledge and a belief that is not knowledge. The only difference between the two is that the action is caused by the piece of knowledge in one and the mere belief in the other. But that difference is not valued by Meno's Thesis. It need not have effects that are valued by Meno's Thesis either. So from Meno's Thesis alone we cannot derive different values to the two states of affairs. Without different values, it is hard to see how we could derive a deontic theory that prescribes the first and forbids the second. Of course we could simply build the Knowledge–Action Principle in our deontic theory ; but that would not be deriving norms from values.[17]

In sum, it appears possible to derive one of Zeno's norms from Meno's Thesis : namely, the Knowledge–Belief principle according to which one ought to believe only what is known. That can be done in a straightforward non-consequentialist way, and perhaps also in a consequentialist manner. But it does not appear possible to derive Zeno's other norm : the Knowledge–Action principle, according to which one ought to act only on what is known. It is not possible to do so because Meno's Thesis only ascribes value to knowledge, not to relations between one's action and knowledge.

3. From Norms to Value

Let us consider the opposite direction instead. We assume that Zeno's norms hold and try to derive Meno's thesis. But before we do this, it is worth spelling out the norms more carefully. We stated them as follows :

Knowledge–Belief Principle One ought to believe only what one knows.

Knowledge–Action Principle One ought to act only on the basis of what one knows.

But *ought* is a notoriously slippery term. It can be used to mean many things, so the claims above should be clarified. I will distinguish two dimensions of variation in what *ought* claims express.[18] First, they may vary along normative

[17]There are further loops one may go through in this argument, but I do not think they alter the conclusion. One may consider adding more assumptions about value. For instance, we may assume that some actions are good. Insofar as these actions are based on beliefs, the total state of affairs of doing those actions based on those beliefs would be better if the beliefs in questions constitute knowledge. Still, pairs like the one above may still be built.

[18]As far as semantics is concerned, we may assume that these variations correspond to various contextually-specified semantic values of "ought". The standard contextualist semantics of

sources. Normative sources are usually put under broad headings such as *moral*, *prudential*, *legal*, *aesthetic*, *all-things-considered*, and so on. But I think *ought* claims may reflect much more fine-grained sources, such as *what is prudent for a given task*, *what is prudent relative to health*, and so on. An attractive hypothesis is that normative sources correspond to values : each dimension or aspect of value is a source of *ought* claims. Second, *ought* claims vary along normative *layers*. A typical distinction of normative layer is the one commonly made between 'objective' *vs.* 'subjective' ought. The distinction is orthogonal to the previous one : if you have mistaken information about the laws, for instance, we can distinguish what you objectively legally ought to do from what you subjectively legally ought to do. The same goes for any other source of value. Thirdly, some *ought* claims are arguably not normative.[19]

The best reading of Zeno's norms, I claim, is that (a) they are normative, though perhaps hypothetically so ; (b) that do not express any specific normative *source*, but a normative *layer* ; (c) that the normative layer is expressed is distinct both from the traditional 'objective' and 'subjective' ones. Let me detail these points.

Genuinely normative

First, not all *ought* claims are normative. When *ought* is used normatively, there is something amiss with somebody who sincerely accepts that something ought to be so but does not in anyway favour its being so.[20] When it is used non-normatively, there is nothing amiss in doing so. In their most natural reading, the sentences below make non-normative claims :

The sky ought to be cloudy tomorrow morning.

Kratzer (2010) could be used. But we need not endorse such semantics here ; we can leave open how exactly the various things that *ought* claims express or convey correlate with the semantics of "ought".

[19] See *e.g.* Broome (2013, chap. 1).

[20] Some characterize an all-things-considered *ought* as the sense of "ought" which makes the following schema true : it is irrational to believe that you ought to ϕ without intending to ϕ (Broome, 2013, 22). Normative *oughts* may be characterized by a weaker schema : it is irrational to believe that *S* ought to ϕ without being *to some extent* in favour of *S* ϕ-ing. (As the phrase is used here, one can be to some extent in favour of something without being overall in favour of it.) For some expressivists the schema holds because believing that something ought to be so in the normative sense just *is* to have a favouring attitude towards it. I do not want to endorse this idea here ; perhaps there are cases where one sincerely believes that one ought to do something without in any way favouring it. I am content with the vague albeit clear enough idea that when one believes that they ought to do something, they would *normally* (they are expected to, meant to, supposed to) favour it to some extent.

> The plural of "mouse" ought to be "mouses". (Broome 2013, 9)

The first would normally be used to express what you expect to be the case — an "epistemic" reading of *ought*. It would not suggest that you somehow favour a cloudy sky. The second may be used not to express one's expectations about English nor one's recommendations for it, but to register instead a regularity.

The simplest view on Zeno's norms is that they are normative. Unfortunately, things are not so simple. For Zeno's norms may also be *hypothetical oughts*, which, if there are any, are neither of the straightforward normative type nor of the straightforward non-normative type. The idea is best illustrated with *have to*. Consider :

> How can one get to the sarcophagus ? — Well, it's not easy. You have to demolish the painted wall in the antechamber.

The dialogue may take place between two people to whom it is very clear that nobody ought in any sense to get to the sarcophagus. So the claim is not a straightforward normative *ought*. On the other hand, the claim has normative implications. For it clearly follows from what the second person says that *if one has to get to the sarcophagus*, then one has to demolish the wall. Thus the claim may be understood as a shorthand for the conditional form such as "if you want to get to the sarcophagus you have to demolish the wall". Similar phenomena may arise with *ought*. Call them *hypothetical oughts*.

Whether hypothetical *oughts* are normative or non-normative is moot. Conditionals of the form "If you want A, you ought to B" have at least in principle two readings, often labelled "narrow-scope" and "wide-scope".[21] On the first reading, the claim is that if some condition obtains (you want A), some norm holds (you ought to B). On the second, the claim is that a norm holds, whose content is : (either you do not want A or you B). On the first reading, the claim is strictly speaking *not* normative, though its combination with additional claims may entail something normative. On the second reading, the claim *is* normative. It forbids a certain combination of attitude and action. The two readings would arise for *ought* claims that are implicitly hypothetical, if there are any.

[21] Broome (1999) ; see Kolodny (2005) and Broome (2013) for further discussion. Talk of scope should not be taken too literally. The two readings may be achieved by several linguistic mechanisms : for instance, one can get the "narrow-scope" reading by having a wide-scope *ought* whose domain of quantification is restricted by the *if* clause.

Now some philosophers would treat vast ranges of *ought* claims as hypothetical. Some would treat all *prudential ought* claims as hypothetical; some would treat all *pro tanto ought* as hypothetical. On a wide-scope account, some could even treat all *oughts* as hypothetical, that is, they could hold that correct *oughts* claims all bear on combinations of attitudes and actions.

I do not want to take a stake on such views. I want to leave open whether Zeno's claims are of the simple normative kind or of the hypothetical one. Since the later may turn out to be not strictly speaking normative, I leave open that Zeno's claims are not strictly normative. All that matters here is that they are no less normative than *e.g.* ordinary prudential *ought* claims are.

A distinct normative layer

Second, normative layers. Let us first illustrate the common distinction between "subjective" and "objective" *ought*. A doctor has a patient with a well-known disease. There are two treatments for it, the old and the new. The old has strong side-effects and is now almost entirely out of use. The doctor naturally prescribes the new. But the patient turns out to have an hitherto unknown allergy to it. The doctor then switches to the old and the patient is cured. Is the following true?

The doctor ought to have given the old treatment straight away.

We are pulled both ways. On the one hand, the right treatment for the patient was the old one. So the doctor ought to have given it straight away. If we had knew in advance of the patient's allergy, that is what we would have told the doctor, and it would have been correct for us to do so. On the other hand, the doctor did what she should have done. Since the new treatment is better, and there was nothing to indicate that the patient would react badly, she had to give the new treatment. Indeed it would have been inappropriate for her to give the patient the old one. So what ought the doctor to have done? A common answer to the puzzle is to distinguish two senses of *ought*, called "subjective" and "objective". What one *objectively ought* to do is what one ought to do in view of the facts. What one *subjectively ought* to do is what one ought to do in view of one's information or one's perspective on the facts. From the doctor's original perspective, the right action was to prescribe the new treatment. But in view of the facts, the right action was to prescribe the old treatment straight away.

There are two misconceptions to avoid here. The first is to think that subjective *ought* claims are not normative. For instance, one may think that "*S*

subjectively-ought to *F*" is roughly equivalent to "*S* believes that they objectively-ought to *F*". The fact that it expresses *S*'s belief about what they in fact ought to do would explain why we expect *S* to act accordingly. But the fact that it merely expresses *S*'s *belief* about what they ought to do would mean that it is not normative. But that picture of the relation of the two *oughts* is wrong. To see this, it is best to consider a case where the two come apart *and the subject knows that they do*. Regan's Mine Shafts story (1980, 265n1, see also Parfit 2011, 159) is one such case. Ten miners are trapped either in shaft *A* or shaft *B*, but we do not know which. The water is rising, and we have three options : open gate *A*, open gate *B*, or open both. If we open only the gate of the shaft where they are, they will all die ; if we open the gate of the other shaft, they will all be save. If we open both gates, one of them, but only one, will die, no matter what shaft they are in. In that case we know that what we objectively ought to do is either to open gate *A* or to open gate *B*. It is *not* to open both gates. But arguably what we subjectively ought to do is to open both gates. Since we do not know which shaft the miners are in, we must minimize risk and avoid the death of all. The cases illustrates several points about the relations between 'subjective' and 'objective' *ought* claims. First, it shows that we subjectively ought to do is not what we believe we objectively ought to do : for we do know that closing both shaft is *not* what we ought to do in view of the facts. Second, it shows that there is something genuinely normative about 'subjective' oughts : there is a clear normative sense in which we ought to close both shafts. It is neither the expression of some non-normative standard nor a mere appearance or illusion. Third, it shows that there is something genuinely normative about 'objective' *oughts* : if we learned that we objectively ought to do close shaft *A*, then that would become what we subjectively ought to do as well. There is a (hard to specify) sense in which 'objective' *oughts* prevail over 'subjective' ones wherever possible. So both *oughts* are genuinely normative. Now once we have said what *S* objectively ought to do and what they subjectively ought to do, it is tempting to react as follows : "granted, what *S* ought to do in view of the fact is this, and what they ought to do in view of their information is that, but what ought they to do in the end ? What ought they to do, *simpliciter* ?". But the question makes no sense ; the two *oughts* both hold, they are both normative, and they do not conflict.

The second misconception is that the 'subjective' and 'objective' *oughts* express different *sources* of normativity. We we are confronted with sources of normativity, there is *conflict* : what we owe to the state *vs.* what we owe to our family, what we ought to do for ourselves *vs.* what we owe to do to others, what we ought to do for the task at hand *vs.* what we ought to do for

our long-term goals, and so on. (Of courses two sources of normativity may prescribe the same thing ; but at least conflict may in principle arise.) Conflict is solved by compromise, prevalence of one norm, or even not resolved at all. But it always involves some considerations in favour of doing something and some considerations against that are balanced against each other. Nothing such arises with 'objective' and 'subjective' readings of *ought* claims. First, 'objective' *ought* is not one category of *ought* alongside *moral*, *legal*, *prudential* and so on. There is no situation where what we "objectively" ought to do is *F* but what we "prudentially" ought to do is not *F*. Rather, for each of the categories *moral*, *legal* and so on, there are objective *oughts*. It may be, for instance, that in view of the facts the *legal* thing to do would be *F* but, still in view of the facts, the *prudent* thing to do would not be *F*. The same holds for 'subjective' *ought* : it is not one category alongside *moral*, *prudential* and so on. One may be tempted to think so, if one calls it the *ought* of rationality ; one could think that in some cases we have a conflict between what is *rational* to do and, say, what is *morally* right to do. But I think this is a confusion. For each normative source such as the legal, the moral and so on, there is a rational way to pursue it ; to each of these correspond distinct 'subjective' *ought* claims. The kind of cases where we seem to pit morality against rationality are in effect cases where we pit what is morally required against what is prudentially required — for instance, what we subjectively ought to do, morally speaking and what we subjectively ought to do in view of our interests alone. Second, 'objective' *ought* and 'subjective' *ought* are not such a conflict with each other. Suppose we observe somebody caught in a dilemma between two moral duties. We will see the two duties in opposition ; we will often look for a compromise ; if one duty prevail, we will still feel the force of the considerations brought by the other. We may say, for instance : "On the one hand she must be true to her mother ; on the other hand she should not break the promise made to her sister to keep their secret ; the only way for her to do both is to avoid talking to her mother at all ; but if it came to that, she would have to betray her sister." Contrast when we observe a "conflict" of 'objective' *ought* and 'subjective' *ought*, as in the Mine Shafts case. We do not say : "On the one hand the miners are in *A* and she objectively ought to open gate B ; on the other hand she subjectively ought to open both gates". We do not try to find a compromise between the two *oughts*. If we are about to offer advice, the 'objective' *ought* alone will matter and considerations of what the agent subjectively ought to do at that time will have no force whatsoever. If we are discussing whether the agent acted in a stupid or evil manner, the 'subjective' *ought* alone will have weight. Each corresponds to a distinctive layer of normative claims. They are

not different sources in conflict within a single layer.

The best way to understand Zeno's norms is that they intend to capture a distinctive *layer* of normativity. The 'subjective' *ought* is supposed to correspond to what is best in view of one's 'information' or from one's 'perspective'. But there are various notions of 'information' or 'perspective' to consider. In our original example, giving the new treatment was what the doctor ought to do in view of what they *knew*, but also what they *believed*. But the two can come apart. In some cases, the doctor irrationally believes that the patient will respond well to the new treatment. Doing so would then be what she ought to do *in view of what she believes* but not in view of what she knows or rationally believes. In some cases, what the doctor ought to do in view of what they *know* may differ from what they ought to do in view of what they *rationally believe*. That may be so, for instance, if all the doctor as ever heard about the new treatment is in fact a fabrication ; she rationally believes it, but there is nothing to it. If so in view of all that she knows — namely, that the *old* treatment works — what she ought to do is to give the old treatment, even though in view of what she rationally believes it is to give the new one. Now one may argue that these difference correspond to distinct but genuine normative *layers*. If someone does what she believes to be best, without rationally believing that it is best, then there is a sense in which they do what they ought to be doing and a sense in which they do not. If someone does what she rationally believes to be best, without it being best in view of what they know, then, it is argued, here as well there is a sense in which they do what they ought to be doing and a sense in which they do not.

So the most charitable way to assume Zeno's norms is to grant that there is a distinctive *layer* of normativity about which they hold. There is a sense of *ought* in which one ought to believe only what one knows and one ought to act only what one knows. It is distinct from some 'subjective' *oughts*, such as (some notions of) rationality : it is sometimes rational to act on something we do not know, for instance when everything misleadingly indicates that we know it.[22] It is not an 'objective' *ought* : in the Mine Shaft case, it matches the 'subjective' one instead.

[22]Philosophers sometimes distinguish "procedural rationality", which is merely a matter of having coherent attitudes and "substantial rationality", which requires more, *e.g.* having beliefs that fit the evidence and desire things that are worth desiring. These would also correspond to distinct layers of rationality ; the one is clearly distinguished from a knowledge-based normative layer ; the second less clearly so.

No distinctive source

We should not read Zeno's norms as expressing a distinctive *source* of normativity. That is most defensible for the Knowledge–Action Principle. We may imagine a case where *in view of what you know*, you morally ought to help someone but it is not in your best interest ; while *in fact*, you morally ought not to help them but it is your best interest. That may happen for instance if an eccentric rich man pretends to be in dire poverty and need your help. In view of what you know, the moral thing to do is to help them, though it is not in your interest. In view of all relevant facts, there is no moral requirement to help him, though doing so will happen to bring you a hefty reward. In such a case we do not have a knowledge-based *ought* that enters in conflict with a moral one and a prudential one. Rather, morality and prudence each generate their knowledge-based ought alongside their 'objective' one. The requirement to act on what you know is a distinctive normative *layer*, not a normative source.

With the Knowledge – Belief principle, the claim is more debatable. Consider standard cases of believing for practical reasons. An athlete knows that they have no chance to win the race, but they also know that believing that they will win will improve their time. One may feel a conflict analogous to a conflict of prudence and morality here. What the athlete *epistemically* ought to do is to believe that they will not win ; what they *prudentially* ought to do is to convince herself that she will win. The two *oughts* conflict ; and both are at the layer of knowledge-based *oughts*.[23] So one may be tempted to count the Knowledge – Belief principle to reflect a particular normative source.

If we did so we would get Meno's thesis fairly straightforwardly. An attractive hypothesis about normative sources is that they all reflect values. What one ought morally to do derives from what is morally good, what one ought prudentially to do derives from what is good for one, and so on. More precisely, what one 'objectively' X-ly ought to do is what is X-ly good ; what one X-ly ought to do in view of what one beliefs is what has higher expected X-ly value, where the expectations are given by one's beliefs ; what one X-ly ought to do in view of what one beliefs is what has higher expected X-ly value, where the expectations are given by one's knowledge ; and so on. Now if the hypothesis holds, one 'objectively' ought to believe only what one knows

[23] To parallel the foregoing case, one can imagine a variant where *in view of what the athlete knows*, they have no chance to win but believing that they do is likely to improve their time, while *in fact* they are likely to win but the belief would lower their chances. We seem to have an epistemic *vs.* prudential conflict both at the 'objective' and knowledge-based layer, and the two layers are distinct.

if and only if it is bad to believe what one does not know. So the norm would entail (part of) Meno's Thesis.[24] The resulting set of claims is virtually indistinguishable form the non-consequentialist derivation of the Knowledge–Belief principle we examined earlier. While we would have derived Meno's Thesis, it would not be clear, however, that we would have explained it. For one may think that the hypothesis holds because values ground norms ; if so, we are in effect explaining the Knowledge–Belief Principle by Meno's Thesis and not the opposite.

Be that as it may, I will focus on another construal of the Knowledge–Belief Principle. It is best seen by rewriting the principle thus :

Knowledge–Belief Principle (rewritten) One ought to believe only what is true in view of what one knows.

We start with the idea that *having true beliefs and not having false beliefs* is a normative source ; that is, something that a source of *oughts* that may in principle conflict with moral, prudential, legal considerations and so on. The source will generate various layers of *oughts*. Roughly : that one 'objectively' ought to believe what is in fact true ; that one purely subjectively ought to believe what is true in view of what one believes ; and that one 'knowledgeably' ought to believe what is true in view of what one knows. The latter is the Knowledge-Belief Principle. It reflects *both* a normative source, in its requirement of believing the truth, and a normative layer, in its focus on what one ought to believe in view of what one knows. To take the dimension of layer apart, we may focus on a more general principle :

Generalized Knowledge–Belief Principle One ought to form one's belief in view of what one knows.

The principle generates epistemic oughts, when combined with the idea that one ought to form true beliefs, but also prudential oughts, when combined with the idea that one ought to form useful beliefs, for instance.

Back to the derivation

The Knowledge–Action principle and the Generalized Knowledge–Belief principle delineate a significant normative role for knowledge. Can we derive

[24] To entail Meno's Thesis we should add the positive principle that one should believe what one is in position to know.

from them the claim that knowledge is better than belief that is not knowledge? Engel (2009) takes the answer to be negative, but does not discuss why.[25] *Prima facie*, there seems to be a way. First, we assume that believing the truth is good and that acting in the light of true propositions is good. These are assumptions about value, but distinct from the straightforward assumption that knowledge itself is good. Second, it follows from Zeno's norms that in order to do these good things, we are in some sense required to have knowledge — and not merely true belief. (The sense in which we are required to do so is the one that corresponds to the sense *ought* has in Zeno's norms.) From this, it seems, we can conclude that knowledge is better than true belief.

The derivation has some appeal. It seems plausible that (normative) conditions for doing good things are themselves good. If knowledge (normatively) allows you to form beliefs and act on them, then insofar as these beliefs and acts are good, knowledge would seem to be a good thing.

As it stands, the derivation fails. It is not in general true that (normative) conditions for doing good things are themselves good. Apologizing for one's faults is good; one ought to apologize for one's mistakes only if one made mistakes; but making mistakes is not good. That being said, there are undeniably many cases where conditions for doing good things seem good, and seem good precisely in virtue of being such. One would need a restricted version of the principle. If the restricted version applies to the case of knowledge, the derivation would succeed.

The route from Norms to Values is more promising. Whether it ultimately succeeds depends on whether we can find a plausible motivation of the idea that knowledge is good because it is normatively required to do good things. It is worth stressing that, if it is ultimately successful, it would yield a picture of the value of knowledge that is at odds with much of the current literature. Much current literature tries to show that knowledge is valuable by showing that it is a worthwhile thing to aim at, either as an end or as means to some end. On the present perspective, knowledge would be good because it is normatively required to aim at anything.

[25] Engel (2009, sec. 5) calls the idea that knowledge is relevant to what we ought to do "pragmatic relevance" and writes: "Of course if the phenomenon of pragmatic encroachment reduces, as I have claimed, to that of pragmatic relevance, question (2) [the question whether pragmatic encroachment explains why knowledge is valuable] has to be answered in the negative". He does not discuss the idea further, but assumes that it is only if pragmatic encroachment is *more* that something like the Knowledge–Action principle that it could explain the value of knowledge.

4. Doing Without Value

We started with a pair of ideas : that knowledge has value, and that it enters certain norms. We have tried to derive one idea from the other. I have argued that the norms cannot be derived from the value and I have expressed doubts as to whether the value can be derived from the norms. Now I want to stress another perspective : namely, that once we assume the knowledge-involving norms, it is unclear whether there is any way to motivate the idea that knowledge has value. I have made the suggestion elsewhere (Dutant, 2012 ; forthcoming). Here I will discuss a couple of objections to it.

Why should we think that knowledge has more value than true belief ? Many philosophers treat it as a (at least *prima facie*) platitude in need of no defence. To them, it is (at least *prima facie*) obvious that there is something commendable about believing something when you know it that there is not about believing it when you do not know it. But note that that is explainable on the basis of the Knowledge – Belief principle alone. If that principle holds, then there is a sense in which you do what you ought to do when you believe p while knowing p and you do not when you believe p while not knowing p. That is enough to explain that there is something commendable about knowledge that mere true belief lacks. That does not require or entail that there is something genuinely *better* about it. What we ought to do and what is good may come apart. If you face a choice between A and B, if A is bad but everything indicates that it is good and better than B, then there is a sense in which you ought to choose A, but we may deny that choosing A is genuinely good.[26]

Some will think that doing what one ought to do has value in itself.[27] They would endorse the straightforward derivation of Meno's thesis from the Knoweldege-Belief principle that we sketched earlier. But there is no need to accept that view. In particular, normative *sources* and *levels* may stand differently in their relation to value. There is some plausibility that normative *sources* reflect different values : the morally required derives from the morally good, the prudentially required derives from the personally good, and perhaps the epistemically required derives from the value of believing the truth.

[26] We use adjectives like "good" fairly liberally. There are very natural uses on which "you should do that" and "doing that would be good" are virtually interchangeable. With these uses we may well say that choosing A was the "good" choice. But in laying out a theory of value and norms one may need use "good" and "value" more strictly. Strictly speaking, choosing A was not good ; it just *seemed* good. That it seemed good made it the choice you ought to make ; but that need not itself make it genuinely good.

[27] Piller (2009) pursues this line to explain the value of knowledge.

But normative *layers* do not require additional values. Rather, each normative layer correspond to a different way to derive an *ought* from a value. Given moral goods, there is when we objectively ought to do for the moral good ; what we ought to do for the moral good in view of what we believe ; what we ought to do for the moral good in view of what we know ; and so on. We have argued that the Knowledge–Action principle and the Generalized Belief–Action principle express a normative layer and no particular normative source. If that is so, they need not be associated with any value at all.

Do we have any other reason to think that knowledge is valuable ? Knowledge is definitely something that matters. But as we have seen, its normative role is enough to explain that it does.

The best reason to think that knowledge has value is, I think, the idea that knowledge is worth aiming at. We want knowledge ; we strive for it and we are ready to make sacrifices for it. We are not foolproof ; but assuming that we are right in this, knowledge is something we ought to *aim at*. Conversely, a theory that denies that knowledge is something we ought to aim at has to claim that we are misguided in that respect. If knowledge is good, we have a straightforward explanation of why we ought to aim at it. For in general, we ought to aim at good things. Conversely, if knowledge has no value, it becomes doubtful whether we should aim at it.

The idea raises a difficulty for the view that knowledge plays a important normative role but has no value. Zeno's norms alone do not entail that we should *acquire* knowledge. One may comply with both norms by avoiding belief and action altogether. Less radically, one who complies with the norms at one point may comply with them onwards by not acquiring any new belief and acting only on what they already know. In reply, I would point out that the norms entails requirements to acquire knowledge *in conjunction* with other norms or values. For instance, if we assume that it is better to act on more relevant facts, we may derive that one ought to acquire more knowledge of relevant facts. Similarly, we may assume that it is good to have true beliefs on a range of topics ; if so, we may derive that one ought to acquire more knowledge. It is not trivial to work the reply out properly. For instance, it is not obvious that it is *always* better to act on more relevant facts. For our purposes, it is sufficient to show that there are some ways to ground requirements to acquire knowledge that do not assume that knowledge is itself valuable.

Another difficulty for the view that grants knowledge a normative role but no value is worth discussing. The view assumes Zeno's norms, or something like them. The assumption is not trivial. As we have argued, the best way to understand them is to postulate a normative layer distinct from the tradi-

tional 'objective' and 'subjective' ones. One may fell uneasy about the very idea of layers of normativity, and about the idea that knowledge-based *ought* claims delineate a distinctive layer. In reply, I will make a couple of points. First, as I have argued, it is difficult to make sense of cases such as Regan's Mine Shafts case without accepting that there are at least two distinct layers of normativity ; each genuinely normative but not conflicting. Second, once we have admitted the idea of layers, it is easy to see how they multiply. Within a belief-based layer alone, we can often distinguish opposite *ought* claims derived from various natural subsets of one's beliefs. For instance, one may have a set of salient beliefs in view of which one ought to do A ; but one may at the same time have some deeply buried beliefs, such that in view of them and the salient ones, one ought not to do A. In such cases, an onlooker may feel both the pull of "they ought to have done A" and the pull of "they ought not to have done A". Or again, one may impeccably infer a conclusion from a set of crazy premises. In such a case we may feel both the pull of "they ought to have inferred that conclusion" and of the opposite ; one is what one ought to do in view of one's premise beliefs, the other what they ought to do in view of some broader background. The divisions may be multiplied : there may be cases where one's conclusion is correct in view of the premises, insane in view of the premises of the premisses, correct again in view of the premisses of the premisses of the premisses, and so on indefinitely. I see no reason to reduce the profusion of these normative layers to one or two. As long as each *ought*-question we care about manages to pick up a specific enough layer, we can leave with many *oughts*. On the backdrop of these many normative layers, the knowledge-based one is not a cost.

Summing up, once we grant that knowledge plays a central normative role along of the lines of Zeno's norms, it is not clear that there is anything left to motivate the idea that knowledge is of distinctive value. So Zeno's norms can be used to explain *away* Meno's thesis. In that perspective, the normative role of knowledge allows us to explain why it matters without having value.

5. Conclusion

I have highlighted two possible answers to the question why knowledge matters. One is that it has value. Another is that it plays a significant normative role. I have granted that if knowledge *had* value, or if it *did* play the alleged normative role, then it would matter. For most of the discussion I have remained neutral on whether knowledge has value or does play that normative

role. My central question has been instead whether we can derive one idea from the other. That is, whether assuming the idea that knowledge has value — and some defensible general hypotheses about norms and values —, we could derive the claim that it plays the alleged normative role. Or whether, assuming that knowledge does play that role — and some defensible general hypotheses —, we could derive the claim that it has value. I have found the route from Value to Norms unsuccessful. The main problem here is that the idea that knowledge has value does not seem enough to derive the idea that one should act on what one knows. I have found the route from Norms to Value more promising, though a complete path is missing. The main idea here is that knowledge is good because it is normatively required to do good things, such as believing the truth and acting in view of true propositions. But since not all normative condition for doing something good is itself good, we still lacked an explanation of why knowledge would be so. Finally, I have suggested an alternative perspective, on which we would not try to derive the idea that knowledge has value from its normative role, but rather use its normative role to explain away the idea that it has value. The general idea is that if knowledge does play the normative role in question, then the fact that it does explains while knowledge *seems* to be something that has value. But there is no need to think that it has ; all that matters about knowledge could be explained by its normative role.

6. References

Selim Berker. Epistemic teleology and the separateness of propositions. *Philosophical Review*, 122(3):337–393, 2013a.

Selim Berker. The rejection of epistemic consequentialism. *Philosophical Issues*, 23:363–87, 2013b.

John Broome. Normative requirements. *Ratio*, 12:398–419, 1999.

John Broome. *Rationality through Reasoning*. Wiley-Blackwell, 2013.

Edward Craig. *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford University Press, 1990.

Julien Dutant. The value and expected value of knowledge. *Dialogue*, 51(01): 141–162, 2012.

Julien Dutant. In defence of swamping. *Thought*, forthcoming.

Pascal Engel. Pragmatic encroachment and epistemic value. In Adrian Haddock, Alan Millar, and Duncan Pritchard, editors, *Epistemic Value*. Oxford University Press, 2009.

Roderick Firth. Are epistemic concepts reducible to ethical concepts. In Troyer, editor, *In Defense of Radical Empiricism: Essays and Lectures by Roderick Firth*, pages 237–249. Rowman & Littlefield, Lanham, Md., 1998a.

Roderick Firth. Chisholm and the ethics of belief. In John Troyer, editor, *In Defense of Radical Empiricism: Essays and Lectures by Roderick Firth*, pages 143–155. Rowman & Littlefield, Lanham, Md., 1998b.

Roderick Firth. Epistemic merit, instrinsic and instrumental. In John Troyer, editor, *In Defense of Radical Empiricism: Essays and Lectures by Roderick Firth*, pages 259–71. Rowman & Littlefield, Lanham, Md., 1998c.

Michael Frede. Stoics and sceptics on clear and distinct impressions. In *Essays on Ancient Philosophy*, pages 151–176. Clarendon Press, 1983/1987.

Michael Frede. Stoic epistemology. In K. Algra, J. Barnes, Mansefld J., and Schofield M., editors, *The Cambridge History of Hellenistic Philosophy*, pages 295–322. Cambridge University Press, 1999.

H.J. Hankinson. Stoic epistemology. In B. Inwood, editor, *Cambridge Companion to the Stoics*. Cambridge University Press, 2003.

Mark Kaplan. It's not what you know that counts. *Journal of Philosophy*, 82: 350–363, 1985.

Thomas Kelly. Epistemic rationality as instrumental rationality: A critique. *Philosophy and Phenomenological Research*, 66:612–640, 2003.

Niko Kolodny. Why be rational? *Mind*, 114(455):509–563, 2005.

Angelika Kratzer. The notional category of modality. In *Collected Papers on Modals and Conditionals*. Oxford University Press, 2010.

Maria Lasonen-Aarnio. Unreasonable knowledge. *Philosophical Perspectives*, 24(1):1–21, 2010.

Anthony A. Long and David Sedley, editors. *The Hellenistic Philosophers*, volume 1. Cambridge University Press, 1987.

Derek Parfit. *On What Matters*, volume 1. Oxford University Press, 2011.

Charles S. Peirce. *The Philosophy of Peirce: Selected Writings*. Harcourt Brace, 1950.

Philip Percival. Epistemic consequentialism: Philip percival. *Supplement to the Proceedings of The Aristotelian Society*, 76(1):121–151, July 2002.

Christian Piller. Valuing knowledge: a deontological approach. *Ethical Theory and Moral Practice*, 12:413–28, 2009.

Donald Regan. *Utilitarianism and Cooperation*. Oxford University Press, 1980.

Robert Stalnaker. Epistemic consequentialism: Robert stalnaker. *Supplement to the Proceedings of The Aristotelian Society*, 76(1):153–168, July 2002.

Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, 2000.

16

Construction ou critique ? Carnap et Kant sur le concept de synthèse

KATSUYA TAKAHASHI

1. Une théorie de la synthèse sans synthétique *a priori*

Dans sa *Construction logique du monde* (*Der logische Aufbau der Welt*), Carnap dit à plusieurs reprises que la théorie avancée dans son oeuvre traite du processus synthétique, non pas celui analytique, de la connaissance de l'objet (*Aufbau*, sec.68, 69, 74, 83,100). Cette théorie, nommée théorie de la constitution, envisage de présenter et illustrer l'idée d'un système dans lequel tous les objets (ou tous les concepts) se redéfinieraient comme « constitutions » logiques réductibles à certains concepts fondamentaux. La procédure de la constitution est un analogue rationnel du processus effectif de la cognition consistant à construire divers concepts d'objet à partir des donnés empiriques. Elle concerne « le traitement du donné pour former et représenter les choses, la réalité ». Or ce dernier n'est autre chose que ce que les philosophes appellent depuis Kant la « synthèse cognitive » (sec,100).

Il existe donc une certaine parenté avec le kantisme dans la théorie carnapienne de la constitution. Comme Michael Friedman le fait justement remarquer, la théorie dans l'*Aufbau* incarne, outre la motivation empiriste souhaitant réduire toute connaissance à des donnés empiriques (avec l'aide de la logique), une autre motivation qui viennent de la tradition kantienne : celle désirant décrire en détail le fonctionnement du pouvoir rationnel et constructif de cognition qui impose ses propres formes aux donnés empiriques (Friedman, 1999, p.125-126, pp.140-141).

Cependant, la parenté avec le kantisme ne doit pas être surestimée, car l'*Aufbau* exclut de sa théorie une idée qui était essentielle au kantisme, à savoir, l'idée de la connaissance *a priori* synthétique (sec.106, 179). Carnap déclare que la théorie de la constitution n'accorde aucune place aux propositions synthétiques *a priori* (sec.106,179). De ce fait, on peut dire que cette théorie est une théorie de la synthèse qui veut se passer de l'idée du synthétique *a priori*.

Comment Carnap a-t-il pu concevoir une théorie de la synthése sans introduire le synthétique *a priori* ? Est-il possible de concevoir en général une théorie de la synthèse sans synthétique *a priori* ? Cet article envisage d'affronter ces questions, mais cela dans une certaine limite. Nous aborderons les questions, notamment la dérnière, par l'intermédiaire de la comparaison de Kant et de Carnap plutôt que par la lecture immanente.

Est-il possible de concevoir une théorie de la synthése en excluant le synthétique *a priori*? Il y a deux voies par lesquelles on peut lire l'*Aufbau* du point de vue de cette question. La première consiste dans l'examen directe du système carnapien assumé pour savoir si ce système est vraiment cohérent et suffisant à la lumière de son propre objectif. Cette sorte de lecture a déjà été entreprise par un bon nombre de commentateurs dont certains semblent avoir montré des défauts non négligeables de ce système. Par exemple, on connaît les « difficultés » signalées par Nelson Goodman et les tentatives de les examiner, entreprises par d'autres auteurs éminents[1]. Nous trouvons fort signifiantes les problèmes de cette sorte, mais nous n'entendons pas les examiner concrètement ici. Nous nous contenterons de suggérer le rapport que ceux-ci peuvent avoir à notre conclusion, laquelle s'obtiendra plutôt à travers la deuxième voie.

La deuxième voie de l'examen, qui est celle que nous choissions, départ de la conception kantienne de la synthèse et de la problématique qu'elle envisage de résoudre. La problématique que Kant désirait affronter par sa théorie de la synthèse concerne, naturellement, le processus et la possibilité de l'élaboration des concepts, portant sur la réalité empirique, susceptibles de l'utilisation scientifique. Par exemple, on peut se demander comment il se fait que le concept de temps capable de formuler les lois physiques ou le concept de couleur maîtrisé par les artistes se construisent à partir des donnés sensibles. Cette question, portant sur la formation des concepts empiriques et rationnels, peut s'exprimer aussi comme celle de la correspondance, ou de la coordination (*Zuordnung*), entre l'empirique et le conceptuel. Celle-ci consiste à savoir com-

[1] Sur ce sujet, les discussions dans Vuillemin 1971, Proust 1986/1989 et Granger 1994 nous semblent surtout importantes.

ment on peut établir et s'assurer la coordination entre le système de concepts et la réalité empirique. Résumons désormais tous ces questions simplement par le terme « la problématique de la coordination ». Le fameux concept de synthétique *a priori*, conçu comme désignant les formes pures de la synthèse par Kant, est introduit pour expliquer et justifier la possibilité de cette coordination. Nous verrons que cette problématique est reprise dans la théorie carnapienne et est reformulée en nouveax termes. Notre tâche consistera alors à mettre en valeur la différence des deux philosophes à légard du concept de synthèse et à tenter de savoir quelles solutions ceux-ci sont censés être donner à la problématique de la coordination.

De la comparaison, il ressortira que le concept carnapien de synthèse ne parle pas de la synthèse mais plutôt de l'analyse. C'est pourquoi ce philosophe viennois a cru possible de rompre avec le synthétique *a priori* dans son épistémologie, dirons-nous. L'absence de la synthèse chez Carnap nous justifiera également de dire que son système de constitution n'atteste pas de la possibilité d'une synthèse sans synthétique *a priori*.

Cependant, ne faut-il pas dire alors que l'entreprise de l'*Aufbau* suggère la possibilité d'affronter la problématique de la coordination en excluant le concept de « synthèse » lui-même? À ce propos, la comparaison nous permettra le scepticisme en montrant que les approches des deux philosophes vers la problématique sont fort différentes et qu'il n'est pas évident que l'*Aufbau* ait résolu la problématique.

Ces résultats mettront en évidence la signification importante que peut avoir la théorie kantienne de la synthèse au sujet du problème de la coordination d'empirique et de conceptuel. Évidemment, il n'en suivra pas que la possibilité et l'existence du synthétique *a priori* soient soutenues. Nous aurons eu alors tout au plus le droit de poursuivre la défence de cette notion. La confrontation de l'*Aufbau* carnapienne et de la *Kritik* kantienne n'en est pas moins avantageuse pour le synthétique *a priori* ; en nous permettant de dire ce qui échappe à la formulation carnapienne de la problématique, la comparaison nous indique vers quelle direction le kantien aujoud'hui peut porter ses pas pour avoir une notion signifiante du synthétique *a priori*.

2. La constitution comme synthèse

La tâche qu'assume l'*Aufbau* est de nature épistémologique[2]. Mais cette oeuvre n'entend pas décrire le processus effectif de la genèse ou du développement de la connaissance. Il s'agit plutôt d'exposer un système dans lequel divers concepts d'objet auraient leurs définitions rigides en terme de certains concepts de base ainsi que de la logique mathématique. « Classe » et « relation » sont les concepts de base choisis par Carnap. Naturellement, toute classe a ses éléments. Le système désire donc réduire les concepts à des classes, à des relations et à certains éléments moyennant le langage formel. Il est à souhaiter que la réduction relie les concepts par degrés afin d'en former un arbre généalogique. Tous les concepts se rapporteront alors, directement ou indirectement, à des éléments qui sont fondamentaux. Or la motivation épistémologique exige que ces éléments soient empiriques et simples. Le système offrira ainsi la réduction des concepts d'objet à des donnés empiriques immédiats. Si l'on suit la procédure de la réduction dans la direction inverse, on aura la procédure de la « constitution ». Aussi le système de l'*Aubau* se nomme « système de constitution ». Comme nous l'avons dit, ce système est envisagé par l'auteur comme « reconstruction rationnelle (*rationale Nachkonstruktion*)» (sec.100, 143) du processus synthétique de la cognition. Il exprime par le langage formel les opérations telles que « la formation de l'objet, son identification ou sa classification en espèces » (sec.100). Quant à la « rationalité » des constitutions, elle peut être garantie, selon l'idée de Carnap, par la logique mathématique (ou la logistique) de Whitehead et Russell ; étant pourvue de la théorie des relations, cette arme est assez puissante pour formaliser les concepts scientifiques.

Quelques remarques terminologiques sont à ajouter avant de voir la « synthèse » carnapienne plus concrètement.

Premièrement, Carnap parle « tantôt des objets tantôt des concepts sans faire de différence essentielle » (sec.3). Cette indifférence tient à une idée de l'auteur selon laquelle à tout concept correspond son objet même s'il s'agit d'un concept dit « général » ou d'un concept de propriété. C'est que le terme

[2]Dans la Préface de la première édition de *l'Aufbau*, Carnap dit : « il s'agit avant tout ici de la question de la théorie de la connaissance, par conséquent de la réduction des connaissances les unes aux autres » (fr.p.53). Mais il ne tardera pas à rejeter l'idée même de la théorie de la connaissance qui avait retenu son « premir livre important » (*Aufbau*, Préface de la seconde édition) dans le voisinage de la problématique kantienne. Dans le *Logische Syntax der Sprache*, il refuse d'employer le terme « théorie de la connaissance (épistémologie) » dans la crainte que sa propre tâche, baptisée maintenant « la logique de la science », soit confondue avec l'épistémologie traditionnelle (Carnap, 1934/1937,pp.279-280).

objet est employé dans l'*Aufbau* dans son sens le plus large ; il désigne « tout ce sur quoi peut porter une proposition (*Aussage*) » (sec.1).

Deuxièmement, le système de constitution exposé dans l'*Aufbau* n'est pas l'unique système possible de la constitution. Il n'est pas imposssible, souligne Carnap, de concevoir différents systèmes construits dans le but similaire. Ceux-ci se différeraient selon diverses conditions qu'on décide de choisir. Ce qui caractérise avant tout le système de l'*Aufbau*, c'est la décision à l'égard des éléments fondamentaux à partir desquels on va constituer d'autres objets. Carnap suggère la possibilité de considérer des entités physiques comme fondamentaux mais il décide, par respect pour l'intérêt épistémologique, de commencer à partir des donnés psychiques rencontrés dans la conscience d'un individu quelconque. Conformément à cette décision, les relations qu'on doit supposer reconnaissables parmi les éléments fondamentaux sont également à déterminer ; ce choix-ci peut aussi influencer la structure du système. On verra Carnap conclure qu'une seule relation (nommée « le rappel de ressemblance ») suffit pour en faire dériver les autres relations importantes et pour constituer divers types d'objets. Suivant son ambition réductrice, il prend cette seule relation pour relation fondamentale. Ces décisions constituent ensemble le problème de la « base » du système. Il nous faut donc retenir que le système de l'*Aufbau* est fondé sur une base vonlontairement choisie. Nous allons entendre désormais par « le système de constitution » ce système choisi dans l'*Aufbau* et non pas le système de constitution en général.

Le problème de la base nous incite à relier la théorie de la constitution à la théorie traditionnelle de la synthèse, car ce problème semble concerner la question sur les matériaux et les formes du traitement conceptuel dans la connaisance. De fait, Carnap explique parfois l'idée de la constitution en termes kantiens. De même que la synthèse a ses matéraux et ses formes, on peut parler des matériaux et des formes de la constitution, dit-il. Quelles sont alors les matériaux et le formes de constitution en tant que « synthèse » ?

Pour avoir la réponse à cette question, il nous est nécessaire de prendre en compte le fait que la constitution comprenne différents niveaux, à partir le niveau fondamental vers les niveaux supérieurs, dont chacun peut également être appelé « constitution ». En d'autres termes, la constitution du système entier se compose des constitutions particulières. Les « formes de la synthèse » chez Carnap sont donc les formes générales qui rend possible à chaque niveau la constitution des objets depuis les objets du niveau précédent. Ces formes générales se nomment « les formes des niveaux (*Stufenformen*) ». D'après Carnap, classes et relations jouent cette fonction. En effet, la constitution des objets d'un niveau se réalise toujours comme classification des objets du niveau

précédent ou comme mise en ordre de ces derniers par l'application d'une relation quelconque. Une fois que de nouveaux objets se constituent ainsi comme nouvelles classes ou nouvelles relations, ils peuvent servir d'éléments (ou de membres) destinés à des constitutions des niveaux supérieurs. Cependant, Carnap dit que seul le concept de relation peut être dit « forme de la synthèse » au sens propre parce que les classes peuvent se réduire en fin de compte aux éléments fondamentaux, lesquels sont plutôt les matériaux de la synthèse. Plus exactement, seules un petit nombre de relations qui sont fondamentales doivent être remarquées : la resseblance partielle, l'dentité partielle, la relation mémorielle, etc. Il n'est pas impossible d'accorder aux relations fondamtentales la qualification de « catégorie », dit Carnap (sec.83). Comme nous l'avons déjà dit, la déduction du nombre des relations fondamentales peut être poursuite encore plus loin jusqu'à ce que seule une relation reste : le rappel de ressemblance. Cette relation est sans doute l'unique catégorie « authentique », dit l'auteur. Le système de constitution se base effectivement sur cette relation au final. Cependant, Carnap fait noter que cette conclusion ne promet aucune réponse défintive à la question des catégories et que l'*Aufbau* ne peut parler que d'une supposition à ce sujet.

Quant aux matériaux de la constitution, les éléments du chaque niveau sont en principe dignes de cette qualification, mais les matériaux ultimes sont, naturellement, les éléments fondamentaux. Les éléments fondamentaux du système de l'*Aufbau* sont les donnés psychiques de la conscience d'un individu. Chaque élément est une expérience qu'a un sujet à un instant. Cette expérience crue, nommée « vécu élémentaire », fait une unité dont les constituants ou les caractéristiques ne sont pas isolées.

Comme un vécu élémentaire est une unité indivisible, la tâche de la première constitution consiste à comparer des vécus et à former par abstraction les concepts de constituants, tel que le concept de la qualité sensible rouge. Le système suppose qu'on ait en vertu de la comparaison une large liste des paires dont chacune indique deux vécus élémentaires qui sont en relation de « rappel de ressemblance » l'un avec l'autre. Le rappel de ressemblance est l'unique relation fondamentale à partir de laquelle Carnap désire dériver d'autres relations importantes. Mais qu'est-ce que c'est que le rappel de ressemblance ? Cette relation est là lorsqu'on reconnaît une ressemblance partielle entre deux vécus élémentaires dont l'un s'est produit dans le temps passé et est conservé dans la mémoire. Par exemple, si un vécu contient l'apparition d'une couleur à un endroit et qu'on se rappelle un autre vécu qui contenait l'apparition d'une couleur similaire, alors on dit que les deux vécus se ressemblent partiellement et sont en relation de rappel de ressemblance.

(Il n'est pas demandé que les endroits où ces couleurs apparaissent dans les vécus soient identiques.) En profitant de cette liste, on pourra classifier les vécus selon les groupes qu'ils font et définir par là divers constituants portés par eux. Telle est la procédure de la première constitution, qui est appelée « quasi-analyse ». On peut dire que la quasi-analyse est la première rencontre qui doit se dérouler entre les formes et les matériaux de la contitution. La « quasi-analyse » semblerait désigner la procécure consistant à décomposer des vécus en leurs constituants, mais elle opère en réalité la contitution des nouveaux objets qui sont dits d'habitude « qualités sensibles ». De même que toutes les constitutions ultérieures, elle est une procédure synthétique : d'où la remarque carnapienne qui souligne que « *la quasi-analyse est une synthèse qui revêt la forme linguistique d'une analyse* » (sec.74).

En résumant l'explication ci-dessus, nous pouvons décrire le déroulement « synthétique » de constitution de la manière suivante : on reconnaît d'abord les relations qu'entretiennent les éléments les uns avec les autres, et puis classifier les éléments selon ces relations ou crée de nouvelles relations en profitant des anciennes. Les classes et les relations ainsi produites sont appelées, dans le système de constitution, « objets » ou « concepts ».

3. L'exclusion du synthétique *a priori*

Bien que la constitution est une procédure synthétique ayant formes et matériaux, la théorie de la constitution ne soutient pas l'existence des propositions synthétiques *a priori*. Cela veut dire que le système de constitution ne contient que deux sortes de proposition : analytique et empirique. Voyons la raison de cette division dualiste.

Le système de constitution est un système formulé en terme d'un langage formel. Mais il ne faut pas prendre ce système pour un système déductif dans lequel toutes les propositions dériveraient logiquement à partir des axiomes. Il n'est pas une axiomatique mathématique mais un système de concepts offrant les formulations rigides des concepts empiriques et des lois empiriques fondamentales. Exprimées ainsi formellement, les concepts et les lois deviendront aptes au discours scientifique rigoureux. Telle est l'ambition de Carnap incarnée dans le système. Cette ambition s'éteint, à son tour, vers un idéal concernant la science en général : celui-ci rêve encourager et faciliter la collaboration des chercheurs de tous les domaines en intégrant les sciences particulières dans un ensemble bien organisé, nommé « la science unifiée (*ein Gesamtwissenschaft*) » (*Aufbau*, Préface de la première édition, sec.179). Dans

la science unifiée, tous les concepts apparaissant dans le discours scientifique seraient devenus clairs grâce à la définition rigide et à la systématisation. Le système de constitution, qui lui-même est encore ouvert à l'élaboration par les philosophes et les scientifiques, s'occupe de la partie formelle de la science unifiée. Ainsi, on se rend compte que le système de constitution est censé être servir d'un fondement pour l'application du langage formel aux sciences empiriques. Les lois logiques et mathématiques ne sont pas exposées mais présupposées dans ce système.

Pour remplir la fonction de médiateur, le système de constitution doit introduire et réconcilier deux facteurs hétérogènes, c'est-à-dire le formel et l'empirique, lors de l'élaboration des propositions. Comment les propositions de ce système se rapportent-elles à ces deux facteurs ? Les propositions du système peuvent être divisées en deux catégories : les définitions et les théorèmes. Les définitions ne sont autre chose que ce que Carnap appelle « constitutions ». Elles concernent des objets empiriques mais dépendent de notre convention, si bien qu'elles sont à qualifier d'analytique. Quant aux théorèmes, leur rôle consiste à indiquer l'introduction ou la dérivation des relations. Or nous avons vu que les relations dans ce système sont les formes de la « synthèse ». Ce sont donc les théorèmes qui peuvent entrer en question quand on discute de la question du synthétique *a priori*, car cette question concerne le statut des propositions affirmant quelque chose sur les formes de la synthèse. Selon ce que souligne Carnap, ses théorèmes impliquent aucune composante qui s'accorderait avec la qualification de synthétique *a priori*.

Les théorèmes se divisent, d'après lui, en deux sortes : les théorèmes analytiques et les théorèmes empiriques. Les théorèmes analytiques sont ceux qui sont déduits logiquement des définitions. Les théorèmes empiriques, qui déterminent les relations ou les structures des objets, ne peuvent être établis qu'au moyen de l'expérience (sec.106). Par exemple, une fois que la relation de ressemblance partielle est définie ou constituée à partir d'une relation plus fondamentale (à savoir celle de rappel de ressemblance), on aura la proposition : « la ressemblance partielle est symétrique » comme conséquence logique de la constitution. Cette proposition est un théorème analytique (sec.110). En revanche, la proposition : « la relation du rappel de ressemblance est asymétrique » est un exemple du théorème empirique (sec.108). « Le champ visuel est bidimentionnel » appartient aussi à cette dernière sorte. Outre ces deux sortes, il n'y a aucun d'autre théorème.

Certes, il reste encore les lois logiques et les lois mathématiques présupposées dans le système de constitution. Mais elles sont toutes analytiques selon l'avis de Carnap. En conclusion, la théorie de la constitution n'accorde

aucune place au synthétique *a priori* dans la connaissance.

> « Les jugements synthétiques *a priori* » qui sont à la base de la problématique kantienne de la théorie de la connaissance n'existent pas du tout du point de vue de la théorie de la constitution. (sec.106. cf.aussi sec.179)[3]

La conclusion ne changera pas même si l'on prend en considération la science unifiée incluant toutes les sciences particulières. Celles-ci n'ajoutent en effet que des propositions empiriques.

Nous venons de voir que la théorie de la constitution est une théorie de la synthèse qui exclut le synthétique *a priori*. Mais est-il vraiment possible de concevoir une telle théorie?

En général, la cohérence de la position carnapienne à l'égard de ce concept traditionel peut être examinée de divers points vue. Il est à s'interroger, par exemple, sur le statut du langage philosophique au moyen duquel Carnap présente toute cette réflexion. Sa conclusion sur le statut analytique des propositions mathématiques n'est pas non plus évidente. Mais nous désirons rester dans le concept de synthèse lui-même pour savoir si l'idée de la synthèse sans synthétique *a priori* peut être cohérente. Évidemment, une telle réflexion dépend de la conception qu'on a à propos de la synthèse. Nous croyons qu'il faut retourner à Kant à ce propos. Dans la section suivante, nous allons jeter un coup d'oeuil sur la théorie kantienne de la synthèse. Il nous est surtout important de savoir quelle nature la synthèse kantienne montre à la différence de son opposé, l'analyse, et à quelle problématique la théorie de la synthèse est censée être répondre.

4. L'intuitif et le discursif – problématique kantienne

Pour Kant, la « synthèse » signifie l'acte de notre pensée consistant à « ajouter les unes aux autres des représentations différentes » et à « saisir leur diversité en une connaissance ». Lorsqu'il s'agit de la connaissance de la réalité empirique, la « synthèse » signifie l'acte qui parcours le divers, donné dans la sensibilité, et le lie pour en faire une connaissance (A77/B102). La « connaissance » formée par cette opération peut être encore brute et confuse, si bien

[3]Dans la section 179, Carnap semble identifie l'analytique avec le conventionnel. « Pour la théorie de la constitution, il n'y a dans la connaissance que ces deux composantes, conventionnelle et empirique ; il n'y a donc pas de composante synthétique *a priori* ».

que l'analyse doit venir pour la rendre claire et distincte. Telle est l'explication initiale que Kant donne à la synthèse.

On peut dire avec justesse que la « synthèse » chez Kant signifie « le traitement du donné pour former et représenter les choses » (*Aufbau*, sec.100). Mais il faut être assez attentif pour ne pas identifier le traitement synthétique avec la procédure de classification, laquelle consiste à diviser des objets en groupes selon caractéristiques communs. Cette dernière procédure, bien qu'étant inséparable de la synthèse, appartient plutôt à l'analyse. Ce qui est essentiel à la synthèse de Kant, c'est qu'on pense l'unité de représentations non pas dans un système de concepts mais « *dans une intuition* » (A79/B105, souligné par Kant).

L'opposition de l'intuition et du concept constitue la présupposition ultime de la philosophie critique ; toute problématique épistémologique et sa solution en elle repose sur elle tant à l'égard de la formulation qu'à l'égard du contenu. Si nous désirons nous rendre compte de la nature de la synthèse et du rôle que joue celle-ci dans la problématique épistémologique, il nous faudra nous référer à cette dichotomie kantienne. Cela nous permettra de mettre en contaste le concept kantien et le concept carnapien de synthèse.

Quelle est la différence de l'intuition et du concept ? Comme on le sait, Kant explique la différence par celle de la connaissance directe et de la connaissance indirecte. Étant connaissance directe, une intuition représente un objet dans son individualité. Par exemple, on rencontre par l'intuition une pomme individuelle dont on peut apprécier la couleur, le toucher et la saveur. Par contre, un concept est une représentation générale qui se rapporte à plusieurs objets (ou plusieures représentations). On ne peut pas manger ni toucher de pomme au moyen d'un concept. En contraste avec la connaissance intuitive qui est de nature directe, la connaissance indirecte moyennant concepts est appelée parfois connaissance « discursive ».

La relation dans l'intuition est extensive ; plusieurs individuels y se rapportent de telle manière que le composé qu'il forment ensemble est lui-même quelque chose d'individuel. L'espace, qui est repésenté par nature d'après Kant, fait un exemple important. Un espace composé des espaces indivduels est lui-même un individuel. Ici, les individuels font un tout qui est de la même nature que la leur. On peut dire que la relation de composition dans l'espace est la relation d'un tout à ses parties ; en termes kantiens, l'espace contient ses parties « en lui ». Il en est de même pour un chien individuel qui se trouve dans l'espace. Il se rapporte à ses membres selon la relation d'un tout à ses parties. En revanche, quand on dit que plusieurs individuels font un concept, le concept constitué n'est pas quelque chose d'individuel. Il se peut que les

éléments sont eux-mêmes concepts. Mais à ce moment-là le concept qui les inclut sera plus general qu'eux. Si une intution comprend ses composantes « *en* lui », un concept inclut ses concept « *sous* lui » (A25/B40, cf.A78/B104).

La distinction analyse/synthèse peut se définir maintenant par référence à deux types de connaissance ou deux types de représentation ainsi mise en opposition. L'analyse et la synthèse ont ceci de commun qu'elles emploient des concepts pour connaître queque chose et envisagent d'établir l'unité de la conscience (ou des représentations). Elles diffèrent en ce qu'elles envisagent des représentations de types différents. C'est que l'analyse apporte l'unité pour les concepts tandis que la synthèse apporte l'unité pour les intuitions. L'unité du premier type est applée « l'unité analytique » et l'unité du dernier type « l'unité synthétique ». L'unité analytique est l'unité de conscience par laquelle nous relions diverses représentations l'une à l'autre selon un caractéristique commun ; elle nous permet de penser à plusieurs choses « sous » un concept général. Quant à l'unité synthétique, elle est l'unité de conscience par laquelle nous nous représentons un objet individuel qui se compose des parties « en lui ». Mais, à la différence de la connaissance intuitive prise isolément, les composantes de l'objet individuel est reliées dans l'unité synthétique à la connaissance discursive, c'est-à-dire, elles sont pensées au moyen des concepts. Ainsi, si l'on se représente une pomme comme ayant propriétés telle que la couleur rouge, la forme ronde, etc., on aura une unité synthétique. Si, par contre, on se représente la couleur rouge de manière général, c'est-à-dire, comme appartenant beaucoup d'autres objets particuliers outre cette pomme, on aura une unité analytique.

Sans craindre la simplification, nous allons dire que l'unité synthétique consiste toujours dans la coopération de la connaissance intuitive et de la connaissance discursive, et que l'unité analytique peut rester dans la connaissance discursive. Car Kant souligne souvent que l'unité synthétique doit précéder l'unité analytique quand il s'agit de connaître des objets empiriques.

> L'unité analytique de la conscience s'attache à tous les concepts communs comme tels : par exemple, si je pense le rouge en général, je me représente par là une qualité qui (comme caractéristique) peut être trouvée quelque part ou liée à d'autres représentations ; ce n'est donc qu'au moyen d'une unité synthétique possible, pensée auparavant, que je puis me représenter l'unité analytique. Une représentation qui doit être pensée comme commune à des *choses différentes*, est considérée comme appartenant à des choses qui, en dehors de cette représentation, ont encore en elles quelque chose

> de *différent* ; par conséquent, elle doit être pensée d'abord en une unité synthétique avec d'autres représentations (même si ce sont des représentations seulement possibles), avant que l'on puisse penser en elle l'unité analytique de la conscience, qui en fait un *conceptus communis*. (B133 note, souligné par Kant)

L'unité synthétique doit précéder l'unité analytique pour reconnaître des choses comme ayant propriétés qui sont à décomposer ultérieurement par abstraction et classification. L'abstraction et la classification, qui visent faire représenter des choses par un système des concepts généraux, ne peuvent venir qu'après ce traitement initial. Telle est l'idée qu'affirme Kant ici. Nous voulons la résumer en disant que la nature de la synthèse consiste dans la médiation qui doit avoir lieu entre l'intuitif et le conceptuel (ou le discursif) pour rendre le premier apte au dernier.

C'est par cette notion que Kant a voulu résoudre sa problématique épistémologique. La problématique épistémologique qui avait motivé la réflexion de ce philosophe est la question de savoir « sur quel fondement repose le rapport de ce qu'on nomme en nous représentation à l'objet » (Lettre à Marcus Herz, 1772). Dans la philosophie critique, la question devient celle concernant le rapport de l'intuition à la connaissance conceptuelle ; il s'agit d'expliquer comment il est possible qu'on crée, à partir des donnés sensibles, des concepts et des propositions qui sont susceptibles de l'usage scientifique[4]. Comme nous l'avons déjà dit, nous appelons tous ces problème « la problématique de la coordination ».

On se plaindrait sans doute que la « synthèse » kantienne n'est pas l'explication recherchée à propos de la médiation mais plutôt une simple dénomination en un autre terme, lequel a encore besoin de l'explication. À quoi la synthèse ressemble-t-elle, demandra-t-on. À ce sujet, nous présenterons notre commentaire plus tard. Avant de plaider pour la théorie kantienne, il nous faut

[4]Il se peut qu'on nous reproche de l'identification des deux problèmes différents : le problème sur le rapport de représentation à l'objet et le problème sur la formation des concepts scientifiques. En réalité, le premier problème peut se traduire par le dernier. Car un concept scientifique a ceci d'essentiel qu'il nous permet de corriger nos jugements éronnés sur ce que la perception nous présente. Par exemple, le concept (plus ou moins) scientifique de longueur nous permettra de dire que les deux flèches apparemment différentes en longueur, qu'on voit dans l'illusion de Müller-Lyre, sont *en réalité* de la même longueur. En général, pourvus des concepts (plus ou moins) scientifiques, nous pouvons juger si notre représentation intuitif correspond bien à son objet ou non. De ce fait, « la problématique de la coordination » peut concerner tantôt le rapport de représentation à l'objet, tantôt le rapport des concepts à l'intuition.

voir quelles métamorhoses son concept et sa problématique subissent dans l'*Aufbau*.

5. La prédominance du discursif – conception iconoclaste de la synthèse chez Carnap

Il existe dans l'*Aufbau* une idée analogue de la distinction kantienne intuitif/discursif. On peut la trouver dans la section 36, où la distinction de la « complexe » et de la « totalité » est introduite. Selon cette idée, lorsqu'un objet est réductible à d'autres objets, il peut être appelé un « complexe logique » ou simplement un « complexe ». Une classe constituée de ses éléments ou une relation constituée des paires de membre sont typiquement des complexes logiques. Par contre, si « un objet se rapporte à d'autres objets comme à ses parties relativement à un milieu extensive, espace et temps par exemple », il se nomme la « totalité extensive » ou simplement le « tout » des autres objets. Il n'est pas injuste, nous semble-t-il, de comparer cette distinction à celle kantienne de l'intuitif et du discursif. Ayant introduit cette distinction, Carnap déclare que ce que la constitution envisage ne sont pas les totalités mais les complexe. Il dit : « ... la théorie de la constitution a justement affaire avec les complexes qui ne se composent pas de leurs éléments comme une totalité de ses parties » (sec.36). Si l'on emploie l'expression figurée de Kant, on pourra dire que la constitution carnapienne veut nous faire connaître les choses par des représentations qui incluent leurs éléments « sous elles ». C'est que la « synthèse » dans l'*Aufbau* veut accomplir sa tâche au moyen de l'unité analytique. Détaché de l'unité synthétique, la synthèse semble être devenue opération purement discursive.

La métamorphose de la « synthèse » tient au fait que la théorie de la constitution soit conçue comme application de la logique mathématique à la théorie de l'objet (sec.2). Tous les objets y sont considérés comme classes ou relations, constituées de leurs éléments ou de leurs membres. La connaissance par intuition n'entre plus en jeu, dans la reconstruction rationnelle de la connaissance, pour se représenter des objets.

Cela ne veut pas dire, toutefois, que la théorie de la constitution traite seulement de ce qu'on appelle habituellement les concepts généraux et qu'elle laisse de côté les concepts individuels. Traditionnellement, on distingue ces deux types de concepts ou, dans le même esprit, le concept et l'objet[5]. Le système de constitution dévalorise cette distinction elle-même. Étant système

[5]Kant préférait cette dernière formulation de la distinction. Diviser les concepts en concepts

compréhensif de concepts, il traite tous les concepts (ou tous les objets), qu'ils soient « généraux » ou « individuels », également comme complexes logiques. Une illustration donnée par Carnap peut nous aider pour saisir cette idée (sec.158).

Supposons qu'on a un chien nommé Luchs. Le chien en tant qu'espèce est une classe à laquelle appartient ce chien Luchs. Par conctre, du point de vue de la manière habituelle de voir, « Luchs » est un concept individuel et correspond à un objet individuel ou particulier. Mais, selon le système de constitution, il est possible de dire légitimement que « Luchs » est une classe ; ses éléments sont alors les états de Luchs. Un état particulier de Luchs (en tant qu'objet de la perception), à son tour, est une classe dont les éléments sont des points du monde (*Weltpunke*) de la perception. Un tel point est une relation dont les membres sont les coordonnées spatio-temporelles et une ou plusieurs qualités sensibles. Une qualité sensible est une classe de « mes vécus ». De cette manière, on peut considérer toujours un objet apparemment individuel comme classe ou relation, à savoir comme complexe. Seuls les éléments fondamentaux (ce sont ici « mes vécus »), refuse la qualification de « complexe logique ». Dans cette série hiérarchique des objets ayant différents niveaux, la différence qu'on voit d'habitude entre l'individuel (le chien Luchs) et le général (l'espèce chien ou la qualité sensible brun) est relativisée[6] ; chaque terme apparaîssant dans la série peut être qualifié tantôt de général tantôt d'individuel selon le niveau qu'on choisit comme point de vue. Du fait de cette relativisation, tous le concepts, y compris ceux des objets individuels existant dans l'espace et le temps, ressortissent à présent à la connaissance discursive. L'exclusion de l'unité synthétique ne signifie donc pas l'exclusion des objets dits « individuels ».

Naturellement, la relativisation de la généralité signifie aussi l'élargissement de la notion d'individualité. De ce point de vue, tous les concepts, même les concetps dits généraux, pourront être considérés comme individuels. Carnap formule cette l'idée par celle de la pluralité des milieux dans lesquels les objets de divers niveaux d'abstraction trouvent respectivement leurs principes

---

généraux et en concepts individuels était populaire dans les manuels de logique du XVIII^{e} siècle (par exemple, Georg Friedrich Meier, *Auszug aus der Vernunftlehre*, 1752, sec.260, 262). Kant aussi parle de cette division dans sa lecture de la logique. Mais il n'aimait pas de dire « concepts individuels » parce que, selon la philosophie critique, seule l'intution peut représenter un individuel et les concepts sont tous généraux. Ce qui peut être qualifié d'individuel n'est pas un concept mais un usage de concept, dit-il (Kant, *Logique Jäsche*, sec.1).

[6] « Contrairement à la doctrine traditionnelle du concept, la généralité d'un concept nous semble relative et par suite, la limite entre concepts généraux et individuels varie suivant le point de vue (voir sec.158) » (*Aufbau*, sec.5).

d'individuation. Un milieu de cette sorte, incarnant l'arrangement struturel des objets d'un certain type, est appelé un « ordre ». Tout objet, qu'il soit de nature spatio-temporel ou non, peut maintenant être conçu comme appartenant à un ordre quelconque dans lequel il prend place parmi d'autres objets du même type. Ainsi, de même que le chien Luchs est qualifié d'individuel relativement à l'ordre spatio-temporel, la qualité sensible brun a son lieu dans un autre ordre, dans lequel toutes les couleurs se distinguent l'une de l'autre et s'ordonnent selon le degré. Cet ordre, l'« ordre des couleurs », a trois dimensions servant de points de vue pour la mise en ordre : le ton, la saturation et la luminosité.

On pourrait dire que les « ordres » sont les représentations schèmatiques des relations conceptuels, lesquels sont par ailleurs de nature discursif. Il arrive même qu'on apporte un ordre dans une représentation intuitive ou imagée. Par exemple, l'ordre des couleurs peut être représentée par l'image d'un corps ayant trois dimensions. Celui-ci s'appelle d'habitude « le corps des couleurs » (sec.81, 90). Le corps des couleurs n'est pas un corps spatial au sens littéraire, mais est néanmoins susceptible d'être exposé intuitivement. De même que le chien Luchs occupe un certain domaine des points spatio-temporels, on peut dire maintenant que le brun occupe un point ou un domaine de corps des couleurs ; s'il s'agit d'un ton tout à fait déterminé, un point de ce corps lui correspond et, s'il s'agit du brun en général, un sous-domaine connexe du corps lui correspond (sec.158). Ce n'est pas seulement les concepts de couleurs qu'on peut mettre en image. Carnap suggère la possibilité de concevoir un corps imaginaire dans lequel les espèces animaux s'ordonnent et se distinguent : « le corps zoologique » (ibid.). Naturellement, on ne peut pas dire que tous les ordres sont susceptibles d'être représentés intuitivement. Si le nombre des dimensions déterminant un ordre dépasse trois, il sera impossible de l'exprimer par un image.

Quoi qu'il en soit, tout en quittant le dualisme de l'intuitif et du discursif, la constitution de Carnap s'assure toujours le parallélisme du concept et de l'objet. Ce parallélisme se reproduit à chaque niveau de constitution en se complétant par un ordre quelconque. La synthèse carnapienne semble n'avoir plus besoin de procédé médiateur que Kant supposait pour la coopération des deux connaissance hétérogènes. Tout semble se dérouler maintenant à l'intérieur de la connaissance discursive. Si la synthèse doit signifier l'opération qui médiatise l'intuition spatio-temporelle et les concepts généraux, alors il faudra dire que la synthèse carnapienne n'est pas synthèse mais plutôt analyse. La théorie de la constitution exclut non seulement le synthétique *a priori* mais aussi la synthèse elle-même.

Mais est-il possible d'affronter la problématique épistémologique, à laquelle Kant voulait répondre par sa théorie de la synthèse, sans recourrant à la synthèse? Il n'est pas illégitime de poser cette question parce que Carnap lui-même a l'intention d'en traiter. Il croit possible de donner la solution à la problématique de la coordination par sa méthode dont l'unique arme est la logistique.

6. La caractérisation structurelle – la problématique de la coordination chez Carnap

Ce que nous appelons la problématique de la coordination se revêtit d'une nouvelle formulation dans l'*Aufbau* et reçoit une solution qui différe radicalement de celle offerte par la *Critique*. Cela n'est pas étonnant, car la métamorphose de la « synthèse » (ou plus précisément, le licenciement de la « synthèse ») ne peut avoir lieu que si la problématique lui-même se métamorphose. La métamorphose de la problématique chez Carnap est motivée sans aucun doute par la puissance remarquable que la nouvelle logique lui paraissait prométtre concernant la question de l'organisation des concepts. On sait maintenant organiser les concepts dans les « ordres » qui leur sont propres et à les spécifier par leurs relations mutuelles ; cela signifie qu'on a un moyen fort sophistiqué pour décrire des objets. Cette situation a permis Carnap de délimiter sa problématique épistémologique à la question sur la possibilité de la désignation univoque des objets par un système de symboles. Nul doute que l'*Aufbau* présente une idée très prometteuse dans le cadre de cette question.

C'est surtout la théorie mathématique des relations, introduite par Russell et Whitehead dans la logistique, qui a inspiré Carnap de la solution. Elle a rendu possible de développer et maîtriser l'arme indispensable de la théorie de la constitution, à savoir le concept de structure. Voyons la nature de cette arme. D'abord, Carnap distingue deux façon de décrire des objets : la « description de relation » et la « description de propriété » (sec.10). La description de propriété indique quelles propriétés appartiennent aux objets parculiers d'un domaine quelconque. La description de relation indique, par contre, quelles relations existent entre les objets sans rien dire des objets en eux-mêmes. Par exemple, supposons qu'il y trois hommes a, b, c. La proposition : « a est agé de vingt ans et grand » offre une description de propriété. En revanche, les propositions comme « a est le fis de b » et « b a trente ans de plus que c » offrent des descriptions de relation. Dans ces dernières, on

n'énonce plus quelque chose d'intrinsèque des objets particuliers. Si l'on poursuit l'abstraction encore plus loin et décide de n'énoncer plus les noms des relations d'objets, on aura des « descriptions de structure » ou des « caractérisations structurelles (*strukturelle Kennzeichnungen*) » (sec.11).

La méthode de caractérisation structurelle est ceci d'avantageux qu'elle nous permet de désigner des objets systématiquement et d'en traiter ainsi dans le discours intersubjectif ou scientifique. Elle peut résoudre la question de la désignation univoque là où les objets en question sont « discernables par des moyens scientifiques » (sec.15). De plus, elle nous permet non seulement de désigner des classes d'objets mais aussi des objets individuels. Un bon exemple est une carte des voies ferrées qui représente un réseau férroviaire sans conserver la similitude avec la forme dans la réalité. Sous conditions favorables, seules les relations que montrent les lignes et les points sur la carte suffiront pour dire de quel région et de quelles stations il s'agit. L'idée de discuter de la connaissance objective en terme de la désignation univoque avait déjà été avancée dans le traité de Schlick sur l'épistémologie. Mais, à la différence de la méthode de la définition implicite employée par Schlick, la méthode de Carnap ne s'adresse pas seulement aux concepts dits « généraux » mais aussi aux objets « individuels ». C'est pourquoi, comme l'indique Richardson, Carnap peut être fier de l'avantage qu'a sa théorie sur celle de Schlick ; la théorie de l'*Aufbau* a la possibilité d'assurer aux concepts, logiquement construits, le contacte avec la réalité empirique, lequel restait problématique dans la *Théorie générale de la connaissance*[7].

La méthode de caractérisation structurelle semble si prometteuse que Carnap n'hésite pas à exprimer sa réponse à la problématique de la coordination sous une formulation fort paradoxale. Selon cette formulation appartenant originairement à Reichenbach[8], il s'agit, dans la connaissance de la réalité empirique, de réaliser une relation de correspodance (ou une coordination) uni-

[7] «Un système des vérités édifiées au moyen de la définition implicite ne repose nulle part sur le sol de la réalité, mais se meut pour ainsi dire librement, portant en lui-même – tel le système solaire – la garantie de sa propre stabilité» (Schlick,1918/1925, sec.7 ; p55, fr.p.84). Cf. *Aufbau*, sec.15 ; Richardson 1998, p.42

[8] Reichenbach avance cette idée en la mettant en contraste avec l'idée ordinaire de la correspondance telle que la correspondance mathématique entre deux ensembles. Dans ce dernier cas, on a d'abord les éléments déterminés pour chaque ensemble et ensuite définit une correspodance entre les deux côtés. Mais, dans la connaissace de la réalité, le côté de la réalité n'a pas, au début, d'éléments déterminés, car ce sont les déterminations de ces éléments qu'on a à établir par la connaissance, dit-il. « Ainsi, nous sommes en face de ce fait singulier, que nous établissons dans la connaissance une relation de correspondance (*Zuordnung*, *coördination*) entre deux ensembles, dont l'un [l'ensemble des objets empiriques] obtient non seulement son ordre mais aussi les définitions de ses éléments en vertu de cette correspondance elle-même » (Reichenbach, 1920/1965,

voque entre les signes et les objets et cela de telle manière que les objets ne deviennent les termes de la correspondance qu'en vertu de cette correspondance elle-même. Carnap dit dans l'*Aufbau* :

> Par la méthode de caractérisation structurelle, il devient à présent possible de faire correspondre de manière univoque des signes aux objets empiriques et de les rendre ainsi accessibles au travail conceptuel, quoique par ailleurs les objets empiriques ne peuvent de toute façon être déterminés individuellement que par cette symbolisation. (Carnap, 1928, sec.15)

Ce passage nous montre combien la théorie des relations et le concept de structure ont encouragé Carnap à affronter « *la question de la théorie de la connaissance* (*die Frage der Erkenntnislehre*) » en excluant non seulement le synthétique *a priori* mais aussi le synthétique lui-même. Si Kant a dû s'attacher à la description de propriété et chercher le moyen rationel qui ferait l'intermédiaire des caractéristiques générales et des objets individuels, Carnap sait relier les concepts mutuellement dans une caractérisation structurelle et rendre le royaume de concepts lui-même un analogue rationnel des objets empiriques « individuels ».

Mais la solution par la caractérisation structurelle peut-elle être vraiment une réponse à la problématique épistémologique de la coordination ? Le synthétique s'est-il avéré inutile devant la nouvelle arme qu'a obtenu la constitution ?

7. De la fiction au pratique

À propos de cette dernière question, le kantien peut repousser la conclusion affirmative. Deux points de vue, qui s'entrecroiseront sans doute au bout de l'examen, peuvent être avancés. En premier lieu, l'idée de la solution par caractérisation structurelle fait ressortir la différence des approches que prennent les deux philosophes vers la problématique, plutôt que l'inefficacité de la vieille solution. En second lieu, il n'est pas évident que la théorie de la constitution ait résolu la problématique de la coordination. Nous allons développer principalement le premier point de vue pour finir cette étude comparative.

Le kantien peut légitimement mettre en question, à notre avis, la délimitation de la problématique en celle des moyens de la description. Certes, la

p.40).

théorie de la constitution maîtrise d'un langage efficace et important à l'égard de la description d'objets qui est apte au discours scientifique. Mais elle ne dit pas comment on distingue les objets en sorte que ceux-ci se manifestent aptes à la caractérisation structurelle. La possibilité de cette dernière dépend sans aucun doute de quelques moyens rationnels ou scientifiques employés pour discerner les différences des objets. De fait, Carnap est bien conscient de cette condition.

> *Il ressort que la caractérisation univoque au moyen de pures indications de structure est en général possible, pour autant qu'est possible en général une différenciation scientifique :* cette caractérisation ne fait alors défaut pour deux objets que s'ils ne sont pas du tout discernables par des moyens scientifiques. (sec.15)

Le passage fait remarquer avec justesse la nécessité des moyens pratiques qu'on applique à des objets pour en reconnaître des différences et des relations. En effet, si l'on peut se rendre compte qu'une carte représente un réseau férroviaire, n'est-ce pas parce qu'on sait *compter* le nombre des stations d'un région réel ou le nombre des lignes qui sortent de ces stations ? Si les scientifiques peuvent légitimement accorder le terme H_2O à l'eau, n'est-ce pas parce qu'ils savant *décomposer effectivement* de l'eau en deux substances ? Naturellement, reconnaître les différences des objets ressort, au final, à la fonction de la perception qui a lieu par nature indépendamment du contrôle intentionnel. Il n'en est pas moins nécessaire de discuter de la nature rationnelle des moyens pratiques, qui sont pris d'habitude consciemment, pour utiliser la perception en faveur de la description scientifique. Or la théorie de la constitution ne semble pas prendre en considération la question de ces moyens pratiques lorsqu'elle détermine les formes des niveaux. En séparant ainsi la question de la description de celle des moyens pratiques de recherche, la théorie carnapienne a pour effet d'écarter une voie, qui appartenait à l'approche kantienne et qui est prometteuse à notre avis, pour faire coordonner l'empirique et le formel.

Pour justifier notre diagnostic, nous allons comparer la conception canapienne et la conception kantienne sur les « formes de la synthèse ». Nous verrons que la conception kantienne repose sur l'idée qu'il y a une interdépendance essentielle entre les moyens pratiques de recherche et les moyens de description. Grâce à cette présupposition, dont la légitimité est indubitable, l'approche kantienne de la problématique de coordination n'a pas besoin de formulation paradoxale telle que celle aimée par les empiristes logiques. Pre-

nons pour exemple la connaissance des relations de couleurs dans le but de mettre en contraste les conceptions des deux philosophes.

Dans l'*Aufbau*, la constitution des qualités sensibles telles que les couleurs vient en premier et précède toutes les autres étapes de constitution. Comme nous l'avons vu, cette constitution initiale se déroule par la procédure de « quasi-analyse » et la seule relation fondamentale disponible à celle-ci est le rappel de ressemblance. Or la supposition que la comparaison s'appuie uniquement sur la mémoire et que les donnés de la comparaison vont se classifier en sorte qu'ils se conforment à l'ordre des couleurs, nécessite que le système convoque ceraines fictions extrêmement imaginaires pour se montrer une théorie de la synthèse. Deux « fictions » introduites par Carnap importent ici : la fiction de « la séparation du donné » et celle de « la rétention du donné » (sec.102). Selon la première fiction, la réception des donnés élémentaires et le traitement cognitif de ces donnés sont séparés et distribués à deux périodes différentes ; il est à supposer qu'une personne reçoit des donnés dans la première partie de sa vie sans effectuer aucun traitement, et qu'elle s'occupe du traitement des donnés dans la seconde partie de sa vie sans recevoir de nouveau donné. Selon la deuxième fiction, la fiction de la rétention du donné, tout donné reçu dans la première période est censé être retenu dans la mémoire ; en conséquence, la personne pourrait opérer le traitement cognitif sur un ensemble suffisamment riche des donnés dans la deuxième période.

C'est sous ces conditions imaginaires que l'esprit assume la tâche de la quasi-analyse. Il doit parcourir les souvenirs dans son stock pour classifier les vécus élémentaires selon diverses ressemblances partielles. Au cours de ce processus, les vécus ayant certaines couleurs similaires constituent un groupe, dans lequel ils s'ordonnent selon le degré de ressemblance (« les cercles de ressemblance », sec.80). Depuis les groupes ainsi formés, la quasi-analyse construit des classes dont chacune incluent les vécus comportant en commun une certaine couleur déterminée (« les classes de qualité », sec.81). On obtient ainsi les concepts déterminés des couleurs.

Manifestement, tout ce processus présuppose la sûreté et la capacité énorme de la mémoire, ce qui n'est pas le cas en réalité. On pourrait dire que le premier travail de l'abstraction, que l'esprit effectue inconsciemment dans sa vie réelle, ressemble à la quasi-analyse dans une certaine mesure. Mais le travail élémentaire de l'esprit ne sait pas encore trouver des relations aussi rigides et systématiques que celles maîtrisées par la quasi-analyse. Notamment, la capacité de notre mémoire ne nous permettra qu'une performance très pauvre à l'égard du jugement sur le degré de similarité.

Nous savons bien que l'objectif de l'*Aufbau* ne consiste pas à décrire le

processus effectif de la cognition mais à reconstruire de toute sorte d'objets par un langage formel. Il n'en est pas moins important de noter l'indifférence que le système carnapien montre pour la finitude de notre mémoire. Car c'est cette indifférence qui caractérise la conception carnapienne des « formes » de la synthèse. On peut la reconnaître surtout par le fait que l'opposition de l'ordre spatio-temporel et d'autres ordres se voie minimisée dans ce système (sec.158).

Quant à la théorie kantienne de la synthèse, elle prend pour essentielles les conditions que les fictions de l'*Aufbau* considèrent comme négligeables. C'est-à-dire qu'elle commence par retenir la limite et la fragilité de notre mémoire et recherche des règles qui permettraient la comparaison systématique *en dépit de* ces défauts. Ce sont les schèmes transcendantaux qui nous indiquent les règles de cette sorte. En effet, ils ont pour fonction de faire l'intermédiaire de l'intuitif et du discursif ou du temporel et de l'intemporel. Selon la philosophie du schématisme, la médiation ne se réalise que quand on intervient intentionnellement dans des empiriques pour y découvrir et exploiter des phénomènes homogènes ou répétables. Compter, produire artificiellement, distinguer ce qui est constant d'avec ce qui ne l'est pas, etc. sont des actes d'intervention qui sont en question. Si la comparaison des donnés se déroule avec une sûreté considérable dans la vie réelle, n'est-ce pas parce qu'on vertu de ces actes stratégiques ? Habituellement, on fait la comparaison non seulement en parcourant dans le stock des souvenirs, mais aussi en cherchant intentionnellement de nouveaux données qui seraient pertinentes. Par exemple, s'il est demandé de mettre en ordre des exemples de couleur selon le degré de ressemblance, on transposera des objets colorés pour les regarder sur la table dans divers arrangements. Ce moyen nous permettra de former un concept (plus ou moins) scientifique et de corriger même des érreurs de la mémoire. Ce qui récomponse la finitude de la mémoire ici, c'est la liberté de la composition artificielle, laquelle s'exerce (non pas sur des souvenirs mais) sur des choses elles-mêmes. Or ce n'est autre chose, à notre avis, que ce que le schème transcendantal de la qualité nous indique. Il enseigne à « produire continuellement un réel » d'un degré à un autre. On peut penser aussi, comme application de ce schème, à la création des dégradés de couleur exercée par les peintres ou par les ingénieurs dans le but d'apprendre à maîtriser les couleurs.

La réflexion ci-dessus nous apprend que le concept de degré chez Kant exprime non seulement une relation dont on a à profiter pour classifier des objets, mais aussi le moyen pratique qu'on peut employer effectivement pour découvrir la relation dans des donnés de l'intuition. En partant ainsi de l'idée

de l'interdépendance essentielle entre la pratique et la description en science[9], autrement dit en partant de la finitude de nos capacités cognitives, la philosophie critique s'assure un fondement qui lui permet de faire raison de sa théorie des formes de synthèse. Dans la théorie de la constitution, au contraire, le choix des relations fondamentales ne doit être sujet qu'à l'idée d'un système de la réduction. C'est qu'elle se décide à ne faire interposer aucun concept de rationalité, sauf logique formelle, dans le dialogue des formes avec les matériaux. Mais cette décision dispense-t-elle Carnap de se préoccuper par la question de la médiation entre l'empirique et le logique?

La comparaison de Carnap avec Kant nous amène ainsi à la question de savoir si l'*Aufbau* a résolu la problématique de la coordination. Nombre de commentateurs ont indiqué l'existence des soucis qui ont plus ou moins un rapport à ce sujet. Il s'agit, dans la plupart des cas, de la suspicion que, pour construire formellement les objets tels qu'ils nous sont connus comme objectifs, la théorie de la constitution serait obligée d'introduire certains principes non logiques sans justification interne[10]. Certains de ces soucis pourraient s'avérer apparents, comme le suggèrent d'autres commentateurs (Proust 1986, p.307 ; Richardson 1998, p.63), si l'on s'empêche de supposer une réalité indépendante qui précéderait la constitution. Mais la question fondamentale subsiste. En se déclarant une reconstruction rationnelle du processus cognitif et en s'appropriant du moyen langagier permettant la transcription du réel en symbolique, le système de constitution semble être obligé de nouveau d'affronter la problématique de la coordination à son intérieur. La coordination des concepts scientifiques et de la réalité empirique y sera reprise, par exemple, dans le cadre de la question concernant la coordination correcte qu'il faut établir entre la constitution du monde perceptible et la constitution du monde physique. Carnap répète que le monde de la perception, construit à partir des vécus élémentaires, a besoin d'être corrigé et complété à la lumière des concepts des objets physiques et des lois physiques (*Aufbau*, sec.136). Il nous faudra demander, avec Friedman, comment il est possible que les concepts physiques de l'objet peuvent nous servir à corriger et com-

[9] Tel est, à notre avis, ce que Kant voulait signifier par l'interdépendance de l'unité analytique et de l'unité synthétique. « La même fonction qui donne l'unité aux représentations diverses *dans un jugement*, donne aussi à la simple synthèse de représentations diverses *dans une intuition* l'unité, qui, exprimée généralement, s'appelle le concept pur de l'entendement ». (A79/B105, souligné par Kant)

[10] Sur les fameux « difficultés » concernant la quasi-analyse, voir Goodman 1951 (p.121, fr.pp.155-156) et Vuillemin 1971 (p.276). D'un point de vue plus général, Granger déclare aussi l'échec de l'*Aufbau* (Granger 1994, p.320-321).

pléter le monde de perception alors qu'il soit demandé de réduire les concepts physiques eux-mêmes à des donnés élémentaires (Friedman 1998, p.122). La discussion portera alors sur la précision et l'examen de la solution que Carnap peut avancer à l'égard de cette question : le conventionalisme (Friedman 1998, loc.cit. ; Proust 1986, p.319, pp.326-327). À notre avis, le kantisme se manifestera capable de se confronter avec le conventionalisme dans la mesure où la conception kantienne de la synthèse telle que nous venons de mettre en évidence se voit approfondie. Mais nous n'entendons pas poursuivre notre discussion à ce sujet ici.

Évidemment, dire que le kantisme a plus d'avantage qu'il n'y paraît ne signifie pas nécessairement qu'il a montré l'existence du synthétique *a priori*. Mais, en remarquant ce qui échappe à l'idée carnapienne de la théorie de la connaissance, nous trouverons la direction vers laquelle le kantien aujourd'hui peut porter ses pas pour avoir une notion signifiante de synthétique *a priori*. Sans faire dévaloriser l'importance de la question sur le langage logique de description, il nous faudra retourner au processus effectif de la connaissance pour tenter de savoir s'il y a quelques principes généraux qui seraient constitutive des moyens pratiques de la recherche empirique, propres à un entendement fini comme le nôtre.

8. Références

Carnap, Rudolf, (1928), *Der logische Aufbau der Welt*, Felix Meiner Verlag, Hamburg, 1998. *La construction logique du monde*, traduction de Thierry Rivain, revue par Élisabeth Schwartz, Paris, J.Vrin, 2002.

Carnap, Rudolf, (1934/1937), *Logische Syntax der Sprache*, Wien, J.Springer, 1934. *Logical Syntax of Language*, translated by Amethe Smeaton, 1937, reprinted 2001, London, Routledge.

Friedman, Michael, (1999), *Reconsidering Logical Positivism*, Cambridge University Press.

Goodman, Nelson, (1951), *Structure of Appearance*, 3rd ed., Dordrecht, Boston, D.Reidel, 1977. *structure de l'apparence*, Traduction française par Pierre Livet, Philippe Minh, Nancy Mulzili, Marc Pavlopoulos, Jean-Baptiste Rauzy, Norma Yunez-Naude, Paris, J.Vrin, 2004.

Granger, Gilles-Gaston, (1994), *Formes, opérations, objets*, Paris, Vrin.

Kant, Immanuel, (1781/1786), *Kritik der reinen Vernunft*. *Critique de la raison pure*, traduit par Alexandre J.-L. Delamarre, et Francois Marty à partir de la traduction de Jules Barni, Paris, Éditions Gallimard, 1980.

Proust, Joëlle, (1986), *Questions de forme : Logique et proposition analytique de Kant à Carnap*, Librairie Arthème Fayard, Paris, 1986.

Reichenbach, Hans, (1920/1965), *Relativitätstheorie und Erkenntnis a priori*, Berlin, Springer, 1920. *The Theory of Relativity and A Priori Knowledge*, translated and edited by Maria Reichenbach, Berkeley and Los Angels, University of California Press, 1965.

Richardson, Alan W., (1998), *Carnap's Construction of the World : The Aufbau and the Emergence of Logical Empiricism*, New York, Cambridge University Press.

Schlick, Moritz, (1918/1925), *Allgemeine Erkenntnislehre*, 2.Auf., repri.in Suhrkampf, Frankfurt am Main, 1979. *La théorie générale de la connaissance*, traduit en français par Christian Bonnet, Paris, Gallimard, 2009.

Vuillemin, Jules, (1971), *La logique et le monde sensible. Études sur les théories contemporaines de l'abstraction*, Paris, Flammarion.

17

Modes of knowledge and vagueness

PIERRE LIVET

1. Introduction

In *Va Savoir!* (Hermann, 2007), Pascal Engel claims that we can know a proposition without necessarily knowing that we know this proposition. This implies that we can know something without being able to give strong inferential and reflexive justifications of our knowledge. In this conception, knowledge is based upon external foundations and not only upon internal reasons. Nevertheless, this externalist conception can give place to justifications, because external justifications are possible, mainly the prima facie justifications given by perception and its non-conceptual contents. This kind of modest dogmatism about knowledge allows Engel to share with Williamson not only the conclusion of his argument on vagueness, the thesis that knowing p does not ensure knowing that one knows p, but also the idea that knowledge cannot be reduced to a kind of belief and that our concept of belief depends on our concept of knowledge.

I agree with all these propositions, except the last one that I find disputable, because it seems difficult to give primacy either to the concept of belief or to the one of knowledge. I will address this question only at the very end of this paper. I will first concentrate on the articulation between the perceptual and the inferential foundations of knowledge and their relation to the problem of vagueness. I will begin by some considerations upon Williamson's argument. Then I will propose a formulation of the problem of vagueness that makes us able to treat the problem of generalized vagueness (vagueness of any procedure used to solve a vagueness problem). This leads to examine more carefully the relation between the perceptive discrimination of a form

or a quality and perceptual comparison. This relation will give us a basis for anchoring a conceptual content (on a perceptive identification and linking the difference between the two content to a move from the perceptual discrimination of a form, a quality or an object towards an inquiry on the reliability of our epistemic access to their identification, a move that I will call an "epistemic ascent". My conclusion will be that our cognition can only reach epistemic states *compatible* with knowledge – and, when we are able to build new methods of inquiry, with knowledge of higher order. Belief, in this perspective, is compatible with a single step of this process, while knowledge (as in Peirce's conception) is compatible with new steps of inquiry.

2. Williamson's argument

Let us briefly recall how Williamson's argument works. In order to be reliable, knowledge requires that the cases in which we are in position to know that p cannot be too close to cases in which p is false. This implies that between cases in which we know that p is true and cases in which we know that p is false, there is an area in which we do not know whether p is true without knowing that p is false, because, would have one a direct access to facts, p is still true in this area[1]. This buffer zone ensures us the safety of our knowledge. In cases very close to the considered case in which p is true, we would still tell without error that p is true. We have a margin of safety, but such margin implies a zone of vagueness.

Now consider M. Magoo. He knows that his visual powers of discrimination are bad. Suppose for example that, if a tree is x cm tall, M. Magoo does not know whether it is x cm or $x + 1$ cm or $x - 1$ cm tall. In this case his margin of error is at least 2 cm wide. When the tree is in reality $x + 1$ cm tall, M. Magoo knows that he cannot exclude, just by using his bad visual powers, that the tree is x cm, because x cm is still inside his margin of error. We can claim: (1) "M. Magoo knows that, if the tree is $x + 1$ cm tall, he does *not* know that the tree is *not* x cm tall". It is a general proposition, relating a hypothesis ("if the tree is $x + 1$ cm tall") and a negative epistemic consequence. It can be true even if "the tree is $x + 1$ cm tall" is not the case.

Suppose that (2) "Magoo knows that the tree is *not* x cm tall". Let us assume (*KK*): "everybody who knows p knows that he knows p" (Kp implies

[1] If it was not the case, a case of the first kind (p known true) could be adjacent to a case of the second one (p known false).

KKp). By KK we go from (2) to (3): "he knows that he knows that the tree is not x cm tall".

Notice that if Magoo's margin of error is 2 cm, (2) would imply that the height of the tree is equal to or more than $x+2$ cm if we go up, or than $x-2$ cm if we go down. In either case, the tree cannot be $x+1$ cm tall. We would know by (2) and the margin of error that proposition q ="the tree is $x+1$ cm tall" is false[2].

(2) implies also that "Magoo does *not* know that the tree is *not x cm* tall" is false. Can be (1) still valid? Yes. The antecedent and the consequent of its implication are assumed both to be false, and the only case in which an implication is false is when the antecedent is true and the consequent false. Therefore by (1), q implies not (2). But by the validity of (3), we obtain again (2).

From "q implies not (2)" and (2), we infer by contraposition that not q. As Magoo is supposed to know a consequence of the set of propositions that he knows, and he knows the content of (1), KK, (2) and (3), he knows not q: "Magoo knows that the tree is *not* $x+1$ *cm* tall".

This one more step than knowing that the tree is *not x cm* tall, a step in the direction of the tree being taller than x cm and $x+1$ cm. The same reasoning can be repeated, leading to the conclusion of an immense tree, a tree that even the myopic Magoo can distinguish from the tree that he is seeing at the beginning of this reasoning. Williamson (p. 115-116, Oxford U. Press, 2000) concludes that the only sensible thing to do in order to avoid the disastrous conclusion of this sorites is to reject KK[3]. We can know something without knowing that we know it.

I agree with the conclusion, but at first sight something in step (2) looks strange with respect to Mister Magoo's epistemic abilities. On one hand he is able to know by (1) that something is under the threshold of his discriminative power: it is a *general* property of his visual abilities that he cannot distinguish x cm and $x+1$ cm. On the other hand, there are *particular* cases in which he is supposed by (2) to be able to know that *not x c*m is true, while not knowing that $x+1$ *cm* is true. Why is M. Magoo unable to use his knowledge of the existence of a margin error and the contrast between knowing and not knowing

[2]This conclusion depends on knowing what margin of error is the one of M. Magoo.

[3]In this case we cannot obtain again (2), and cannot conclude not q (not $x+1$ cm tall) by contraposition.

in order to conclude that $x + 1$ cm is inside his margin of error, and is in this respect a plausible measure?

I think that the oddity here is only apparent. But some elaboration is needed to clear it away.

3. Breaking the symmetry of uncertainty

The difference between x cm and $x + 1$ cm is a particular case of a general problem of categorization: has item i to be put in category A or in category B (supposedly disjoint)? Our ancestors, the gatherer-hunters, have to solve this kind of problem. Is this forked form in the bush a sign of the horns of an antelope (category h) or the fork of a stump, to be put in the not h category? There is an epistemic state in which the hunter is uncertain: neither he knows h, nor he knows not h. In this state, there is an epistemic symmetry between the two possible propositions h and not h, and the uncertainty state can be written: (not Kh and not K not h). When the form seems to move, the hunter gets a clue that breaks the symmetry in favour of h — if he has no other clue in favour of not h. Now, his epistemic state includes that he knows that he does not know not h: (K not K not h). Does he know that he knows h? He has a sign in favour of h, but he has no proof that this sign is decisive, because he has noticed in the past that he could believe to see a move of an object, while in fact this impression was due to a move of his head or to one of his visual saccades. Therefore he does not know that he knows h : not KK h. But (not KKh) is still compatible with Kh. By contrast, (K not K not h) is not compatible with K not h. The symmetry that characterizes the two parts of the epistemic state of uncertainty is broken.

Breaking symmetry opens the possibility of another epistemic move, an inquiry about a second order knowledge. Does the hunter know that he knows h? Remember that even a hunter may have to answer to this apparently sophisticated question: if he tries to run and hunt down every thing that seems to move and discovers that in a lot of the cases it was a misperception, he will waste a lot time and effort in unsuccessful attempts. But notice that by asking this new question, he shifts from a question about in which category to put the form or the thing towards a question about the reliability of his perception and interpretation of the move. Was it a real move or an apparent move due to a move of his head or to his visual saccades? The question is now about his epistemic access and not about the category of the thing the form of which he was seeing. We can call this shift an "epistemic ascent". Notice that it

becomes reasonable to make this epistemic ascent in an inquiry about of the second order knowledge of h only when there is a dissymmetry in favour of h, that is, when the epistemic state of the hunter becomes compatible with knowing h.

Suppose that in his second order inquiry, the hunter finds new clues that break the symmetry between the conclusion: "the previous epistemic access is reliable" and "it is not reliable" access. These new clues do not yet imply that $KKKh$, because our hunter has not yet tested his methods to assess reliability. We are still in the state not $KKKh$. But since these new clues are not compatible with (K not KKh) they are compatible with KKh.

Our hunter has no need to go further, but scientists may have: they may not only require, for example, that a proposition is demonstrated, but that the theoretical framework in which it is demonstrated is the right one for the considered topic. For example, Euclidian geometry is not the right framework for the theory of relativity.

We can generalize: at each step of the epistemic ascent, we are in an epistemic state that is guaranteed to be compatible with an order of knowledge that is lower than the level of the present step in the ascent. This compatibility state is not just a knowledge "by default" (knowledge in the absence of a demonstrated contradiction). Knowledge by default requires only that the epistemic state is compatible with Kh. In our third step of ascent, for example, the epistemic state is not only compatible with Kh, but also with KKh. This is more robust than the basic knowledge by default.

After some of these steps, our epistemic situation is the following: it is neither compatible with (K not h), nor with any degree of ($K.(n).K$ not h). Therefore, it implies (not $K.(n).K$ not h). It is compatible with Kh and with some degree of $K.(n).Kh$, but a higher degree $K.(n+1).Kh$ is not guaranteed. Are the lower degrees of $K.(n-m).Kh$ guaranteed? We can notice that in counterexamples like the one of the Euclidian geometry which is not relevant for relativity, we have a proof of the discrepancy between the theory of relativity and the Euclidian geometry: we know that we do not know the validity of the postulate of parallels in the theory of relativity. As we climb higher in our epistemic ascent, our positive knowledge is more selective, and we accumulate negative knowledge of higher and higher degree. It is the conjunction of these dual positive and negative movements that ensures us a more and more robust guarantee, even if we cannot get a guarantee that at a higher step, we would not have to restrict againour positive knowledge.

Can we claim that this conjunction justifies the first order assertion Kh? If for such a justification we require to have at our disposal the infinite se-

rie of $K \ldots \infty \ldots . Kh$, the answer is surely no. But this requirement seems to be itself an unjustified demand, because we can have cases of Kh without $K \ldots \infty \ldots . Kh$ — if it was not possible, why to distinguish the levels of knowledge? The other approach, a more modest and sensible one, leads to say that our first step, in which we have both (not KKh) and (K not K not h), is compatible with Kh, and that our second step, in which we have (not $KKKh$) and (KK not K not h) is compatible with KKh and then can imply Kh, and so on and so forth.

In this approach, we cannot consider questions about positive knowledge without taking into account the side of negative knowledge. This holds even if knowledge, as a modality, does not ensure the simple management of negation that would allow us to conclude from "I not know not p" that "I know p". The root of these troubles seems that knowledge, as factive, is related to simple truth, leading us to be satisfied by Kp without taking care of KKp, but as epistemic, is related to methods of justification, leading us to a quest for higher levels of knowledge. Whatever aspect you put the accent on, the common fact is that in any actual conditions of knowledge, Kp does not imply KKp.

This conclusion is the one endorsed by Williamson and Engel. Our approach adds a particular flavour to this proposition. We can use it to give solutions to vagueness problems. As soon as we have asymmetric clues (clues in favour of h, and no corresponding clues in favour of not h) and can go a step further by testing the robustness of our epistemic access to the clues for h while not noticing any clue for not h, we are justified to break the chain of the sorites reasoning. Similar solutions can be applied to higher order vagueness – when the results of higher order tests are vague at a first examination. This kind of solution does not require crisp data: the asymmetry between the clues for h and the clues for not h can be itself a vague asymmetry. This situation just leads us to a higher order inquiry about the clues in favour of a real asymmetry and the clues in favour of a fake asymmetry, and so on and so forth. As our modest and sensible approach does not require to get the whole stack of higher justifications, but just to build the following higher level in order to assert a knowledge of lower degree, this solution of vagueness does not lead us to an infinite regress.

4. Is perceptive knowledge based on comparisons?

This approach has been centred on examples in which knowledge consists in knowing to which of two categories a phenomenon belongs. It relates knowledge with an operation of comparison: comparing clues in favour of a category and clues in favour of the other one. Is every knowledge grounded on such comparison operations between bases for Kp and bases for K not p? The basic source of our knowledge, perception, seems to give us a simple knowledge of p, without any comparison. For example, we see this red spot and know that it is red, full point. According to Engel, such a knowledge Kr is given. Engel acknowledges that this is a dogmatic stance, but claims that this kind of modest dogmatism is inescapable. It is only when we ask for the justifications of this basic knowledge that we have to acknowledge that it has only *prima facie* justifications, which are defeasible ones, as they can be defeated if our inquiry goes further. Comparison, could Engel say, implies relations and the concept of difference, and perceptual knowledge is mainly non-conceptual.

I agree that perceptive knowledge is mainly non-conceptual, but argue that nevertheless it implies relation and comparisons. It is well known, for example, that given a tessellation of squares of different colours, the colour of a square located in the centre of the tessellation is perceived differently in relation with changes of the colours of squares that can be located far from the centre. The relation between a form and its background is central for perception, and focussing on different elements in the same picture can exchange their roles (for example in the Necker's cube). We perceive the same part of a landscape (grass with a few trees) as wooded or as a meadow when it is surrounded by fields without any tree or bush or by a dense forest.

Taking one element or another as the focussed clue gives rise to a content of perception that is in a sense the opposite of the other. These clues are not explicit parts of the perceptive content, but they are decisive for shifting from one content to the other. In the same way, in Peacocke example, the same form is perceived as a square or as a diamond, even when they are presented together. The discriminatory clue here is the parallelism with the vertical or the horizontal of either the sides (for the square) or the diagonals (for the diamond). It is not explicit in the perceptive content of each form, but it is decisive for the discrimination of one form as a square and the other as a diamond.

The difference between perception and judgment is neither the absence of relation and comparison (for perception), nor the presence of a balancing pro-

cess of taking into account a clue in favour of A and a clue in favour of not A- in judgements made in the uncertainty related to vagueness. The difference is that the process leading to the judgment can be made explicit, while we are most of the time unaware of the process leading to the content of perception and unable to explicit it. In both cases, the main operation is a process of discriminating data by using clues. But the clues can be made explicit in the conceptual judgment, while in perception the decisive clues are so intimately integrated in the perceptual content that they cannot be isolated. For example, in the Muller-Lyer illusion (lines with arrows directed towards the line or away from it) we can judge that the illusion is caused by this inversion of the direction of the arrows without being able to avoid to perceive one line longer than the other. The clues given by the arrows are compared, but the result of this comparison is integrated at very basic cognitive levels in the perceptual content and at a very higher speed that the one of the process of explicit judgment.

The difference underlying the distinction between conceptual and non-conceptual content seems to be the following: perception is a discrimination process using signs and clues. The relations between these clues give rise to the formation of the perceptive content, and by the way to a perceptive identification. In this identification, the clues are non-explicit and already integrated. This identification gives a basis for concepts. The process of judgement presupposes that several such identifications are possible. When the process of judgment works on clues, it has not only to integrate them in a unique identification, but also to take into account the clues that are in favour of an identification or in favour of another one. The one that the judgment selects has by this very selection process a conceptual content. A concept makes sense in a network of other concepts, while a perceptual identification makes sense in a network of clues. The judgment process may require the possibility of making explicit some of the clues that are used in order to discriminate two conceptual contents. To take again the example of the square and the diamond, the perceptual identification does not make explicit the clues given by the relation of the sides or the diagonal with the vertical and the horizontal, while the concepts of a form as a square and as a diamond (the same form in the example) require to make them explicit. If by "comparison" we refer to the explicit discrimination of two conceptual contents, there is no "comparison" in this sense in our basic perception. Nevertheless a perceptual content can also be said to imply comparisons in a non conceptual sense, as the process of integrating clues includes comparisons -but non explicit ones.

To say that the perceptual content has a *prima facie* justification is just to say

that the clues that justify the identification that gives this content are already integrated in the identification. In the case of a judgment, the notion of *prima facie* justification is slightly different; it refers to a specific stage of the process of evaluating the strength of the different clues: the stage in which a dissymmetry appears between the pro- and the contra-clues. As we have seen, at this stage an epistemic shift is possible: we can no longer focus on the category of the object of the judgment, but on the robustness or reliability of our access to the clues that entitle us to put it in this category, or on the validity of the relation between the clues and the categorization; we can make the first move of the epistemic ascent. By doing so, we start a (virtually endless) process of justification that goes beyond the *prima facie* justification, which appears now as an end in the perceptual process and a beginning in the conceptual and judgmental process. The perceptive *prima facie* justification is compatible with Kp and incompatible with (KK not p), but it is still not compatible with KKp. The judgmental *prima facie* justification is compatible with Kp and KKp, but still not with $KKKp$.

The dissymmetry of the perceptive clues between the ones that are pro-p and the ones that are contra-p ensures the identification of the item i as a p-object. On the basis of this identification, the inquiry about the validity of our epistemic access to this p-property of i can be triggered – it would have no sense to trigger such inquiry without any previous identification. This inquiry implies also to keep watching out for possible clues contra-p. But our watching is dissymmetric: regarding p, we are testing the robustness of our epistemic access to the contra-p clues; regarding not p we are just keeping watching out for possible contra p-clues. Regarding not p, we are still in the process that can result in identification, while regarding p we are involved in a higher level process.

5. Scepticism, belief and knowledge

As Engel says, we are entitled to take our knowledge as valid, even if our justification is a prima facie justification. The sceptic can attack this knowledge as defeasible, of course, and is tempted to extend this attack to every level of conceptual knowledge. But his attack wins only against the dogmatic that claims that knowing p implies an absolute knowledge, given by the infinite chain of $K \ldots \infty \ldots Kp$. His attack is not relevant against a theory of knowledge in which levels of knowledge are distinguished and "knowledge of p" is taken as a summary for " is valid for p the epistemic modality based on

the dissymmetry between (not KKp) and (K not K not p) and compatible with KKp, even if for the present time not compatible with $KKKp$ (no inquiry has been made at this higher level, so not $K\ KKp$ holds) while no longer compatible with (KK not p) – any of these formulas being possibly extended to similar expressions of higher levels of the epistemic ascent". In this approach, Kp can be true at its level even if $KKKp$ will prove not to hold.

We have to be more precise about justification, since *prima facie* justification is slightly different at the non-conceptual perceptive level and at the conceptual level. The sceptic may believe that he can attack perceptive knowledge by showing that the information given by the clues is an incomplete and insufficient one for concluding p. But his attack is irrelevant: this information and the dissymmetry between the clues is sufficient to identify the perceptive quality, form or object, even it is not sufficient to assign one conceptual category and not another one to them. Any attack on this conceptual assignment needs to presuppose a previous identification, and the sceptic has no power on the perceptive processes of integrating the clues, as he is, like us, unaware of them.

The sceptic's attack is relevant at the higher level of the conceptual judgment, when he tries to defeat the *prima facie* conceptual justification. But as we have seen, his attack cannot be valid once for all levels of knowledge, because the higher levels of the epistemic ascent cannot be built all together at the same time. In order to build a new level of epistemic inquiry, we have first to be given the evidence of an asymmetry between the pro- and contra-clues at our disposal at the previous level. This condition blocks the infinite ascent of the sceptical argument. As attacking the higher level of justification requires to presuppose the asymmetry at the previous level, the status of *prima facie* justification that rests on this asymmetry is enforced by the very move of epistemic ascent that the attack requires.

The sceptic can make a more general objection: what you describe in this dynamic of epistemic ascent is not a real knowledge, but only a belief. Belief that p is compatible with the truth of p and not compatible with the truth of not p, but does not ensure the truth of p, and this is all that you have got by your comparisons between clues for p and clues for not p. Our answer is that belief does not require and does not imply the possibility for any epistemic level to trigger a higher level inquiry about the robustness of our epistemic access. This possibility is present for knowledge from the first step, from the first asymmetry between pro-p clues and contra-p clues. Actualizing this possibility once the asymmetry has be recognized is a requirement of knowledge, is it not a requirement of belief.

Noticing this situation is an argument for Engel's claim that knowledge involves a normative aspect. But does it support Williamson's and Engel's claim that knowledge is more basic than belief? The answer is yes if we have a minimalist conception of belief, in which belief is reduced to the recognition of the asymmetry between the pro-and contra-clues, without any mention of a possibility to trigger an epistemic ascent. In order to understand conceptual cognition, we need to go further than this minimal belief. The answer is no if we define belief as a conjunction of the asymmetry and the possibility of triggering the following step, the epistemic ascent, but do not include in this conjunction the actualization of this possibility. In this conception, believing p implies the possibility of an inquiry for deeper justifications, but only knowledge requires that this following step of the cognitive process have been actualized. One given degree of knowledge shares a similarity with belief since it does not actualize levels of epistemic inquiry that are higher than the immediately following one. Our epistemic state is compatible with Kp if we have actualized an inquiry about our epistemic access to p rather than not p, leading to a conclusion that excludes (KK not p). Regarding the inquiry about higher levels, this epistemic state requires only its possibility. But it differs from belief since it involves the actualization of an inquiry of higher order than the previous step.

If we take two real epistemic states and want to justify to call one of them a belief that p and the other a knowledge that p, what will be their difference? Each of them is of course compatible with p, each of them implies the possibility of an epistemic ascent to the immediately higher level. Each of them can even pretend to be compatible with the knowledge of p! But this compatibility is only based on a pure possibility in the case of the belief, and is confirmed by the actualization of the epistemic inquiry at the higher order in the case of knowledge. When we take the same given level of epistemic state as a common reference for belief and knowledge, knowledge requires at least one step further than a belief, but can be considered as an improved stage of belief from the point of view of a higher level of knowledge.

18

Knowledge, Perception, and the Art of Camouflage *

JÉROME DOKIC

1. The Epistemic Conception of Perception

The relationship between perception and knowledge is notoriously complex. Many epistemologists are happy to acknowledge that perception can be a *source* of knowledge in a rather minimal sense: some of our perceptual experiences can convert some of our empirical beliefs to knowledge, or at least partly justify or warrant them. I know that there is a cat near the fence because this is what I see; my seeing the cat near the fence is an essential determinant of my favourable epistemic position.

In this essay I am concerned with a different and bolder claim, which is that perception itself is best conceived as an essentially epistemic state. Let us call this claim the Epistemic Conception of Perception, or ECP for short. The strongest version of ECP may be formulated as follows:

> (ECP-strong) Perception is a *form* of knowledge of what is perceived.

*As part of this modest homage (or should I say *hommage*) to Pascal Engel, I would like to add that my epistemological outlook owes much to him, first through his teachings in Geneva and Paris IV-Sorbonne, then through our philosophical discussions and collaboration over the years. I hope that what I have to say here will please his neo-Moorean ears.

On this version, seeing that there is a chair over there, or hearing the voice of a friend, amounts to or constitutively involves knowing that there is a chair over there, or that a friend is around. More generally, seeing a state of affairs is already a way of knowing that it obtains. Perception is a species of knowledge on a par with other species, such as memory or testimony.

Perhaps the most famous recent defence of ECP is Williamson (2000). As Williamson puts it: "If you really do see that it is raining, which is not simply to see the rain, then you know that it is raining; seeing that A is a way of knowing that A" (2000, p. 38). According to Williamson, seeing that something is the case is not simply seeing something or seeing a situation in which something is the case. When the subject's visual perception has a genuine propositional content, it amounts to a form of propositional knowledge. In a similar vein, Dretske (1969) notoriously calls "epistemic perception" cases in which the subject perceives that something is the case, and although he acknowledges that there are cases of non-epistemic perception, he insists that the latter lacks propositional content (I can see a cat near the fence without seeing that there is cat near the fence).

ECP-strong should not be confused with the claim that perceptual experience is a form of knowledge, which is plainly wrong. Perceptual experiences can be either veridical or non-veridical (in the latter case, they can be illusions or hallucinations). Obviously, non-veridical perceptual experiences cannot be a form of knowledge. Macbeth visually hallucinates a dagger before him, but he does not know that there is a dagger before him, for the good reason that there is none. ECP-strong is about veridical perceptual experiences.

The standard argument against ECP-strong turns on the belief-independence of perception, and in general perceptual experience. I can have the visual experience that there is a dagger before me while believing that I am hallucinating and thus that no dagger is actually there before me. Now suppose that my belief is false; my visual experience is in fact fully veridical. Even so, I cannot be said to know that there is a dagger before me since I do not even believe it. Belief is a necessary condition (although arguably not a constituent) of knowledge.

Williamson (2000) responds to the standard argument by suggesting that it puts more pressure on the link between knowledge and belief than on the thesis that perceiving a state of affairs is a way of knowing that it obtains. I am not convinced. How can I be said to *possess* knowledge that there is a dagger before me if I believe that there is none? At best my visual experience puts me in a *position* to know that there is dagger before me (see McDowell, 1998a, p. 390, n. 37). On this view, perception is still an essentially epistemic state. It can

yield knowledge provided that the subject forms the appropriate beliefs and there are no other, defeating beliefs (such as the belief that I am hallucinating). What is at stake here is a weaker version of ECP:

> (ECP-weak) Perception puts the subject in a *position* to gain knowledge about what is perceived.

Both versions of ECP entail that perception is a source of knowledge. Either perception is already knowledge (ECP-strong), or it is so to speak virtual knowledge (ECP-weak). In either case, it can ground perceptual beliefs and convert them to knowledge.

ECP is a widespread view among contemporary epistemologists, but there are dissenting voices. Consider for instance McGinn (1999, p. 20), who claims that "it is a mistake to make perception one's model or paradigm of knowledge". Here is McGinn's argument for his claim: "The concept of perception is plausibly understood as both causal and local: to perceive the fact that *p* requires that that fact should play an appropriate causal role in the production of an experience, and whether a particular experience counts as genuinely perceptual depends solely upon how that experience is related to the fact that causes it. [...] Knowledge depends upon the status of other relevant beliefs, but perception is possible without such global reliability" (1999, p. 20).

I wholeheartedly concur both with McGinn's claim and his general argument. What I would like to do here is to formulate what I see as a concrete and hitherto neglected challenge to ECP. According to both versions of ECP, there can be no gap between a perceptual experience that veridically presents a given state of affairs and an experience that can ground the knowledge that the state of affairs obtains. In other words, being perceptually related to the world is always epistemically significant. Against ECP, I shall present a particular case of perceptual experience in which (as I shall argue) the following triad of claims is true:

> (i) The experience presents a given state of affairs (it has propositional content).
>
> (ii) The experience is veridical.
>
> (iii) The experience cannot ground the knowledge that the state of affairs obtains (even in the absence of relevant defeaters).

To my mind the case to be presented casts doubt on the claim that our perceptual relation to the world is in essence an epistemic relation. Perception and

knowledge are different kinds of achievement, although of course the former can give rise to the latter in appropriate circumstances.

2. The challenge of perceptual hysteresis

Consider a phenomenon that is pervasive in the field of perception, namely what psychologists call "perceptual hysteresis". Broadly speaking, perceptual hysteresis is the maintenance of a perceptual experience with a relatively stable content over progressively degrading sensory stimulations.

Here is a typical perceptual hysteresis experiment (see Kleinschmidt et al., 2002). The subject faces a uniformly grey screen. A letter is segregated from the background by gradually increasing its contrast relative to the background. Following a short phase in which nothing changes while the letter is plainly visible (the "plateau"), the contrast is gradually decreased until the letter disappears out of view. Perceptual hysteresis arises as the threshold at which "pop out" occurs (i.e., when the subject becomes visually aware of the letter) is higher than the threshold at which "drop out" occurs (i.e., when the subject ceases to be visually aware of the letter).[1]

The critical hypothesis is that in the envisaged experiment, the boundary between the experience being veridical and it being non-veridical does not coincide with the boundary between knowledge and ignorance. There is some point between plateau and drop out – let us call it t^* – at which the subject veridically sees the letter but cannot know, just on the basis of her visual experience, that there is a letter on the screen, even if she believes that she is not hallucinating.

Knowledge is plausibly constrained by margin for error principles (Williamson, 2000). One cannot know that there is a letter on the screen if one's method of knowledge (looking at the screen) would have too easily produced the false belief that there is a letter on the screen. Let us assume that at t^* the relevant margin for error principle is not satisfied. On the basis of her visual experience, the subject can form a true belief that there is a letter on the screen, but this belief does not count as knowledge because the state of the screen at t^* is too close to states where the contrast is inexistent or invisible (i.e., the contrast is too weak or the screen has become uniformly grey again). In other words,

[1] Although perceptual hysteresis in such an experiment is typically quite high, it may vary across subjects. For some indication that perceptual hysteresis is significantly higher in schizophrenia (and especially for hallucinatory patients), see Martin et al. (submitted), which adapts Kleinschmidt et al. (2002)'s paradigm to the auditory case.

the state of the screen at t^* is too close to the boundary between the presence and the absence of the letter.

On the hypothesis under consideration, the subject at t^* can see *that* there is a letter on the screen. She does not merely see the letter or a situation in which there is a letter. Her experience has a genuine propositional content, which can be specified by a sentence such as "There is a letter on the screen". Given that the subject's experience is fully veridical, the boundary between veridical and non-veridical experiences should be placed *after* t^*, and thus after the boundary between knowledge and ignorance.

The point about perceptual hysteresis is that the subject's ability to see that there is a letter on the screen depends on her perceptual history. If the subject had looked at the state of the screen at t^* right away, i.e., independently of having seen the same letter before in better viewing conditions (e.g., during the plateau phase), she would not have seen the letter still to be seen (given the actual state of the screen). On this view, perceptual hysteresis *enhances* the subject's visual abilities.

If the hypothesis is true, the subject is at t^* in a situation that ECP predicts is impossible. First, the experience visually presents a letter on the screen. Second, the experience is veridical, and thus a letter is really there on the screen. Third, the experience cannot ground the visual knowledge that there is a letter on the screen, because the relevant margin for error has not been provided.

3. Three rejoinders

The foregoing argument rests on several assumptions that a defender of ECP might want to reject. Here I shall focus on what I think are the main objections that can be levelled from her perspective.

One option is to reject altogether the claim that perceptual knowledge is constrained by margin for error principles. Given that such principles can be shown to be derived from tight connections between knowledge and safety, which in turn correspond to robust epistemic intuitions (for a discussion see Engel 2007, Ch. 3), this option strikes me as particularly costly. As a consequence, I won't discuss it further here.

Another option available to the defender of ECP is to challenge the interpretation of the subject's experience at t^*. Even if it is accepted that due to perceptual hysteresis the subject has a visual experience as of a letter, one might object that this experience is not fully veridical, contrary to what I assumed above. If the experience is non-veridical, no counter-example to ECP

has been provided. Note that the experience can be non-veridical even though there is actually a letter on the screen. The objection is that the subject is not perceptually *related* to the letter. In other words, the subject's experience at t^* might be what Lewis (1980) calls (perhaps misleadingly) a "veridical hallucination". Its propositional content is correct (there is a letter on the screen), but by accident so to speak, so that there is no appropriate causal connection between the presence of the letter on the screen and the experience as of a letter on the screen.

In response to this objection, it should be conceded that perceptual hysteresis can in principle produce hallucinations, even "veridical" ones in Lewis's sense. For instance, in the experiment described above, a subject might still have the experience as of a letter after the screen has gone uniformly grey again. In this case, her visual experience is clearly hallucinatory. There might also be cases of veridical hallucinations in Lewis's sense: the subject seems to see a letter, a letter is still visibly there on the screen, but the subject is no more visually related to it.[2]

What is less clear is *why* the subject's experience at t^* must already be classified as hallucinatory. Perceptual hysteresis has evolutionary advantages, among which is the ability to detect camouflaged predators. As Large et al. (2004) write, by way of introducing the phenomenon of perceptual hysteresis, "a camouflaged animal might not be noticed until it moves, but once it has moved, it becomes clearly visible, and *can remain so* when it resumes a fixed position" (p. 3453; my italics). The value of perceptual hysteresis is to extend one's perceptual abilities in some situations, at the cost of producing false positives, i.e., perceptual hallucinations, in other situations.

On the hypothesis under consideration, perceptual hysteresis can extend one's perceptual abilities in epistemically unsafe situations. For instance, even if the subject cannot know that there is a camouflaged animal over there, just by looking at the bush, she might be able to visually detect the animal thanks to her favourable perceptual history. The subject has a veridical experience of the animal, even if her experience is not *safe*, i.e., is not reliably connected to the animal.

A third possible option is to bite the bullet and insist that at t^* the subject has perceptual knowledge that there is a letter on the screen, despite the fact that at that time the relevant margin for error principle is not satisfied. This is

[2] In the study of Martin et al. (submitted), 40% of schizophrenic patients report hearing the signal when the signal-to-noise ratio has gone down to -30db. At this point, they clearly have auditory hallucinations, perhaps even "non-veridical" ones in Lewis's sense (although the signal is still physically there, it is much below the normal threshold of audition).

not necessarily to go back to the first option. An advocate of this option can acknowledge that perceptual knowledge is constrained by margin for error principles. What she claims is that the *previous* satisfaction of the relevant margin for error principle guarantees that the subject perceptually knows, at the later time t^*, that there is a letter on the screen.

In a nutshell, the idea is this. During the plateau phase, the subject visually knows that there is a letter on the screen. The relevant margin for error principle is satisfied: the subject might not easily have been wrong. Then perceptual hysteresis allows the subject's perceptual belief to *retain* its original epistemic status so that it still counts as knowledge at t^*. Here an analogy with memory beliefs can be drawn. A memory belief may count as knowledge even if the subject has long forgotten the reasons that have justified or warranted it in the past. Thanks to the subject's faculty of long-term memory, her belief has retained its original epistemic status.[3]

However, the analogy between perceptual hysteresis and memory might not be a good one. First, perceptual hysteresis should not be conceived as a personal-level memory phenomenon. It is much more primitive than that, and even though it depends on top-down activity, the relevant "top" level is still internal to our perceptual systems. Second, the analogy breaks down at a crucial point. Memory beliefs are (typically) beliefs about the past, while the perceptual beliefs at stake concern present states of affairs, so that their "Fregean" contents are different at each time. It is much less plausible that a justification for a given perceptual belief (that there is a letter on the screen at t_1) can be recruited as a justification for a logically independent belief (that there is a letter on the screen at t_2, where t_1t and t_2 are different times).

McDowell (1998b) allows for knowledge of "changeable, though reasonably durable, states of affairs", such as the fact that François Hollande is the President of France, independently of repeated acquisition of updated pieces of evidence. The idea is that "like a living thing, such knowledge needs something analogous to nutrition from time to time, in the shape of intermittent confirmation that the state of affairs known to obtain does still obtain" (1998b, p. 426). However, the case of perceptual hysteresis seems quite different, at least on the face of it. What is lacking in the latter case is precisely anything analogous to epistemic nutrition. The subject's perceptual experience at t^* is entirely dependent on a previous experience with high epistemic status (during the plateau phase), but is otherwise epistemically detached from the world

[3] See for instance what Bernecker (2009, Section 3.2) calls "the principle of continuous justification".

(since by definition there is no margin for error at t^*). It is hard to believe that its epistemic status can still be the same as that of the previous experience. Moreover, the notion of doxastic responsibility, which McDowell argues is necessarily involved in the preservation of the relevant kind of knowledge, does not apply at the level where perceptual hysteresis arises. Perceptual hysteresis is best seen as a low-level, pre-rational phenomenon.

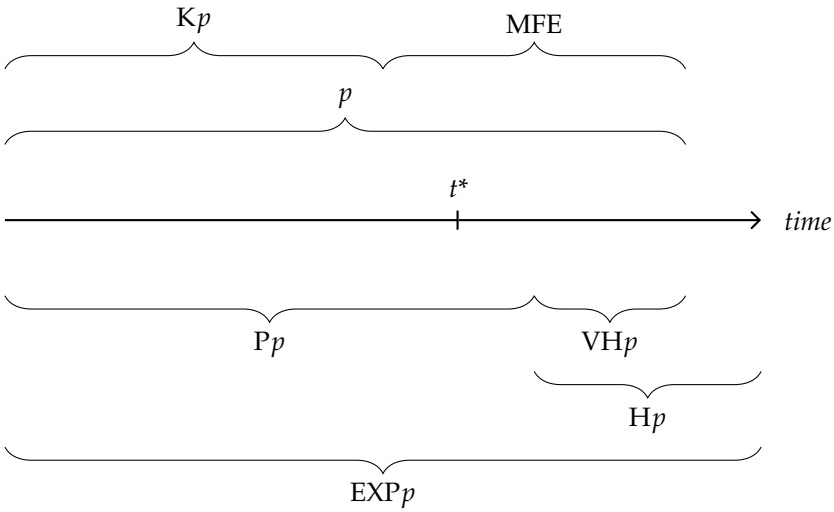
4. Conclusion: perceptual content *vs* perceptual confidence

I have dealt with three main options that a defender of ECP can pursue in order to accommodate perceptual hysteresis within her conception. No doubt there are others but at least these are the most important ones I could think of.

Now of course I have not given a direct argument against ECP. I have only pointed out that it is *a priori* plausible that we can see state of affairs that we cannot know to obtain just on the basis of our visual experience. The defender of ECP has the burden of proving that the boundary between perception and illusion/hallucination coincides with the boundary between knowledge and ignorance. If these two boundaries come apart, then perception cannot be considered to be epistemic in either ECP-strong's or ECP-weak's sense (see Figure 18.1 for a schematic recapitulation of the main distinctions at work here).

Rejecting ECP does not entail denying that perception can be a source of knowledge. Some perceptual experiences are such that in appropriate circumstances (including for instance the absence of relevant defeaters) beliefs formed on their basis amount to knowledge. If the argument from perceptual hysteresis sketched in this essay is correct, not any perceptual relation to the world can do the trick, though. Only some perceptual experiences will have the required epistemic property.

As I have tried to show elsewhere, the formation of warranted perceptual beliefs is sensitive not merely to the *content* but also to the *quality* of experience. Even though the content of the subject's experience can remain the same from the plateau phase to t^* ("There is a letter on the screen"), the quality of the experience may move from optimality to sub-optimality. At the plateau phase, the subject feels quite confident that there is a letter on the screen, while at t^* the subject may say something like: "I am not quite sure, but I would still say that there is a letter on the screen". Whether the subject forms safe "outright" perceptual beliefs thus depends on metaperceptual feelings of confidence or certainty (see Dokic, forthcoming, and also Dokic & Martin,



N.B. The time arrow goes from the plateau phase to some point after drop out.
p: there is a letter on the screen.
Kp: the subject knows p.
Pp: the subject perceives that p.
Hp: the subject hallucinates that p.
VHp: the subject veridically hallucinates that p.
EXPp: the subject seems to perceive that p.

Figure 18.1: Graphic representation of some of the relevant distinctions.

2012).

There is some empirical evidence that metaperceptual feelings are based on subpersonal mechanisms which monitor the quality of our current experience independently of its content. Assuming the Bayesian model of perception, these mechanisms may assess the extent to which the perceptual response is stimulus-driven rather than modulated by top-down processes or "priors" (see Barthelmé & Mamassian, 2010). For instance, during the plateau phase, sensory stimulations are given more weight in the generation of the perceptual response, while at t^*, priors partly determined by the subject's pre-

vious experience get predominance.

Now perceptual hysteresis extends one's perceptual abilities but not necessarily one's perceptual confidence. If perceptual hysteresis operates at the level of the content of experience, which it contributes to stabilize against the relative degradation of sensory stimulation, it might not affect further, meta-perceptual processing.

5. References

Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proc. Natl Acad. Sci.* 107:48, pp. 20834-20839.

Bernecker, S. (2009). *Memory. A Philosophical Study*. Oxford, Oxford University Press.

Dokic, J. (forthcoming), Feelings of (un)certainty and margins for error. *Philosophical Inquiries*.

Dokic, J., & Martin, J.-R. (2012). Disjunctivism, Hallucinations, and Metacognition. *WIREs Cogn Sci*. 3:533-543. doi: 10.1002/wcs.1190

Dretske, F. (1969). *Seeing and Knowing*. Chicago, The University of Chicago Press.

Engel, P., (2007). *Va Savoir !* Paris, Hermann.

Kleinschmidt, A., Buchel, C., Hutton, C., Friston, K.J., and Frackowiak, R. S. (2002). The neural structures expressing perceptual hysteresis in visual letter recognition. *Neuron* 34, 659-666.

Large, M.-E., Aldcroft, A., and Vilis, T. (2005). Perceptual Continuity and the Emergence of Perceptual Persistence in the Ventral Visual Pathway. *Journal of Neurophysiology* 93, 3453-3462.

Lewis, D. (1980). Veridical Hallucinations. *Australasian Journal of Philosophy* 58, 239-249.

Martin, J.-R., Dezecache, G., Bruno, N., Dokic, J. Demily, C., Pacherie, E., Franck, N. (submitted). Perceptual hysteresis uncovers the presence of sensory persistence biases in people with schizophrenia.

McDowell, J. (1998a). Criteria, Defeasibility, and Knowledge. In *Meaning, Knowledge, and Reality* (pp. 369-394). Harvard, Harvard University Press.

McDowell, J. (1998b). Knowledge by Hearsay. In *Meaning, Knowledge, and Reality* (pp. 414-443). Harvard, Harvard University Press.

McGinn, C. (1999). The Concept of Knowledge. *In Knowledge and Reality. Selected Essays* (pp. 7-35). New York, Oxford University Press.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford, Oxford University Press.

19

Il n'y a pas de croyances « gettierisées »

BENOIT GAULTIER

Résumé Edmund Gettier est censé avoir montré qu'il ne suffit pas qu'une croyance soit vraie et justifiée pour être une connaissance. Je soutiens que l'argumentation de Gettier peut être rejetée indépendamment de toute théorie de la justification ou de la connaissance. Je défends l'idée que s'il n'y a pas de volontarisme doxastique, les croyances dites « gettierisées » ne peuvent en réalité tout simplement pas être formées (à moins d'adopter une conception téléologique de la croyance). Après avoir répondu à quelques objections, je procède à l'analyse de cas de Gettier classiques sur la base de la thèse que j'ai défendue. J'essaie d'identifier pour finir ce que les cas imaginés par Gettier ont de spécifique et qui pourrait expliquer l'extraordinaire postérité de son article.

Abstract Edmund Gettier is supposed to have shown that it is not enough for a belief to be true and justified for it to constitute knowledge. I contend that the reasoning at work in Gettier cases can be rejected independently from any theory of knowledge and justification. The claim I argue for is that if doxastic involuntarism is true, the so-called gettierized beliefs simply cannot be formed (unless one subscribes to a teleological conception of belief). After having answered some objections to my view, I carry out a detailed analysis of the classic types of Gettier cases on the basis of the thesis I have defended. To conclude, I try to identify what is distinctive about Gettier's paper which could explain its extraordinary posterity.

Mots-clés : Cas de Gettier ; croyance ; involontarisme doxastique ; éléments de preuve.

Edmund Gettier est censé avoir montré dans « Is Justified True Belief Knowledge ? » qu'il ne suffit pas qu'une croyance soit vraie et justifiée pour constituer une connaissance. L'argumentation ou, plus exactement, les contre-exemples qu'il imagine pour parvenir à cette conclusion permettent-ils cependant de l'établir ? Deux angles d'attaque semblent possibles : soit contester que les croyances vraies dont il est question dans les contre-exemples imaginés par Gettier soient réellement justifiées ; soit soutenir qu'on a bel et bien affaire à des connaissances. Dans les deux cas, Gettier n'aurait pas montré qu'une croyance vraie et justifiée n'est pas nécessairement une connaissance.

Parce qu'il semble particulièrement contre-intuitif de soutenir que les croyances imaginées par Gettier constituent bel et bien des connaissances — comme le fait par exemple Crispin Sartwell (1992) pour qui la simple vérité d'une croyance peut suffire à en faire une connaissance —, les critiques formulées contre l'article de Gettier ont consisté dans leur immense majorité à contester que les croyances en question soient réellement justifiées[1]. Autrement dit, les contre-exemples de Gettier reposeraient implicitement sur une conception, sinon erronée, du moins contestable de la justification. De sorte que, d'une manière plus générale, l'article de Gettier ne devrait pas faire davantage autorité que n'importe quelle position épistémologique à propos de la justification ou de la connaissance.

La position que je vais défendre est que les cas imaginés par Gettier ne permettent pas d'établir qu'une croyance peut être vraie et justifiée sans être une connaissance. Mais je n'affirmerai ni que les croyances qu'il considère, dans ces cas, être vraies et justifiées constituent en réalité des connaissances, ni qu'elles ne sont pas justifiées. Ce que je soutiens est que, dans les situations que Gettier imagine, ces croyances ne peuvent tout simplement pas être formées. Autrement dit, c'est indépendamment de toute théorie de la justification ou de la nature de la connaissance que l'on peut avancer que l'auteur de « Is Justified True Belief Knowledge ? » n'a pas montré ce qu'il prétendait avoir montré.

[1] Cf. par exemple les réserves de Keith Lehrer (1965) ou d'Alvin Goldman (1967) sur la conception de la justification implicitement contenue dans l'article de Gettier (réserves qui, dans leur cas, ne sont cependant pas associées à un rejet de la conclusion de Gettier).

1. Contre la possibilité de croyances « gettierisées »

Je partirai, à des fins de clarté d'exposition, d'un cas de Gettier classique qui ne se trouve pas dans l'article de 1963 mais qui instancie clairement le principe fondamental d'après lequel sont construits tous les cas de Gettier[2].

Imaginons que je visite une entreprise et qu'un de ses salariés, Jean, me dise qu'il possède une Ford. Imaginons par ailleurs que j'aie de très bonnes raisons de le croire : Jean me montre avec fierté son porte-clés Ford ainsi qu'une carte grise indiquant la puissance fiscale du modèle de Ford dont il me parle ; un de ses clients l'appelle au téléphone et j'entends ce client lui demander s'il est toujours satisfait de sa Ford ; plusieurs de ses collègues de longue date me disent à propos de Jean, pendant qu'il est au téléphone, qu'il est extrêmement aimable et l'honnêteté faite homme. Je crois en conséquence que Jean possède une Ford. J'en infère alors la croyance que quelqu'un dans l'entreprise possède une Ford. Il s'avère cependant que 1) Jean m'a menti et ne possède pas de Ford et que 2) un autre employé de l'entreprise, Martin, à qui je n'ai pas parlé et dont je ne connais rien, en possède une. Il s'ensuit que ma croyance que Jean possède une Ford est justifiée et, *ipso facto*, ma croyance que quelqu'un dans l'entreprise possède une Ford. Mais tandis que la première est fausse, la seconde est vraie, puisque Martin en possède une. Cette seconde croyance est donc vraie et justifiée mais, intuitivement, ne constitue pas une connaissance : je ne sais pas que quelqu'un dans l'entreprise possède une Ford.

L'idée selon laquelle, dans un tel cas, je crois de façon vraie (et justifiée) que quelqu'un dans l'entreprise possède une Ford est cependant contestable. Partons d'un point relativement trivial qui, en lui-même, ne suffit pas à rejeter cette idée mais qui ne doit pas pour autant être négligé : la croyance que je pourrais exprimer en disant que quelqu'un dans l'entreprise possède une Ford pourrait être une croyance non pas à propos de n'importe quel employé de l'entreprise mais à propos de Jean en particulier. C'est Jean et lui seul que je pourrais viser à travers la description vague « quelqu'un dans l'entreprise ». Imaginons, pour illustrer ce point, que je voie mon frère Paul fréquenter depuis quelque temps une ravissante jeune fille et adopter, avant de se rendre aux rendez-vous qu'il a avec elle, un comportement typique d'un jeune homme amoureux. Imaginons ensuite qu'au cours d'un repas de famille dominical, j'aie envie de taquiner Paul et que je dise à haute voix, un sourire

[2]Comme un relecteur anonyme me l'a fait remarquer, je dois préciser que ce cas constitue une légère variante de celui de Lehrer (1965).

en coin : « Quelqu'un est amoureux autour de cette table… ». Imaginons enfin que Paul ne soit en réalité nullement amoureux de cette jeune fille mais que mon cousin Pierre, présent à ce repas, le soit secrètement. Il est évident en ce cas que la croyance que j'exprime en disant « Quelqu'un est amoureux autour de cette table » porte sur Paul. Elle est donc fausse et ne saurait être rendue vraie par le fait que Pierre soit amoureux de cette jeune fille. Je serais de bien mauvaise foi si j'apprenais la vérité et que je disais quelque chose comme : « Je ne me suis pas trompé, puisque Pierre en était bel et bien amoureux »[3].

Sans doute remarquera-t-on alors *et à juste titre* : « Soit. La croyance que Jean possède une Ford est fausse et il se pourrait qu'en *disant* que quelqu'un dans l'entreprise possède une Ford, ce soit Jean que vous visiez, exprimant par là votre croyance que *Jean* possède une Ford. Mais en quoi ce point évident est-il pertinent pour l'analyse philosophique des cas de Gettier ? Dans ceux-ci, il s'agit d'imaginer que vous formiez non seulement la croyance que Jean possède une Ford mais, *en plus*, la croyance *différente* que quelqu'un dans l'entreprise possède une Ford (en vous disant, de façon absolument incontestable, que si Jean possède une Ford, alors nécessairement quelqu'un possède une Ford[4]). Et c'est avec l'introduction de cette seconde croyance que tout se joue. »

Tout se joue effectivement sur ce point. Mais alors que les débats auxquels a donné lieu l'article de Gettier ont tous porté sur *le statut épistémologique* de cette seconde croyance, la question de *sa possibilité*, spontanément admise par Gettier, n'a pas, à ma connaissance, été discutée. Or, à cette question, il faut répondre par la négative : il n'est pas possible que je forme la croyance que *quelqu'un* dans l'entreprise possède une Ford lorsque je crois déjà que *Jean* possède une Ford. Plus exactement, ma croyance que *Jean* possède une Ford, telle qu'elle a été formée, ne peut pas entraîner la formation de la croyance que

[3] Autrement dit, le sens auquel « Quelqu'un est amoureux autour de cette table » peut être rendu vrai par le fait que Pierre ou n'importe quel autre convive soit effectivement amoureux n'est pas celui auquel j'ai prononcé cette phrase : en m'exprimant comme je l'ai fait, avec le ton et l'attitude qui était la mienne, je voulais justement dire que je croyais de quelqu'un *en particulier* autour de cette table, Paul en l'occurrence, qu'il était amoureux. La littérature relative à la manière de caractériser ou de conceptualiser ce genre de point est, comme toujours, aussi importante en volume et en enjeux philosophiques que les positions qui s'y affrontent sont fines et précises. Mais tout ce qui m'importe ici est que le point en question apparaisse évident.

[4] Il reviendrait absolument au même de dire que dans les cas de Gettier il s'agit d'imaginer que l'on forme non seulement la croyance que Jean possède une Ford mais, en plus, la croyance différente que le nombre de possesseurs de Ford dans l'entreprise est plus grand que zéro (en se disant, de façon absolument incontestable, que si Jean possède une Ford, alors nécessairement le nombre de possesseurs de Ford dans l'entreprise est plus grand que zéro).

quelqu'un possède une Ford. Mais si cette croyance ne peut pas être formée, la question de son statut épistémologique est sans pertinence, ce dont il suit que Gettier n'a pas montré qu'une croyance vraie et justifiée peut ne pas être une connaissance.

La raison pour laquelle ma croyance que *Jean* possède une Ford ne peut pas entraîner la formation de la croyance que *quelqu'un* possède une Ford peut s'énoncer de façon de la façon suivante : tout comme, à propos de la question de savoir si p, je ne peux pas croire *davantage* que ce que m'apparaissent établir ou supporter à ce propos les données que je considère pertinentes, je ne peux pas croire *moins* que ce qu'elles m'apparaissent établir ou supporter. *Plus précisément*, s'il est vrai que l'on ne peut pas croire à volonté — autrement dit, s'il est vrai que je ne peux pas croire *autre chose* à t à propos de la question de savoir si p que ce que les éléments de preuve E dont j'estime disposer à t à propos de la question de savoir si p me semblent établir à t —, alors il s'ensuit que je ne peux pas croire à t à propos de la question de savoir si p quelque chose *de plus faible* (de plus indéfini ou indéterminé) que ce que E m'apparaît établir à t à propos de cette question. *Plus exactement*, je ne peux pas, *en plus* du fait de croire ce que E m'apparaît établir, croire en m'appuyant sur E et sur E seulement quelque chose de plus faible que ce que E m'apparaît établir. Pour que, dans le cas en question, je puisse croire que *quelqu'un* possède une Ford, il faudrait, autrement dit, que je dispose d'autres éléments de preuve que ceux qui m'apparaissent établir que quelqu'un possède une Ford *uniquement en m'apparaissant établir que tel individu en possède une*. Ce qui n'est ici pas le cas. Si par exemple j'avais remarqué, en arrivant sur le parking de l'entreprise, une Ford garée dans la partie du parking dont l'accès est contrôlé et strictement réservé aux employés, j'aurais eu à ma disposition une preuve que quelqu'un dans l'entreprise possède une Ford ; et elle aurait été capable de me conduire à croire que *quelqu'un* dans l'entreprise en possède une, parce qu'il ne se serait pas agi d'une preuve que *tel* individu de l'entreprise en possède une.

Il importe ici d'écarter deux objections que l'on peut être assez spontanément enclin à opposer à l'idée que je viens d'exposer de façon synthétique, et ainsi d'en clarifier la signification et d'en dégager les conséquences[5].

La première de ces objections est la suivante : si quelqu'un arrive dans l'entreprise et demande « Qui ici croit que quelqu'un possède une Ford ? »,

[5] Je remercie le même relecteur anonyme d'avoir pris la peine de formuler l'essentiel des deux objections qui vont être mentionnées et de m'avoir ainsi conduit à réaliser que je n'aurais pas dû compter sur le fait que leur réponse était déjà plus ou moins implicitement contenue dans les deux derniers paragraphes de cette section pour emporter la conviction du lecteur.

dois-je lever la main ou pas ? Le point soutenu dans le paragraphe précédent n'implique-t-il pas que je doive ne pas la lever dès lors que je crois que *John* possède une Ford ? Mais ceci est parfaitement contre-intuitif. De la même manière, si l'on me demande, au tribunal par exemple, si je crois que quelqu'un habite à l'étage supérieur de mon immeuble, et que je réponds « non » parce que mes données supportent davantage – c'est-à-dire parce que je crois que Jeanne, Karl et Lucy y habitent – on m'accusera à juste titre de parjure. Or l'idée qui a été exposée ne conduit-elle pas à soutenir, de façon clairement inacceptable, qu'il n'y a pas de parjure en ce cas ?

La seconde objection est la suivante : en toute rigueur, d'après l'idée en question, il n'est même pas possible que je croie que *Jean* possède une Ford ; si l'on ne peut pas croire moins que ce qu'apparaissent supporter les données dont on dispose, ce que je crois est que Jean, qui a un porte-clés Ford, qui est assis à ce bureau, qui est brun, qui à l'air d'avoir moins de soixante ans, qui porte aujourd'hui un polo bleu, qui est humain, qui vit sur Terre, etc., possède une Ford. Et ceci fait qu'il est faux de dire de moi, d'après l'idée qui a été exposée, que je crois que Jean possède une Ford. La totalité de ce que je crois à propos de Jean (mais aussi des véhicules Ford et de ce que signifie en France le fait d'être le « possesseur » d'une voiture) est la seule chose qui soit réellement crue lorsque l'on dit (ou que je dis) de moi que je crois que Jean possède une Ford[6] – ce qui est franchement contre-intuitif. Je dois pouvoir croire que S est P même quand je crois, étant donné la totalité des données E dont je dispose, que S est P, que S est Q et que S est R. Ne pas pouvoir croire moins que ce que E apparaît supporter signifie simplement, de façon quasi tautologique, que je ne peux pas ne pas croire quelque chose qui m'apparaît supporté ou établi par E.

D'une manière plus générale, lorsque je crois qu'un fait – celui que quelqu'un possède une Ford – est une conséquence logique de ce que je crois étant donné les éléments de preuve E dont j'estime disposer – Jean possède une Ford –, il n'est pas réellement question en ce cas de croire *autre chose* que ce que je crois étant donné E. Dès lors, il n'est pas nécessaire de pouvoir croire à volonté pour pouvoir croire que quelqu'un possède une Ford lorsque les données dont je dispose m'apparaissent établir que Jean possède une Ford.

La réponse à la première objection est simple : je dois lever la main si l'on me demande « Qui ici croit que quelqu'un possède une Ford ? » et je dois répondre « oui » à la question « Croyez-vous que quelqu'un habite à l'étage

[6] Ce qui, d'après cette objection, limite de façon assez drastique le nombre de mes croyances et donc, corrélativement, de mes connaissances.

supérieur de l'immeuble ? ». Je ne dois cependant pas répondre ainsi parce que je croirais non seulement que Jean possède une Ford ou que Jeanne, Karl et Lucy habitent à l'étage supérieur mais *également* que quelqu'un en possède une ou que quelqu'un y habite : je dois dire « oui » au président du tribunal tout simplement parce que « … que quelqu'un habite à l'étage supérieur de l'immeuble » constitue une caractérisation vraie ou correcte de ce que je crois. Elle est bien plus vague, bien moins précise, que « … que Jeanne, Karl et Lucy habitent à l'étage supérieur de l'immeuble » mais elle n'en est pas moins vraie. Si le président m'avait demandé « Croyez-vous que Jeanne, Karl et Lucy ici présentes habitent à l'étage supérieur de l'immeuble ? », j'aurais pu répondre quelque chose comme « Oui, *exactement* », parce qu'il s'agit d'une bien meilleure caractérisation de ce que je crois que « … que quelqu'un habite à l'étage supérieur de l'immeuble ». Mais cela ne rend nullement fausse cette dernière. Telle est la raison pour laquelle j'aurais à juste titre été accusé de parjure si j'avais répondu « non » à la question du président.

La deuxième objection soulève en partie la difficile question de l'individuation et de l'identification des croyances. On peut cependant déjà remarquer que, de la réponse qui vient d'être apportée à la première objection, il suit que la position défendue dans cet article n'implique nullement que, du fait que ce que je crois soit plus précis ou riche que cela, il soit faux de dire que je crois que *quelqu'un* possède une Ford ou faux de dire que je crois que *Jean* possède une Ford. Ce qu'énonce cette position est qu'il n'est rien de tel, dans le cas en question, qu'une croyance que quelqu'un dans l'entreprise possède une Ford qui serait distincte de la croyance que Jean possède une Ford. Mais rien n'interdit de dire que je crois que *quelqu'un* possède une Ford si l'on prétend fournir par là une caractérisation correcte de ce que je crois.

Qu'en est-il alors de la nature de la croyance que Jean possède une Ford relativement au fait de croire qu'il soit brun, qu'il porte un polo bleu aujourd'hui, qu'il possède un porte-clés Ford, qu'il ait l'air d'avoir moins de soixante ans, qu'il soit un employé de l'entreprise, qu'il soit humain, qu'il vive sur Terre, etc. ? La position défendue ici conduit-elle à soutenir, de façon clairement contre-intuitive, que dire que je crois que Jean possède une Ford consiste à caractériser de façon correcte (mais partielle) une unique croyance, quasi infinie, englobant tout ce que je crois à propos de Jean ? Nullement. Sans prétendre ni avoir à proposer une théorie générale de l'individuation et de l'identification des croyances, ce qu'énonce cette position est uniquement que lorsque les seuls éléments de preuve dont je dispose à l'appui du fait que p^* soit le cas (par exemple que quelqu'un possède une Ford) sont ceux qui m'ont conduit à former la croyance plus précise, riche, ou déterminée que p

(par exemple que Jean possède une Ford), je ne crois pas que *p* en plus du fait de croire que* p. Autrement dit, dans une situation de ce genre, on ne peut dire de façon véridique que je crois que *p** qu'à la condition de ne pas vouloir dire par là que j'ai formé, en plus du fait de croire que *p*, la croyance distincte que *p**. Par ailleurs, parce que les seuls éléments de preuve dont je dispose à l'appui du fait que *quelqu'un*, ou même que *Jean*, possède une Ford ne sont pas ceux qui m'ont conduit à croire que Jean est brun, est humain, ou porte aujourd'hui un polo bleu, il ne suit nullement de la position défendue dans cet article que croire que quelqu'un, ou même que Jean, possède une Ford consiste à croire que Jean, qui est brun, qui porte un polo bleu, etc., possède une Ford. Autrement dit, il ne suit nullement de cette position que dire que je crois que quelqu'un, ou que Jean, possède une Ford consiste à caractériser (correctement mais partiellement) une unique croyance comprenant tout ce que je crois à propos de Jean.

Considérons, pour illustrer les points qui viennent d'être avancés, une autre situation : imaginons que, de façon tragique, je voie, en entrant dans le bureau de mon cousin Pierre, ce dernier assassiné par mon frère Paul. Dans ce cas, je ne peux pas former, *en plus* de la croyance que *Paul* a assassiné Pierre, la croyance que *quelqu'un* a assassiné Pierre. Ici, la croyance que *quelqu'un* a assassiné Pierre ne pourrait être formée qu'en niant (pathologiquement) l'évidence, c'est-à-dire en niant que mon frère soit l'assassin. Quand, à la suite de Gettier, les épistémologues soutiennent qu'il est parfaitement possible que je forme la croyance additionnelle que quelqu'un a assassiné Pierre quand je crois déjà qu'il a été assassiné par Paul, *ce que je crois en réalité est que le fait que Paul ait assassiné Pierre peut être décrit de façon véridique en disant que quelqu'un l'a assassiné.* Je crois – parce que je crois Paul a assassiné Pierre – qu'il s'agit d'une description vraie de ce qui s'est passé et de ce que je crois ; je crois que l'affirmation ou l'énoncé que quelqu'un a assassiné Pierre est vrai. Mais cela ne prouve nullement qu'en plus du fait de croire que Paul a assassiné Pierre je croie que quelqu'un l'a assassiné. Cela ne le prouverait qu'à la condition de soutenir que, principiellement, à toutes les manières dont je peux exprimer ou décrire ma croyance que *p*, ou le fait que *p*, correspondent des croyances venant s'ajouter à ma croyance que *p* – ce qui ne semble pas avoir la moindre plausibilité[7]. La position que je soutiens est que lorsque ces croyances addi-

[7] Ce point s'applique, *ceteris paribus*, aux sentiments : si je suis enchanté *par les bonnes nouvelles que j'ai reçues de mon vieil ami Duncan*, mon sentiment peut être décrit de façon correcte en disant que je suis enchanté par quelque chose, et je peux parfaitement considérer qu'il s'agit là d'une description vraie de ce que je ressens. Mais cela ne prouve en aucune façon qu'un sentiment additionnel explique la vérité de cette description : celui d'être enchanté *par quelque chose.*

tionnelles supposées n'ajoutent rien au contenu de ma croyance que *p*[8] et que *les seuls éléments de preuve que j'aurais à l'esprit et que j'invoquerais* si l'on me demandait la raison pour laquelle je suis prêt à affirmer que *quelqu'un* a assassiné Pierre sont ceux qui m'ont conduit à former la croyance que *Paul* a assassiné Pierre, il ne s'agit pas de croyances additionnelles, mais de caractérisations additionnelles de ma croyance que Paul a assassiné Pierre.

Considérons pour finir la variante suivante du cas précédent : imaginons cette fois que je découvre, en entrant dans son bureau, que mon cousin Pierre a été assassiné ; imaginons ensuite que je mène l'enquête pour découvrir l'assassin et que les éléments de preuve que j'accumule me conduisent à croire que mon frère Paul a assassiné Pierre. Puis-je, en ce cas, non seulement croire que *Paul* a assassiné Pierre mais *en plus*, comme je le croyais avant de commencer mon enquête, croire que *quelqu'un* a assassiné Pierre ? La réponse à cette question est la suivante : parce que les éléments de preuve que j'ai collectés en entrant dans le bureau de Pierre (comme le fait de l'avoir trouvé mort, frappé à la tête par un chandelier laissé sur les lieux du crime) qui m'ont conduit à croire que quelqu'un l'a assassiné ne m'apparaissent pas, même après mon enquête, établir que quelqu'un a assassiné Pierre *uniquement en m'apparaissant établir que Paul l'a assassiné*, je peux cette fois croire simultanément les deux choses. En effet, dans cette situation, la croyance que quelqu'un a assassiné Pierre ne constitue pas un simple affaiblissement de la croyance que Paul a assassiné Pierre, contrairement à ce qui est censé se passer dans le cas de Gettier de la Ford.[9]

[8] Il s'agit même précisément du contraire, puisqu'elles consistent en un affaiblissement de son contenu.

[9] Une autre manière d'appuyer la position que je défends pourrait être celle-ci : à supposer que l'on accepte l'idée (qui me semble pour l'essentiel peu contestable) qu'un individu a formé la croyance que *p* à *t-1* si, à *t0*, il serait surpris d'apprendre ou de réaliser qu'il n'est pas le cas que *p*, la question que l'on peut poser au philosophe qui tient pour évidente l'argumentation de Gettier est la suivante : est-il réellement concevable que, en apprenant que Jean ne possède pas de Ford, vienne s'ajouter à ma surprise que Jean ne possède pas de Ford une autre surprise – celle que personne dans l'entreprise n'en possède (à supposer que je croie par ailleurs que tous les autres employés possèdent des BMW) ? S'il faut répondre par la négative, c'est que ce ne sont pas deux croyances distinctes qui avaient été formées mais uniquement la croyance fausse que Jean possède une Ford, qui peut se trouver (correctement mais partiellement) caractérisée comme croyance que quelqu'un possède une Ford.

2. Une autre objection : croire que *p* et croire que « p » est vrai

Il est possible que l'on soit enclin à objecter que l'argument que j'ai avancé ne peut pas être correct puisque si je crois que l'affirmation que quelqu'un a assassiné Pierre est vraie, il s'ensuit que je crois que quelqu'un a assassiné Pierre. Cette objection repose cependant sur une confusion qui doit être dissipée : croire que *p* ne consiste pas à croire que l'affirmation ou l'énoncé « p » est vrai[10]. Il est possible d'établir ce point de bien des façons. J'utiliserai la suivante : je peux croire que « p » est vrai même si je ne comprends pas la signification de « p » ; mais dans un tel cas je ne peux pas croire que *p* du fait de croire que « p » est vrai.

Imaginons par exemple que ma connaissance de la physique soit quasi nulle et que, pénétrant dans la salle de cours de Serge Haroche au Collège de France, je l'entende énoncer, *en n'ayant à peu près aucune idée de ce dont il s'agit* : « Or, vous le savez, la valeur du nombre quantique de spin des fermions est demi-entière ». Imaginons par ailleurs que je sache que Serge Haroche est,

[10]Plus généralement, ce n'est pas croire qu'est vrai quelque chose dont le fait que *p* soit ou non le cas détermine la vérité ou la fausseté. L'argument que je vais utiliser pour établir ce point est indépendant de ceux que Donald Davidson, Peter Hacker ou Bede Rundle par exemple ont formulé contre l'idée que croire quelque chose consiste à entretenir une attitude d'un certain type vis-à-vis d'une *proposition*. Ces arguments ne sont pas nécessaires pour établir le point plus limité qui m'importe ici et qui peut l'être, me semble-t-il, de façon plus immédiatement convaincante qu'en y ayant recours. L'argument de Peter Hacker est le suivant : « puisque ce que vous craignez ou suspectez peut être ce que je crois quand je crois que *p,* et puisque craindre ou suspecter que p ce n'est pas craindre ou suspecter la proposition que p, ce que je crois quand je crois que p ne peut pas être une proposition ». (Hacker, 2004, 186) Peut-être voudra-t-on objecter à cet argument que ce que je crains, espère ou suspecte ce n'est bien évidemment pas une proposition (sauf à soutenir que les propositions sont des faits), mais plutôt qu'une proposition soit vraie. Et l'on pourra alors ajouter que, de la même manière, ce que je crois est qu'une proposition est vraie. A cette objection, il est possible de répondre, en suivant la logique du raisonnement de Hacker, que dire que j'espère que la proposition que *p* soit vraie, ce n'est (sinon de façon alambiquée et potentiellement trompeuse) rien dire d'autre que : « j'espère que *p* soit le cas ». Autrement dit, ce que l'on espère est un fait, mais pas un fait portant sur une proposition (comme le fait que la proposition que *p* soit vraie). L'argumentation de Bede Rundle peut, quant à elle, être reconstruite (et prolongée) de la façon suivante : tout comme il est absurde de dire que l'on rêve un rêve, ou que l'on pense une pensée, il est absurde de dire que l'on croit une croyance ; or si par « proposition » on entend le *contenu* d'une croyance, dire que croire consiste à croire une proposition serait comme dire que l'on croit une croyance. Et si l'on ne veut pas dire cela mais plutôt que 1) croire consiste à adopter une certaine attitude vis-à-vis de quelque chose qui préexiste à la croyance et que 2) ce quelque chose est une proposition, alors on conçoit, qu'on le veuille ou non, la proposition comme un énoncé. Or croire que *p* ne saurait être identique au fait de croire quelque chose à propos d'un énoncé. (Cf. Rundle, 1997, 53)

sur les points qu'il aborde dans ses cours, extrêmement fiable. Je croirai (et même saurai) alors qu'est vrai l'énoncé : « La valeur du nombre quantique de spin des fermions est demi-entière ». Puis-je cependant croire (ou savoir) que la valeur du nombre quantique de spin des fermions est demi-entière ? Imaginons cette fois, afin de trancher cette question, que Serge Haroche ait (dans un accès de fantaisie) affirmé cela non pas en français mais en l'écrivant au tableau dans une langue dont je ne parle pas un traitre mot, l'islandais par exemple (« Or, vous le savez : [*énoncé en islandais écrit au tableau*] »). Dans ce cas, il serait manifestement absurde de soutenir qu'après avoir assisté à ce moment du cours de Serge Haroche je crois (ou sais) que la valeur du nombre quantique de spin des fermions est demi-entière. Or il n'y a pas de raison de supposer qu'il en aille différemment lorsque l'énoncé est en français mais que je n'ai aucune idée de ce dont il est question – le fait de saisir la structure grammaticale de l'énoncé en français mais pas celle de l'énoncé en islandais ne saurait suffire à faire une différence sur ce point. Je peux donc bien croire que je dis quelque chose de vrai en répétant l'énoncé français ou islandais, et l'adopter ou le mémoriser pour cette raison, sans être pour autant en mesure de croire ce qu'il énonce. Autrement dit, je peux bien croire que « p » est vrai mais ne pas croire que *p*.

Il est donc parfaitement possible, pour en revenir au cas initial de la Ford, que je puisse croire que l'énoncé « Quelqu'un dans l'entreprise possède une Ford » est vrai, croire que cela suit du fait qu'il soit vrai que Jean possède une Ford, mais que je ne puisse pas avoir la croyance additionnelle que quelqu'un possède une Ford. Pour dire les choses un peu différemment, si, après avoir écouté Jean, on me demande ce que je crois et que je réponds : « Eh bien, que Jean possède une Ford… et donc que quelqu'un dans l'entreprise possède une Ford, ou que quelqu'un possède quelque chose », la croyance que j'exprime après mon moment de réflexion est que ce que je crois quand je crois que Jean possède une Ford peut être décrit de façon véridique de cette manière – ce que je n'avais pas forcément réalisé avant ce moment. Autrement dit, « … et donc que quelqu'un dans l'entreprise possède une Ford, ou que quelqu'un possède quelque chose » n'est pas l'expression de deux croyances additionnelles qui viendraient s'ajouter à ma croyance que Jean possède une Ford.[11]

[11] Je mentionne ici une autre objection qui pourrait être opposée à l'argument que j'ai défendu : « Il n'importe nullement que les propositions dont il est question dans les *Gettier cases* soit ou non crues ; ce qui importe est que ces propositions soient indubitablement vraies et justifiées et qu'elles ne constituent pas, pour autant, des connaissances ». À cette objection, on peut répondre que s'il n'est pas nécessaire à une proposition d'être l'objet d'une attitude doxastique pour

En soutenant qu'un processus d'indéfinition ou d'indétermination d'une croyance du genre de celui que l'on rencontre dans les cas de Gettier ne peut en réalité suffire à conduire à la formation de croyances additionnelles plus générales[12], il ne s'agit bien évidemment pas de soutenir qu'aucune croyance portant sur un ou plusieurs particuliers déterminés ne peut jamais y conduire. N'importe quelle croyance peut être généralisée, mais de façon *définie* : je pourrais parfaitement former, en plus de la croyance que Jean possède une Ford, la croyance que les employés de l'entreprise possèdent des Ford. Mais pour former cette croyance plus générale (et, éventuellement, qu'elle soit justifiée ou constitue une connaissance), les éléments de preuve m'ayant conduit à croire que Jean possède une Ford ne saurait bien entendu suffire, contrairement à ce que la généralisation par indéfinition était censée permettre. Il faudra que je dispose d'autres éléments de preuve pour croire une telle chose, par exemple que j'aie entendu un des collègues de Jean dire : « Nous conduisons tous des Ford dans cette entreprise », ou que je n'aie vu que des Ford dans la zone du parking strictement réservée aux employés.

3. Le téléologisme doxastique au secours de l'argumentation de Gettier

Pour pouvoir accepter l'idée qu'il n'y a pas de volontarisme doxastique tout en rejetant l'idée qu'être incapable de croire *autre chose* que ce qu'apparaissent

posséder la propriété épistémique d'être justifiée (on peut estimer par exemple qu'elle est justifiée par d'autres propositions qui rendent sa vérité probable), il n'en va pas de même dans le cas de la propriété épistémique de constituer une connaissance : le fait qu'il soit su que *p* implique constitutivement un sujet de cette connaissance. Or, classiquement, savoir que *p* implique que *p* soit cru (et cru d'une certaine manière). Il n'y a donc pas de sens à dire qu'une proposition peut constituer ou non une connaissance même si elle ne peut pas être crue.

[12] Et ceci même si ce processus d'indéfinition ou d'indétermination est plus léger qu'en passant de *Jean* à *quelqu'un*. De la même manière, il n'importe nullement que le processus d'indéfinition de la croyance soit censé partir d'une croyance démonstrative ou d'une croyance descriptive, ni que la croyance initiale possède tel ou tel degré de détermination ou de finesse. Enfin, il n'importe pas non plus que l'on prétende indéfinir le prédicat plutôt que le sujet de la croyance : dans le *Gettier case* en question, je ne peux pas davantage former, en plus de la croyance que Jean possède une Ford, la croyance que *quelqu'un* possède une Ford que je ne peux former la croyance que Jean possède *une voiture* ou *quelque chose*. Cette dernière variante du cas de la Ford est structurellement identique à celle-ci, imaginée par Pascal Engel (Engel, 2007) : « Je déclare faussement que j'ai fait des conférences en Algérie ; vous croyez alors, de façon justifiée, que j'ai fait des conférences en Afrique du Nord ; il s'avère que c'est vrai parce que j'ai fait des conférences en Tunisie. Mais pour autant vous ne savez pas que j'ai fait des conférences en Afrique du Nord, même si c'est vrai et justifié. »

établir les données disponibles implique d'être incapable de croire quelque chose *de plus faible* que ce qu'elles apparaissent établir, il me semble que la seule option philosophique disponible soit d'endosser une conception téléologique de la croyance.

Pourquoi ? Parce que si l'on soutient que la croyance *vise* la vérité, la connaissance ou la justification, le fait que l'on ne puisse pas croire à volonté peut se comprendre et s'expliquer de la façon suivante : on ne peut pas croire quelque chose que les données disponibles ne nous apparaissent pas permettre d'établir, quelque chose *de plus fort* que ce qu'elles nous semblent supporter, parce que dans ce cas on croirait, à nos yeux mêmes, de façon injustifiée, hasardeuse ou peu fiable, ce qui n'est pas possible si la croyance vise la vérité, la connaissance ou la justification. On pourrait cependant croire quelque chose *de plus faible* que ce que ces données nous apparaissent établir, car il s'agirait en ce cas d'un bon moyen de croire de façon justifiée, d'une façon qui a de fortes chances d'être vraie, ou de constituer une connaissance. Ainsi par exemple, si je n'arrive pas à déterminer, d'où je suis placé, si l'animal blanc que je vois se déplacer sur le flanc de la montagne est un mouton ou un chien de berger, je ne peux pas me mettre à croire *davantage* que ce que les données à ma disposition me semblent établir (qu'il s'agit d'*un animal blanc qui peut être soit un mouton, soit un chien*) – par exemple qu'il s'agit d'un mouton. Mais je peux me mettre à croire *moins* que ce qu'elles me semblent établir – par exemple qu'il s'agit d'un *animal*. En croyant *moins* je croirais même *mieux* : en croyant qu'il s'agit d'un animal, j'ai moins de chance d'être dans l'erreur qu'en croyant qu'il s'agit d'un animal blanc qui peut être soit un mouton, soit un chien ; et si je sais qu'il y a *un tel animal* sur le flanc de la colline, je sais *a fortiori* qu'il y a *un animal*.

Le partisan d'une conception téléologique de la croyance se trouve donc en mesure de rejeter à la fois l'idée que l'on puisse croire à volonté et l'idée que l'on puisse croire moins, ou quelque chose de plus faible, que ce que les données disponibles apparaissent établir. Le problème est que cette conception n'a rien d'indiscutable et se trouve au contraire exposée à un certain nombre de difficultés qui sont, sinon insurmontables, tout du moins extrêmement sérieuses.

J'indiquerai simplement ici le cœur de la critique qu'en fait David Owens (2003), telle que Asbjorn Steglich-Petersen le synthétise :

> Si notre adhésion aux normes épistémiques s'explique par un *but* que nous entretenons, on peut s'attendre à ce que, au moins dans certaines occasions, quand il s'agit de décider de la façon d'adhérer

> aux normes générées par ce but, nous mettions en balance le but de la croyance avec d'autres buts. Mais la délibération épistémique ne fonctionne pas comme ainsi. [...] Quand nous nous demandons quoi croire, nous nous concentrons exclusivement sur des considérations épistémiques portant sur la vérité de la proposition que l'on pourrait croire. Par contraste, la délibération relative à des activités motivées par des buts implique typiquement la mise en balance de nos buts. [...] Si une de mes délibérations n'a pas pour caractéristique de mettre des buts en balance, il est difficile de penser réellement qu'il s'agit d'une délibération relative à la manière d'accomplir mes buts. Les buts sont *essentiellement*, semble-t-il, des choses qui peuvent entrer en conflit et être mises en balance les unes avec les autres. [...] Or le soi-disant but de la croyance semble précisément ne pas avoir cette propriété : quand je me demande si la Terre est plate, seules comptent des considérations portant sur la *vérité* de la proposition concernée, *à l'exclusion* d'autres considérations. Il n'y a pas de sens à mettre des buts en balance quand il s'agit de décider quoi croire. Ainsi, bien qu'il puisse y avoir un sens métaphorique ou trivial auquel les croyances 'visent la vérité', à savoir qu'une croyance n'est correcte que si elle est vraie, aucun sens réellement explicatif ne peut être attribué à cette affirmation. (Steglich-Petersen, 2009, 396-7)

Mon propos dans cet article n'est cependant pas de revenir sur les difficultés de la conception téléologique de la croyance. Il est qu'il faut être prêt à endosser cette conception pour pouvoir accepter l'idée qu'il n'est pas possible de croire *autre chose* à t que ce que les données disponibles semblent établir à t tout en rejetant l'idée qu'il est impossible de croire à t *quelque chose de plus faible* que ce qu'elles semblent établir à t. Autrement dit, à moins que le téléologisme de la croyance soit correct, il faudrait pouvoir croire à volonté pour que je parvienne à former la croyance que *quelqu'un* possède une Ford quand je n'ai d'autres éléments de preuve en faveur de cela que ceux qui m'ont conduit à croire que *Jean* en possède une.

4. Examen de cas de Gettier classiques

Il peut être éclairant à présent de procéder à l'analyse détaillée des autres types classiques de cas de Gettier en s'appuyant sur l'argument que j'ai défendu. Je commencerai par les deux cas séminaux proposés par Gettier dans « Is Jus-

tified True Belief Knowledge ? ». Le premier peut être (re)décrit de la façon suivante : je crois de façon justifiée que Jean est l'employé préféré du directeur de l'entreprise ; je crois de façon également justifiée que Jean a quatre filles ; j'en tire la croyance justifiée que la personne qui est l'employé préféré du directeur de l'entreprise a quatre filles ; mais en réalité ce n'est pas Jean qui est l'employé préféré du directeur de l'entreprise mais Martin, qui s'avère à mon insu avoir également quatre filles. Ma croyance justifiée que la personne qui est l'employé préféré du directeur de l'entreprise a quatre filles se révèle donc être vraie. Néanmoins, à l'évidence, elle ne constitue pas une connaissance.

Ce cas peut être abrégé ainsi :

1) J'ai la croyance justifiée (CJ) que S est P&Q.

2) J'en tire la CJ que quelqu'un est P&Q.

3) En réalité S n'est pas P, mais il s'avère que S* est P&Q.

4) Ma CJ que quelqu'un est P&Q est donc une croyance vraie et justifiée (CVJ), mais elle n'est pas une connaissance (K).

Ainsi conçu, ce cas de Gettier tombe directement sous le coup de l'argument défendu dans les sections précédentes : il n'est pas possible de former en plus de la croyance (justifiée) que S est P&Q la croyance (justifiée) que quelqu'un est P&Q.

Peut-être objectera-t-on qu'il ne s'agit pas dans ce cas de Gettier de tirer, de la croyance que *S* est P&Q, la croyance que *quelqu'un* est P&Q, mais d'en tirer la croyance que *celui*, quel qu'il soit, qui est P est (également) Q. De sorte qu'il ne serait pas *évident* qu'on en tire une croyance plus indéfinie ou moins déterminée. Cette objection doit cependant être rejetée : le fait que la propriété P soit singularisante ne suffit pas à faire que la croyance que celui qui est P est également Q porte sur l'individu, quel qu'il soit, qui est le seul à posséder cette propriété, *au sens où* ma croyance que Jean est l'employé préféré du directeur de l'entreprise porte sur Jean. Dans le cas imaginé par Gettier, la croyance que celui, quel qu'il soit, qui est P est également Q n'est pas en ce sens davantage susceptible d'être formée que la croyance que quelqu'un est P&Q.

Le second cas présenté dans l'article de 1963 peut être présenté (et abrégé[13]) de la façon suivante :

1) J'ai la CJ que S est P.

[13]Si l'on préfère que les cas soient rendus concrets, on pourra substituer à « S est P » « Jean possède une Ford » et à « S* est Q » « Gianmario est à Rome avec Katia ».

2) J'en tire la CJ que S est P ou S* est Q. (J'ai choisi la proposition que S* est Q au hasard, mais je sais que de la vérité de *p* (S est P) on peut logiquement inférer la vérité de *p ou q*).

3) En réalité S n'est pas P, mais il s'avère que S* est Q.

4) Ma CJ que S est P ou S* est Q est donc une CVJ, mais ce n'est pas une K.

Ce cas n'est cependant pas concluant non plus. Il ne le serait que s'il était réellement possible de former la croyance que S est P ou S* est Q ; mais elle ne peut pas l'être. La raison pour laquelle elle ne le peut pas ne consiste pas simplement à dire, comme on le fait parfois, qu'il n'y a rien de tel que *la* croyance que S est P ou S* est Q mais, en réalité, *deux* croyances dont l'une est fausse et justifiée (S est P) et l'autre vraie et non justifiée (S* est Q). La raison pour laquelle elle ne peut pas être formée est, plus fondamentalement, que je ne peux pas croire que S* est Q parce que je ne peux pas croire à volonté n'importe quoi. Je ne peux pas croire n'importe quel fait susceptible d'être énoncé en remplaçant « S* » et « Q » dans « S* est Q » par n'importe quel nom d'individu et n'importe quel prédicat. Ce que je crois réellement (et de façon justifiée) quand Gettier prétend que je crois que S est P ou S* est Q est donc uniquement que S est P et que la vérité d'un énoncé (« S est P ») en implique celle d'un autre (« S est P ou S* est Q ») parce que de la vérité de « p » on peut déduire la vérité de « p ou q ».

Qu'en est-il à présent de la variante du *Ford case* imaginée par Keith Lehrer – variante dite du *Penseur Intelligent* ? Robert Fogelin la résume ainsi : « Etant prudent et intelligent, [je] me dis que quelqu'un d'autre dans l'entreprise pourrait bien avoir une Ford ; de sorte que pour augmenter [mes] chances d'avoir raison, [je] me replie consciemment sur une affirmation plus faible, celle que quelqu'un dans l'entreprise possède une Ford. » (Fogelin, 1994, p. 24).

Est-il possible de former, *de cette manière*, la croyance que *quelqu'un* dans l'entreprise possède une Ford ? S'il s'agissait uniquement, comme dans le *Ford case* original, d'affaiblir ou d'indéfinir la croyance que Jean possède une Ford que les données disponibles m'ont conduit à former, il faudrait à nouveau répondre négativement. Mais il ne s'agit pas uniquement de cela car une nouvelle donnée se trouve introduite dans ce cas : je réalise que quelqu'un d'autre que Jean pourrait tout aussi bien posséder une Ford. Or cette nouvelle donnée m'apparaît établir qu'il se pourrait que *quelqu'un* dans l'entreprise possède une Ford *autrement qu'en m'apparaissant établir que Jean possède une Ford* ;

autrement dit, cette nouvelle donnée, qui m'est subitement venue à l'esprit, est indépendante de celles qui me sont apparues établir que Jean possède une Ford.

Mais le fait qu'il soit possible, dans le cas du *Penseur Intelligent*, de former la croyance qu'il se pourrait que quelqu'un dans l'entreprise possède une Ford ne prouve nullement qu'une croyance vraie et justifiée n'est pas nécessairement une connaissance. Il n'est pas évident en effet que cette croyance ne constitue pas une connaissance. Le statut épistémique de cette croyance ne saurait en effet varier selon que ce soit Jean ou quelqu'un d'autre dans l'entreprise qui possède une Ford. Par ailleurs, la croyance qu'il est possible de former (de façon justifiée) dans ce cas est la croyance qu'*il se pourrait* que quelqu'un dans l'entreprise possède une Ford, non la croyance que quelqu'un dans l'entreprise en possède une. Contrairement à ce que la description du cas du *Penseur Intelligent* suggère implicitement, la croyance que *quelqu'un* dans l'entreprise *pourrait posséder* une Ford ne peut pas se combiner à la croyance que *Jean possède* une Ford pour former la croyance que *quelqu'un* dans l'entreprise *possède* une Ford[14].

Passons à présent au *Sheep Case* imaginé par Roderick Chisholm : « un homme estime qu'il y a un mouton dans un champ, et ceci dans des conditions telles [qu'il soit justifié à le croire]. Cet homme a cependant pris un chien pour un mouton ; ainsi ce qu'il voit n'est nullement un mouton. Cependant, à son insu, un mouton se trouve dans une autre partie du champ. » (Chisholm, 1977, 105) Dans un tel cas je forme la croyance, immédiatement justifiée par mon expérience perceptuelle[15], qu'il y a un mouton dans le champ, mais je ne

[14] Ces deux croyances sont néanmoins parfaitement compatibles.

[15] Le *sheep case* est, entre autres choses, conçu pour permettre aux *Gettier cases* d'échapper à l'objection selon laquelle les croyances censées être à la fois vraies et justifiées et ne pas être des connaissances ne sont, en réalité, nullement justifiées parce qu'elles sont inférées de prémisses fausses. Dans le *Sheep Case*, l'*evidence* perceptuelle est supposée justifier de façon immédiate (ou *prima facie*) la croyance qu'il y a un mouton dans le champ. S'est néanmoins inévitablement posé, dans la littérature épistémologique, la question de savoir si cette croyance perceptuelle ne repose pas également sur une prémisse fausse – la prémisse que ce que je crois être un mouton est effectivement un mouton. La position que je défends est, on l'a compris, qu'il n'est nul besoin de trancher cette question pour montrer que les *Gettier cases* n'établissent pas ce qu'ils sont supposés établir : mon argument contre les *Gettier cases* ne repose pas sur une critique de la conception de la justification implicitement défendue par Gettier, ni d'ailleurs sur aucune théorie particulière de la justification (énonçant, par exemple, que pour être justifiée, une croyance ne doit pas être inférée d'une prémisse fausse ; ou reposer sur une inférence incorrecte ; ou ne pas être effectivement établie par les données sur lesquelles elle repose, etc. Ce dont il suivrait, selon l'une ou l'autre de ces théories, que les croyances gettierisées ne s'avèrent pas justifiées.) Par ailleurs, ne change rien à l'affaire le fait d'ajouter à la description du *Sheep Case* la clause qu'un mouton soit par ailleurs

sais manifestement pas qu'il y a un mouton dans le champ.

Ce contre-exemple ne permet cependant pas, lui non plus, de montrer qu'une croyance vraie et justifiée n'est pas nécessairement une connaissance, car la croyance perceptuelle que je forme en ce cas est tout simplement fausse : elle porte en effet sur ce que j'ai pris pour un mouton et qui n'en était pas un, mais un chien. On pourrait objecter que la croyance censée faire apparaître que toute croyance vraie et justifiée n'est pas une connaissance n'est pas, bien entendu, la croyance démonstrative erronée portant sur ce que j'ai pris pour un mouton mais la croyance générale et indéfinie qu'il y *un* mouton dans le champ. Le problème est alors, encore une fois, qu'il n'est pas possible de former cette croyance (ou, par exemple, la croyance qu'il y a *un animal* dans le champ) lorsque les seules données dont je dispose en sa faveur m'apparaissent établir que *ceci*, dans le champ, est un mouton.

Imaginons enfin un autre type de cas de Gettier, où une croyance indéfinie se trouve cette fois bel et bien formée, semble être vraie et justifiée, mais ne pas constituer une connaissance : supposons que quelques accointances, à l'humour passablement potache, veuillent me faire croire qu'il y a des sangliers dans la forêt qui se trouve à côté de chez moi et qu'ils y créent en conséquence de fausses traces de sangliers. Je crois alors, en percevant ces fausses traces, qu'il y a des sangliers dans cette forêt. Or il s'avère y avoir des sangliers à l'autre bout de la forêt, mais je ne les ai jamais vus et n'en ai jamais perçu aucune trace[16]. Dans cette situation, il semble donc que je croie de façon vraie et justifiée qu'il y a des sangliers dans la forêt mais, intuitivement, je ne le sais pas.

Cette croyance est-elle cependant réellement vraie ? Mais amis ne jugeront-ils pas qu'ils sont parvenus à me tromper et que j'ai commis une erreur en formant cette croyance – c'est-à-dire en considérant la présence des sangliers dans la forêt comme établie par les éléments de preuve à ma disposition ? Et s'ils jugeront les choses ainsi, n'est-ce pas parce qu'il s'agit d'une croyance portant sur *ce qui a causé ces traces*[17] ? Peut-être voudra objecter, une fois en-

vu dans le champ mais pas reconnu ou identifié comme tel – ceci pour que, en plus du fait qu'une croyance vraie et justifiée soit formée, un mouton soit effectivement vu

[16]Si l'on préfère continuer à raisonner en s'appuyant sur des cas où il est question de Ford et de locaux d'entreprise, imaginons la situation suivante, structurellement identique : je trouve des clés de voiture Ford, des factures à entête d'un garage Ford, des *Ford Magazine*, etc., un peu partout dans les locaux de l'entreprise. Je forme alors la croyance que quelqu'un dans l'entreprise possède une Ford. Mais il s'avère que seul en possède une l'employé travaillant aux archives, au sous-sol, que ce dernier ne monte jamais dans les étages supérieurs de l'entreprise et que ceux qui travaillent à ces étages n'ont jamais eu affaire à lui et ne savent rien de lui.

[17]Je note au passage que la situation suivante ne permet pas d'établir la conclusion visée par

core, que ce n'est pas cette croyance fausse qui est supposée montrer qu'une croyance vraie et justifiée n'est pas nécessairement une connaissance, mais la croyance affaiblie – tirée par indéfinition de cette croyance fausse mais qui ne porte pas sur ce sur quoi porte en particulier cette dernière – qu'il y a des sangliers dans la forêt[18]. Mais, une fois encore, la réponse sera que cette croyance affaiblie ne peut tout simplement pas être formée, ou, de façon équivalente, n'est qu'une manière de caractériser ma croyance fausse[19].

5. Distinguer les cas de Gettier entre eux

Les cas de Gettier ne sont donc nullement en mesure de prouver que toute croyance vraie et justifiée n'est pas une connaissance. Seuls sont en mesure d'y parvenir, parmi les cas classiques que l'on catégorise parfois comme cas de Gettier, ceux qui n'ont pas pour principe l'affaiblissement par indéfinition d'une croyance fausse (et justifiée), comme le cas des *Façades de Fermes* ou de la *Mort du Dictateur*.[20]

Gettier : imaginons, avec un autre relecteur anonyme, que je croie que quelqu'un dans l'entreprise possède une Ford parce que 1) une personne extrêmement fiable m'a dit à propos d'un employé de l'entreprise qu'il possède une Ford mais que 2) je ne me souviens plus à présent de qui il s'agit. Dans ce cas, je ne sais pas que quelqu'un dans l'entreprise possède une Ford même si cette croyance est justifiée et même si, d'après ce relecteur, elle est vraie. Ce cas ne permet cependant pas de prouver, contrairement à ce qu'est enclin à croire ce relecteur, qu'une croyance vraie et justifiée n'est pas nécessairement une connaissance. Ceci parce que la croyance en question est, en réalité, tout simplement fausse : elle porte sur l'individu dont on m'a donné le nom que j'ai oublié, sur la personne dont on m'a parlé mais que je ne saurais plus identifier – autrement dit, sur Jean. Et Jean ne possède pas de Ford.

[18] Il est bien entendu regrettable, pour la clarté du propos, que ces deux croyances différentes, la fausse et l'affaiblie, soient caractérisées de la même manière, à savoir comme « croyance qu'il y a des sangliers dans la forêt ».

[19] Et si l'on objecte encore : « Très bien, mais imaginons plutôt que l'on cherche et que l'on réussisse à vous tromper non pas avec *des traces* ou *des signes de la présence* de sangliers, mais avec des hologrammes de sangliers apparaissant devant vous. Que se passe-t-il ? », on répondra : « Eh bien, il s'agit alors d'un type de *Sheep Case*, dont l'analyse a été faite plus haut. »

[20] Si l'on considère que l'utilisation de ce principe constitue la caractéristique essentielle des *Gettier cases*, il peut cependant apparaître assez immotivé de catégoriser ainsi ces deux cas. Rien n'interdit bien évidemment de décider de qualifier toute situation dans laquelle une croyance est vraie et justifiée sans être une connaissance de *Gettier case*. Mais pourquoi faudrait-il classifier toute croyance vraie justifiée qui n'est pas une connaissance comme « gettierisée » ? Si l'explication du défaut de connaissance dans un cas (celui des façades de fermes par exemple) n'a quasiment rien en commun avec l'explication du défaut de connaissance dans un autre (celui de la Ford par exemple), et si ces cas sont construits d'après des principes différents, on a alors affaire à une simple stipulation verbale dépourvue de la moindre utilité et de la moindre raison d'être épistémologique. Si l'on choisit de soutenir que l'usage du principe d'indéfinition d'une

Le premier cas, imaginé par Carl Ginet, est le suivant :

> A son insu, Barney circule dans une région remplie de décors de façades de fermes. Il décide de faire quelques pas, arrête son véhicule en face de la seule véritable ferme de la région et forme la croyance qu'une ferme se trouve en face de lui. Même si elle est vraie et justifiée, cette croyance ne semble pas être une connaissance. (Goldman, 1976, 781)

Le second, imaginé par Gilbert Harman est, tel que le décrit Robert Nozick, le suivant :

> Le dictateur d'un pays est assassiné ; dans une première édition, tous les journaux du pays impriment le véritable déroulement des événements, mais plus tard tous ces journaux ainsi que les autres media démentent faussement cette histoire. Toutes ceux qui sont exposés à ces démentis y croient (ou ne savent que croire et suspendent leur jugement). Seule une personne dans tout le pays n'y est pas exposée et continue de croire la vérité. [...] Nous ne serions pas enclins à dire que cette personne connaît la vérité. Car si elle avait lu ou entendu les démentis, elle aussi les aurait crus, comme tout le monde. (Nozick, 1981, 177)

On notera cependant qu'il n'est pas *aussi évident* dans ces deux cas que dans les autres cas de Gettier que les croyances qu'il n'y ait pas connaissance. Autrement dit, ces deux cas ne semblent pas en mesure de *démontrer* la conclusion épistémologique que les autres cas de Gettier semblaient être capables d'établir *de cette manière*.

Est-ce à dire que l'on ne saurait espérer montrer qu'une croyance peut être vraie et justifiée sans être une connaissance sans recourir au principe d'affaiblissement par indéfinition d'une croyance fausse ? Certainement pas. Le cas, imaginé par Russell, de l'horloge arrêtée, où les aiguilles d'une horloge se sont par chance arrêtées pendant la nuit sur l'heure exacte à laquelle je la consulte pour la première fois le lendemain, semble clairement être une instance de croyance vraie et justifiée qui n'est pas une connaissance. De façon similaire, imaginons que je croie que Jean sera en possession d'une Ford Mustang ce soir à 18h sur la base de son témoignage bien étayé (l'honnêteté

croyance fausse est caractéristique des *Gettier cases*, on doit alors soutenir que ce n'est que de façon passablement relâchée, et finalement assez trompeuse, que les cas des façades de ferme ou de la mort du dictateur sont des *Gettier cases*.

et la sincérité de Jean n'ont jamais été prises en défaut et je remarque sur son bureau le devis d'un concessionnaire ainsi qu'un chèque de banque d'un montant correspondant au prix du modèle). Imaginons encore que Jean m'ait menti mais qu'il sera néanmoins bel et bien en possession d'une Ford Mustang ce soir à 18h parce que, à mon insu comme au sien, il vient d'être tiré au sort dans l'annuaire et que le prix qui lui a été attribué, qu'il ira retirer ce soir à 18h, est une Ford Mustang. Dans cette situation, il apparaît évident que *je ne sais pas* que Jean sera en possession d'une Ford Mustang à 18h, même si je le crois de façon véridique et justifiée.

En s'appuyant sur de tels cas, on pourrait alors vouloir dégager un principe épistémologique permettant à la fois de rendre compte de nos intuitions de connaissance et de générer à volonté des cas de croyances vraies et justifiées qui ne sont pas des connaissances. Ce principe pourrait être le suivant : lorsque, par malchance[21], le fait qu'une croyance soit vraie ne s'explique pas par les données qui expliquent qu'on l'ait formée, elle n'est pas une connaissance[22]. Construire des cas d'après un tel principe et en construire d'après le principe d'indéfinition d'une croyance fausse ressortit cependant de logiques et d'ambitions différentes. Tandis que, d'un côté, on est censé disposer, grâce à Gettier, d'une sorte de formule magique permettant de générer automatiquement et indiscutablement des instances de croyances vraies et justifiées qui ne sont pas des connaissances – le fait d'imaginer un cas concret ne consistant, pour l'essentiel, qu'à imaginer deux noms propres et deux prédicats –, on a, de l'autre, dégagé un principe qui n'est ni plus ni moins qu'une thèse épistémologique vague et contestable à propos de la nature de la connaissance.[23] [24]

[21] Ce « par malchance » peut être compris soit en un sens internaliste, soit un sens externaliste – autrement dit, soit comme indiquant qu'il est épistémiquement raisonnable de former cette croyance étant les données disponibles, soit comme indiquant qu'elle est formée d'une façon qui est généralement fiable ou sûre.

[22] Je note au passage que le principe apparemment similaire d'après lequel la croyance vraie et justifiée que p n'est pas une connaissance lorsque les données ayant conduit à la former n'ont, par malchance, rien à voir avec le fait que p soit le cas ne permettrait pas de rendre compte des cas dits de « chance épistémique environnementale », comme celui des façades de fermes ou de la mort du dictateur mort.

[23] Il semble de plus impossible que l'on puisse, pour chaque croyance, déterminer de façon indiscutable si elle instancie ou non un tel principe : dans le cas des façades de fermes par exemple, la vérité de ma croyance qu'une ferme se trouve en face de moi s'explique-t-elle par le fait que ce soit une ferme que j'aie en face de moi ?

[24] Je remarque au passage – au cas où l'on se demanderait à quoi bon prendre la peine de rejeter l'argumentation de Gettier si l'on admet par ailleurs ce qu'elle vise à établir (à savoir qu'il y a des croyances vraies justifiées qui ne sont pas connaissances) – que les arguments ont une

Ce dernier point peut d'ailleurs, me semble-t-il, contribuer à expliquer l'extraordinaire postérité de l'article de Gettier. Son apparente neutralité épistémologique et son apparente capacité à démontrer qu'une croyance vraie et justifiée n'est pas nécessairement une connaissance ne sont pas en effet des caractéristiques distinctives : le cas russellien de l'horloge arrêtée possédait déjà, un demi-siècle plus tôt, ces deux caractéristiques. Ce qui distingue spécifiquement l'argumentation de Gettier est le genre de formule magique qu'elle semble mettre en les mains des philosophes. Cependant, si ce que j'ai avancé est correct, elle ne permet pas d'établir la conclusion souhaitée, car les seules croyances qui se trouvent véritablement formées dans les situations qu'elle nous conduit à imaginer sont des croyances fausses, éventuellement justifiées, qui, de façon assez peu remarquable, ne sont pas des connaissances.

6. References

CHISHOLM, R. (1977), *Theory of Knowledge*. Englewood Cliffs, N.J., Prentice Hall.

ENGEL, P. (2007), *Va savoir !*, Paris, Hermann.

FOGELIN, R. (1994), *Pyrrhonian Reflections on Knowledge and Justification*, Oxford, Oxford UP.

GETTIER, E. (1963), « Is Justified True Belief Knowledge? », *Analysis,* 23, 121–123.

GOLDMAN, A. (1967), « A Causal Theory of Knowing », *Journal of Philosophy*, 64, 357–72.

GOLDMAN, A. (1976), « Discrimination and Perceptual Knowledge », *Journal of Philosophy*, 73, 771-91

HACKER, P. (2004), « Of the ontology of belief », dans Siebel M. & Textor M. (éds), *Semantik und Ontologie*, Frankfurt, Ontos Verlag, 185–222.

LEHRER, K. (1965), « Knowledge, Truth and Evidence », *Analysis*, 25, 168–75.

LEHRER, K. & Paxson, T. (1969), « Knowledge: Undefeated Justified True Belief », *The Journal of Philosophy*, 66, 225–237.

NOZICK R. (1981), *Philosophical explanations*, Oxford, Clarendon Press.

importance intrinsèque en philosophie et qu'ils charrient avec eux les thèses substantielles dont on peut trouver assez triste que des philosophes se réclamant pourtant d'une orientation analytique continuent parfois de les chercher à proximité du point final d'une section d'article ou de chapitre.

OWENS, D. (2003), « Does Belief Have an Aim? », *Philosophical Studies*, 115, 283–305.

RUNDLE, B. (1997), *Mind in Action*, Oxford, Oxford UP.

SARTWELL, C. (1992), « Why knowledge is merely true belief », *Journal of Philosophy* 89, 4, 167–180.

STEGLICH-PETERSEN, A. (2009), « Weighing the aim of belief », *Philosophical Studies*, 145, 3, 395-405.

20

Knowledge as *De Re* True Belief? *

PAUL ÉGRÉ

Abstract Kratzer proposed a causal analysis of knowledge in which knowledge is defined as a form of *de re* belief of facts. In support of Kratzer's view, I think the *de re/de dicto* distinction can be used to integrally account for the original Gettier cases, but in contrast to Kratzer, I think such an account does not fundamentally require a distinction between facts and true propositions. I then discuss whether this account might give us a reductive analysis of knowledge as *de re* true belief. Like Kratzer, I think it will not, for the distinction seems inadequate to account for Ginet-Goldman cases of causally connected but unreliable belief. Nevertheless, I argue that the *de re* belief analysis allows us to account for a distinction Starmans and Friedman recently introduced between *apparent evidence* and *authentic evidence* in their empirical study of Gettier cases, in a way that questions their claim that a causal disconnect is not operative in the contrasts they found.

*This paper is written in honor of Pascal Engel, on the occasion of his 60th birthday. I have a special intellectual debt to Pascal Engel, whose exciting research seminar on the analysis of knowledge, held at Paris-Sorbonne in 2003, was a tremendous source of inspiration, and gave considerable spin and velocity to my PhD work at the time on the topics of epistemology and vagueness. I was very lucky to be Pascal's teaching assistant in that seminar. Thank you Pascal, and Happy Birthday! Several of the ideas presented here occurred to me back in 2003, when I read Kratzer's paper for the first time, but they got only partly transcribed into my dissertation. I am happy to take this opportunity to present them. I am also indebted to Bryan Renne for discussions and joint work in preparation based on this material, and to Jennifer Nagel for directing my attention to some of the recent psychological literature on Gettier problems and for comments. Neither of them, of course, is responsible for the ideas presented in this paper.

1. Facts vs. True propositions

The view has been proposed by several authors that one way of capturing the difference between knowledge and justified true belief might be in terms of the kind of objects they select for (see Russell 1918, Vendler 1972, Kratzer 2002 among others). The idea is that knowledge selects for *facts*, whereas belief selects for *propositions*. On that view, a fact is more than a true proposition, and to know a fact is more than to merely believe a true proposition. But what does it mean to know a fact, as opposed to believing a true proposition? Kratzer (2002) considers a version of Goldman's 1967 causal account of knowledge, in which knowledge is defined in terms of *de re* belief about some fact:

(1) *S* knows *p* if and only if *S* believes *p de re* of some fact exemplifying *p*.

Kratzer gives an application of this analysis to Russell's 1912 pre-Gettier scenario, in which a man has a true belief, but which falls short of constituting knowledge:

> "If a man believes that the late Prime Minister's name began with a B, he believes what is true, since the late Prime Minister's last name was Sir Henry Campbell Bannerman. But if he believes that Mr. Balfour was the late Prime Minister, he will still believe that the late Prime Minister's last name began with a B, yet this belief though true, would not be thought to constitute knowledge."

On Kratzer's analysis, the belief that the late Prime Minister's name begins with a B is true *qua* propositional belief, but the corresponding fact is the fact that Bannerman's name starts with a B. The man simply fails to have the propositional belief he has as a *de re* belief about that fact.

That analysis is suggestive, but it partly begs the question: what is it to have a *de re* belief of some fact? Kratzer in her paper offers to capture facts in a framework of situation semantics. In this short paper, I propose to examine more closely the idea that knowledge could be treated as *de re* belief of some kind, but in a more conservative framework than Kratzer's, and in particular without committing myself to an ontological distinction between facts and true propositions. Like Kratzer, I believe that the *de dicto/de re* distinction is entirely relevant for the analysis of Russell's puzzle, and of Gettier's puzzles, but I think the distinction can be captured entirely in terms of binding and scope mechanisms. I first present an analysis of both puzzles in epistemic logic, and then ask whether we can identify knowledge with *de re* true belief.

In agreement with Kratzer, we will see that Goldman's fake barn cases do not seem amenable to an analysis in terms of *de re* belief alone. On the other hand, I will argue that such an analysis casts light on a distinction recently proposed by Starmans and Friedman (2012) between what they call apparent vs. authentic evidence in their empirical study of laymen's judgments about Gettier scenarios.

2. Russell's puzzle and *de re* belief

Let a denote Bannerman, and b denote Balfour. Let S stand for the complex predicate: "having a name starting with a B", and let P stand for the complex predicate: "being a late Prime Minister". Let us assume that both names a and b are part of our man's mental repertoire, like the corresponding predicates. To make things realistic with regard to Russell's scenario, we may assume that our man (let us call him Ralph) believes the propositions expressed by the four following sentences:

(2) (a) Pb
(b) Sb
(c) Sa
(d) $\forall x \forall y (Px \wedge Py \supset x = y)$

That is, Ralph believes: "Balfour is a late Prime Minister", "Balfour has a name starting with a B", but also "Bannerman has a name starting with a B", and "at most one person is late Prime Minister". We also suppose Ralph's beliefs to be closed under logical consequence.

To deal with definite descriptions, given a predicate Px I will use $P'x$ as an abbreviation for the complex predicate $Px \wedge \forall y (Py \supset x = y)$. Thus, $P'x$ means that x is the only object satisfying P.[1] One way of representing "Ralph believes that the late Prime Minister is Balfour" is by the following *de dicto* ascription, in which the belief operator scopes above the existential operator:

(3) $B_r \exists x (P'x \wedge x = b)$

[1] An alternative would be to use the quantifier $\exists! x$ with its usual meaning, but it would do less service to use it. We could also use a non-Russellian treatment of the definite description altogether. This does not matter for the main point at issue, so long as the scope relations to be discussed in what follows are operative.

Similarly, one can represent "Ralph believes that Balfour's name starts with a B":

(4) $B_r Sb$

From those two ascriptions, it follows by closure that "Ralph believes that the late Prime Minister's name starts with a B", which we can represent as:

(5) $B_r \exists x(P'x \wedge Sx)$

The expression "the late Prime Minister" is read purely *de dicto* here, however. Given 2 c, we also have:

(6) $B_r Sa$

From the assumption that $P'a$ is true in the actual world, it follows by standard assumptions in quantified epistemic logic that:

(7) $\exists x(P'x \wedge B_r Sx)$

That is: "of the actual late Prime Minister [who turns out to be Bannerman], Ralph believes that his name starts with a B".[2] What fails to hold, however, is the following:

(8) $\exists x(P'x \wedge B_r(P'x \wedge Sx))$

That is: "of the late Prime Minister, Ralph believes that he is the late Prime Minister and that his name starts with a B". We can be even more explicit, and see the following to fail:

(9) $\exists x(P'x \wedge Sx \wedge B_r(P'x \wedge Sx))$

That is: "of the late Prime Minister, whose name starts with a B, Ralph believes he is the late Prime Minister and that his name starts with a B". It is easy to construct an epistemic model of Ralph's situation in which all of the previous assumptions are true relative to the actual world, but where 8 and 9 come out as false.

The comparison between 5 and 9 is instructive: basically, the *de dicto* content of 5 is true in the actual world, but Ralph fails to refer it to the right entity.

[2] I am assuming a standard quantified epistemic semantics, with a constant domain assumption, where for example $w \models \exists x B_s Sx$ is true provided some element in the domain of w is such that in every possible world w' compatible with s's belief, that element falls in the extension of S in w'. See Fitting and Mendelsohn (1998) for details.

So Ralph has a correct *de dicto* belief, but an incorrect *de re* belief. Putting them side by side, what we see is a contrast between:

(10) (a) $\exists x(P'x \wedge Sx) \wedge B_r\exists x(P'x \wedge Sx)$ (true)

(b) $\exists x(P'x \wedge Sx \wedge B_r(P'x \wedge Sx))$ (false)

10 a says that Ralph believes a true proposition, whereas 10 b says that Ralph believes a proposition which is true of some actual objects. Under adequate assumptions, 10 b entails 10 a, but not conversely. Prima facie, this looks like a promising way of capturing the difference between an accidentally true belief, referred to the wrong objects in Ralph's belief worlds, and a true belief holding as a fact in virtue of being anchored to the right objects. This difference is obtained without postulating a primitive difference between facts and true propositions, but merely in terms of scope mechanisms.

3. Gettier's puzzles

Two puzzles appear in Gettier's celebrated paper. The first case has a structure very similar to Russell's case, since it also involves a definite description. Smith has "strong evidence that Jones is the man who will get the job, and Jones has ten coins in his pocket" (Gettier 1963). Let us represent what Smith believes by the following sentences:

(11) Gj [Jones will get the job]

(12) Cj [Jones has ten coins in his pocket]

(13) $\forall x\forall y(Gx \wedge Gy \supset x = y)$ [at most one man will get the job]

As it turns out, Smith is the man who will get the job, but also "unknown to Smith, he himself has ten coins in his pocket" (Gettier 1963). Hence the following two sentences hold in the actual world:

(14) Gs

(15) Cs

The following *de dicto* ascription can be made truly about Smith:

(16) $B_s\exists x(G'x \wedge Cx)$ [Smith believes the man who will get the job has ten coins in his pocket]

However, the following *de re* ascription is false given the scenario:

(17) $\exists x(G'x \wedge B_s Cx)$ [of the man who will get the job Smith believes he has ten coins in his pocket]

Parallel to the previous analysis, we can distinguish between the true proposition Smith believes, and the failure of Smith to refer the components of that true proposition to the right objects as follows:

(18) (a) $\exists x(G'x \wedge Cx) \wedge B_s \exists x(G'x \wedge Cx)$ (true)

(b) $\exists x(G'x \wedge Cx \wedge B_s(G'x \wedge Sx))$ (false)

Hence Gettier's first puzzle too can be captured in terms of scope mechanisms. What about Gettier's second case? In the second case, Smith has some reason to believe that Jones owns a Ford, and he furthermore infers about his other friend Brown, "of whose whereabouts he is totally ignorant", that "either Jones owns a Ford, or Brown is in Barcelona". As it turns out, Brown is indeed in Barcelona, but Jones does not own a Ford. So Smith's beliefs can be represented by the following sentences:

(19) (a) Fj [Jones owns a Ford]

(b) $Fj \vee Ab$ [either Jones owns a Ford, or Brown is in Barcelona]

Prima facie, Gettier's second case does not appear to rest on a *de re/de dicto* ambiguity. On the other hand, there is a strong analogy between the occurrence of a disjunction under the scope of the belief operator in 19 a and the occurrence of an existential quantifier in our previous examples, since an existential quantifier is nothing but a generalized disjunction. One way in which we can build a tight parallel with the previous cases is first to observe that the following contrast holds in the scenario. Although the following sentence is true:

(20) $(Fj \vee Ab) \wedge B_s(Fj \vee Ab)$

the following version, in which disjunction takes scope over belief, is false:

(21) $(Fj \wedge B_s Fj) \vee (Ab \wedge B_s Ab)$

Indeed, the first disjunct is false, since Smith falsely believes Jones to own a Ford; and the second disjunct is false too, since Smith does not have the belief that Brown is in Barcelona. The distinction between those two sentences is exactly congruent with a scope distinction. To make a tighter parallel with our

previous representations, one option is to handle disjunction as an existential quantifier ranging over sentences, and to use a truth predicate.[3] The previous two examples are then matched by the following *de dicto* vs. *de re* counterparts:

(22) (a) $\exists p(p \in \{Fj, Ab\} \wedge True(p)) \wedge B_s \exists p(p \in \{Fj, Ab\} \wedge True(p))$
(b) $\exists p(p \in \{Fj, Ab\} \wedge True(p) \wedge B_s(p \in \{Fj, Ab\} \wedge True(p)))$

22 a says that one of the two sentences "Jones owns a Ford", "Brown is in Barcelona" is true, and Smith believes that one of those two sentences is true (this holds in the scenario, since Smith believes "John owns a Ford" to be true). 22 b too says that one of those two sentences is true, and that Smith believes of that sentence that it is indeed one of those two sentences and that it is true. 22 b, however, is false, unlike 22 a, for *Ab* is the only actually true sentence, and of that sentence Smith has not formed the belief that it is true.

4. Ginet-Goldman cases

Both for Russell's problem case, and for Gettier's two cases, we thus have a uniform mechanism which appears to do justice to the intuition of a difference between a merely true belief, and a belief true in virtue of corresponding to a fact. This mechanism, which only involves binding and scope relations, allows us to capture Kratzer's proposed analysis without having to distinguish facts and true propositions. Since the intuition that Russell and Gettier intended to convey is that there is a fundamental difference between justified true belief and knowledge, the question is whether we can equate knowledge with true belief *de re*.

If true belief *de re* is meant to mean a true belief whose truth rests on the right connection to some element in the actual world which makes the belief true, then an argument in favor of this analysis is that we have at least factored in an important externalist component, the one that is arguably missing from so many analyses in which the notion of justified belief is referred only to internal reasons for belief. In all the *de re* logical forms we have proposed, for which the belief ascriptions turned out false, what fails is an adequate connection with the actual *truth-maker* of the proposition believed.

As emphasized by Kratzer in the same paper, however, an analysis of knowledge based exclusively on the notion of true *de re* belief does not ap-

[3] It may seem preferable to quantify over propositions, rather than sentences, and to dispense with a truth predicate. I am skipping details and a more sophisticated treatment here.

pear to be general enough. Besides proper *Gettier* cases, there are also *Ginet-Goldman* cases. In Goldman's 1976 scenario (based on a case proposed by Ginet), Henry is traveling across the land and points to a particular building of which he thinks that it is a barn. It turns out the building is a real barn, but that "unknown to Henry, the district he has just entered is full of papier-mâché facsimiles of barns. These facsimiles look from the road exactly like barns, but are really just façades" (Goldman 1976). The intuition in this case is that Henry is simply lucky: Henry has formed a true belief, and that belief has the right connection to its truth-maker, but it fails a reliability condition. Because of that, Kratzer's analysis of knowledge is as follows:

> "S knows p if and only if
>
> (i) There is a fact f that exemplifies p,
>
> (ii) S believes p *de re* of f, and
>
> (iii) S can rule out relevant possible alternatives of f that do not exemplify p."

An observation Kratzer makes is that knowledge ascriptions violating conditions (i) and (ii) are "clearly false", whereas knowledge ascriptions violating condition (iii) are "vulnerable and context-dependent". This observation implies that knowledge ascriptions involve at least two dimensions of evaluation, a dimension of *causal connection* to the right truth-maker, and a dimension of *counterfactual sensitivity* to the truth-maker (basically: if the circumstances had been different, the belief would have varied accordingly).

What Kratzer suggests, moreover, is that a failure along the dimension of causal connection is in a sense more dramatic than a failure along the dimension of counterfactual sensitivity. Whether this is so or not is a difficult question, about which recent experimental studies on laymen's judgments bring conflicting evidence, and about which experts disagree. For example, Starmans and Friedman (2013: 663) write the following about the Ginet-Goldman cases:

> "So although such 'fake barn' cases are often referred to as Gettier cases, they do not feature the disconnect characteristic of most other Gettier cases. And in fact, philosophers themselves are quite divided on whether to attribute knowledge in these cases (e.g., Lycan, 2006; Sosa, 2007; Turri, 2012)."

Their view of the typology of Gettier cases is in support of Kratzer's intuition in this case. Nagel, San Juan and Mar (2013a), on the other hand, acknowledge

that Ginet-Goldman cases are a new species of Gettier cases, but they also disagree that the distinction is of fundamental importance to the analysis of Gettier cases in the broad sense (2013b).

My own inclination here is to agree with Kratzer and Starmans and Friedman that Gettier cases in the narrow sense (original Gettier cases) and Ginet-Goldman cases really manipulate distinct dimensions of evaluation. On the other hand, we cannot seek much support from Starmans and Friedman's study in favor of Kratzer's judgment, for what they conclude from their findings is that denials of knowledge in Gettier scenarios are not as sensitive to the causal disconnect between a belief and what makes it true as to the kind of evidence that the belief is formed on. Let us consider their distinction, and see where it leaves us.

5. Apparent evidence and Authentic evidence

Starmans and Friedman (2012) conducted a series of experimental studies which appear to challenge the idea that ordinary ascriptions of knowledge are sensitive to the causal disconnect that features in the original Gettier scenarios. However, they found that ascriptions of knowledge were sensitive to whether an agent's belief was formed based on what they call "apparent evidence" vs. "authentic evidence", but they also found that participants ascribed knowledge in cases involving "authentic evidence" for which such ascriptions were prima facie unexpected. The distinction is explained as follows (Starmans and Friedman 2013: 663):

> "Authentic evidence is informative about how the world actually is when the belief is formed, and basing a belief on authentic evidence necessarily makes the belief true when it is formed. Apparent evidence only appears to be informative about how the world actually is, and basing a belief on apparent evidence does not guarantee that the belief is true when it is formed."

An example of a Gettier case based on apparent evidence which they used in their study is one in which a character named Corey believes that he has a quarter coin from 1936 in his piggy bank. His belief is in fact formed on the basis of the wrong perception of a 1938 coin whose date is hard to read. As it turns out, Corey does have an actual quarter coin from 1936 in his piggy bank, but is not aware of it. The corresponding Gettier case involving authentic evidence is one in which Corey believes he has a coin from 1936 in his piggy

bank because he inserted a 1936 coin himself. Unbeknownst to him, there is another quarter from 1936 in his piggy bank. Corey takes a 10 minute nap during which his friend Scott who needs a quarter picks the one Corey has just inserted, leaving the other untouched. In each scenario, the question for which Starmans and Friedman probed participants' judgments was whether, "at the end of the story", Corey "really knows" or "only believes" that there is a coin from 1936 in his piggy bank.

For this pair of scenarios and another such pair, what Starmans and Friedman found was a striking contrast, with an average of 67% of participants attributing knowledge in the Authentic Evidence conditions, vs. only 30% in the Apparent Evidence conditions.[4] The question for us is: can we capture this difference in terms of the *de re/de dicto* distinction used previously, or is there again a genuinely distinct dimension of evaluation at stake? Consider the following sentence:

(23) Corey believes that there is a coin from 1936 in his piggy bank.

Let us consider our two paraphrases, the *de dicto* paraphrase, and the *de re* paraphrase (with "Cx" for "x is a coin from 1936", and "Px" for "x is in Corey's piggy bank"):

(24) (a) $\exists x(Cx \wedge Px) \wedge B_c \exists x(Cx \wedge Px)$

(b) $\exists x(Cx \wedge Px \wedge B_c(Cx \wedge Px))$

The *de dicto* ascription 24 a is true in the two scenarios, the Apparent Evidence as well as the Authentic Evidence scenario. What about 24 b? Whether in the Apparent Evidence case or in the Authentic Evidence case, the only actual coin from 1936 that is left in the piggy bank at the end of the story is not one about which Corey has any *de re* belief. So in both cases, 24 b is false *at the of the story*. However, consider the same belief ascription relative to the beginning of the story. At the beginning of the story, 24 b is this time true in the Authentic evidence case: for Corey has a correct *de re* belief about the coin from 1936 he inserted. But 24 b is false in the Apparent evidence case.

This observation raises two questions. One is whether participants sufficiently paid attention to the adverbial "at the end of the story" (which does

[4]The contrast was even more pronounced in just one of the two pairs, with 76% vs. 14% of knowledge attribution from Authentic to Apparent condition. From Starmans and Friedman's presentation, this appears to be the "Coin" story, but the text talks of "the first story", and I was not sure whether this is indeed so, or whether this refers to the "Yogurt" Story, which is presented first in the appendix to their paper.

not feature in the target sentence, but right before it). Another, assuming that participants did pay attention to the temporal adverbial as they should have, is how much, toward evaluating the sentence "Corey really knows that there is a coin from 1936 in his piggy bank", the temporal reference of the embedded clause needs to be constrained by the expression "at the end of the story". What could be happening is that participants refer the knowledge state to the correct *de re* belief Corey had *at the beginning of the story*, but then judge it to be reliable enough relative to normal circumstances to count it as knowledge at the end of the story, given the rather exceptional circumstances of the Authentic evidence scenario. If that were the case, it might give support to the idea that knowledge is ascribed in the way suggested by Kratzer: first by checking that the causal condition is satisfied, and then by accepting the belief in question to be reliable *enough* in general, in spite of the abnormal circumstances described.

Whether or not any of these explanations is the case, it remains an interesting observation that we can at least capture the distinction between apparent evidence and authentic evidence proper in terms of the distinction between *de re* and *de dicto* belief. If only for that reason, I think this mitigates Starmans and Friedman's claim that "while people sometimes do deny knowledge in Gettier scenarios, it is not because of the disconnect described above".

6. Conclusions

My first two conclusions will be to stress lessons from Kratzer's important paper. The first is that an analysis of knowledge in terms of correct *de re* belief accounts surprisingly well for the original Gettier cases. The second, again in agreement with Kratzer, is that the *de re*-belief-of-facts analysis gives us only part of the truth conditions for knowledge. Admittedly, we must prise apart Gettier cases in the strict sense from Ginet-Goldman cases: both manipulate different dimensions of evaluation, and the weight of those dimensions in knowledge attributions remains an open question.

Let me add two more personal conclusions: the first is that we did not have to distinguish facts and true propositions to account for the original Gettier cases. It was enough to use mechanisms of binding and scope. This is not to say that thinking of facts as distinct from true propositions is necessarily misguided, but merely to say that we do not have to worry about the definition or construction of facts to account for the Gettier cases. The second point is that we were able to use the *de re/de dicto* distinction about belief to account

for Starmans and Friedman's distinction between authentic vs. apparent evidence in their study of Gettier cases. Of course, it remains surprising that knowledge is ascribed to such a large extent for cases of authentic evidence, but at the very least this suggests that some notion of causal connection is still operative in the contrast uncovered by Starmans and Friedman.

7. References

Fitting M. and Mendelsohn R. (1998). *First-Order Modal Logic*. Synthese Library, vol. 227, Kluwer academic publishers.

Gettier E. (1963). Is Justified True Belief Knowledge? *Analysis* 23, 121-123.

Goldman A. (1967). A Causal Theory of Knowing. *The Journal of Philosophy* 64, 357-372.

Goldman A. (1976). Discrimination and Perceptual Knowledge. *The Journal of Philosophy* 73, 771-791.

Kratzer A. (2002). Facts: Particulars or Information Units? *Linguistics and Philosophy* 25: 655-670.

Lycan W. (2006). On the Gettier Problem Problem. In Stephen. Hetherington (Ed.), *Epistemology Futures*: 148-168. Oxford: Clarendon Press.

Nagel J. and San Juan V. and Mar R. (2013a). Lay Denial of Knowledge for Justified True Beliefs. *Cognition* 129: 652-661.

Nagel J. and San Juan V. and Mar R. (2013b). Authentic Gettier cases: A reply to Starmans and Friedman. *Cognition* 129: 666Ð669.

Russell B. (1912). *The Problems of Philosophy*. Oxford University Press.

Russell B. (1918). The Philosophy of Logical Atomism. *The Monist* 28 (4): 495-527.

Sosa E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge* (Vol. 1). Oxford University Press.

Turri J. (2012). Is Knowledge Justified True Belief? *Synthese*, 184(3), 247-259.

Starmans C. and Friedman O. (2012). The Folk Conception of Knowledge. *Cognition*, 124: 272-283.

Starmans C. and Friedman O. (2013). Taking 'Know' for an Answer: A reply to Nagel, San Juan, and Mar. *Cognition* 129: 662-665.

Vendler Z. (1972). *Res Cogitans*. Cornell University Press Ithaca.

21

Contextual Logic and Epistemic Contexts

YVES BOUCHARD

Abstract In this paper, I analyze the notion of context developed by McCarthy and Buvač (1994) and their contextual logic in order to characterize, from an epistemological point of view, a workable notion of epistemic context. This analysis contributes to showing how epistemological contextualism can be formally modeled, and how it can constitutes a general epistemological framework for epistemic normativity.

Keywords Contextual logic; Epistemic context; Epistemic normativity; Epistemic standard; Contextualism

1. Contextual Logic

The notion of context and the contextual logic originally developed by John McCarthy in the field of artificial intelligence aim at providing a solution to the problem of generality, i.e., the problem of representing ordinary knowledge and its integration into inferential processes operating on knowledge bases. The contextual logic of McCarthy and Buvač (CL_{MCB}) can be defined generally as $FOL \cup \{ist(c, \phi)\}$, where FOL is classical first-order logic and $ist(c, \phi)$ is an operator meaning that the formula ϕ is true in context c. The operator ist expresses a relation between a formula and a set of first-order true formulas which is reified as a formal object, a context. In CL_{MCB}, the completeness of FOL is preserved (Buvač and Mason 1993; Buvač, Buvač, and Mason 1995), and even though this contextual logic is not strictly speaking an epistemic logic, comparable for instance to Lemmon and Henderson (1959) or Hintikka (1962, 1975), it can be nonetheless represented in a standard multimodal logic (Buvač, Buvač, and Mason 1995).

Buvač (1996) defines the syntax of CL_{MCB} by means of the following axioms and rules[1]:

(PL) $\vdash_k \phi$, where ϕ is an instance of a propositional tautology

(UI) $\vdash_k (\forall x)\phi(x) \supset \phi(a)$

(MP) $\dfrac{\vdash_k \phi \quad \vdash_k \phi \supset \psi}{\vdash_k \psi}$

(UG) $\dfrac{\vdash_k \phi \supset \psi(x)}{\vdash_k \phi \supset (\forall y)\psi(y)}$, where x is not free in ϕ

(K) $\vdash_k ist(k', \phi \supset \psi) \supset (ist(k', \phi) \supset ist(k', \psi))$

(D) $\vdash_k ist(k_1, ist(k_2, \phi) \vee \psi) \supset ist(k_1, ist(k_2, \phi)) \vee ist(k_1, \psi)$[2]

(Flat) $\vdash_k ist(k_2, ist(k_1, \phi)) \supset ist(k_1, \phi)$

(Enter) $\dfrac{\vdash_{k'} ist(k, \phi)}{\vdash_k \phi}$

[1] Instead of $\vdash k : \phi$, I simply use $\vdash_k \phi$ to mean that a formula ϕ is provable (or assertable) in the context k.

[2] Buvač used Δ instead of D to refer to this propositional property of contexts. I shall use D in order to avoid confusion with the usual symbol for knowledge bases, Δ.

(Exit) $\dfrac{\vdash_k \phi}{\vdash_{k'} ist(k, \phi)}$

(BF) $\vdash_k (\forall v) ist(k', \phi) \supset ist(k', (\forall v)\phi)$

The first group (PL, UI, MP, UG) comprises axioms and typical rules of *FOL*. In the second group ($K, D, Flat, Enter, Exit$), the axioms and rules express propositional properties of contexts; axiom K is a principle of deductive closure (an analogue of the axiom K in modal logic), axiom D (which Buvač called *contextual omniscience*) permits the qualification of any information accessible from any given context, axiom $Unif$ is a principle of information preservation through contexts, and the rules *Enter* and *Exit* permit to access or to leave a context. Finally, in the group of quantificational properties of contexts, there is one axiom (BF) analog to the Barcan formula specifying the relation between the *ist* operator and the universal quantifier.

Classes of Contexts

Buvač (1996) makes a distinction between two classes of contexts, the *knowledge base* contexts (c_{kb}) and the *discourse* contexts (c_d). Whereas in c_{kb} predicates are univocal, in c_d predicates may be ambiguous. A c_{kb} is a set of true propositions, or facts, in a given knowledge base. A c_d is characterized by two components, a set of *epistemic states* and a set of *semantic states*. In an epistemic state, one finds typical elements of a knowledge base, i.e., facts. A semantic state sets the interpretation of a predicate by means of a relation to another predicate in a knowledge base. It is by virtue of such a relation that an ambiguous predicate in a c_d can be disambiguated.

The main motivation behind CL_{MCB} consists precisely in providing a formal framework for eliminating ambiguity.[3] This is where CL_{MCB} presents a special interest for epistemology, in particular for contextualism. Since the knowledge operator has to be interpreted as an indexical term, according to epistemological contextualism (Cohen 1987), it is an operator that requires disambiguation in function of its context of utterance, and thus an epistemic context has to be conceived as a c_d. In this view, CL_{MCB} can shed light on the dynamics at play between the interpretation of the knowledge operator and the epistemic contexts of utterance.

[3]It can also be extended to other types of contexts (Guha and McCarthy 2003).

2. Epistemic Contexts

In order to take advantage of CL_{MCB}, I will need to load the notion of c_d with some epistemological content. The notion of epistemic context (c_ε) that I will be using rests on the idea that *an epistemic context c is a context defined by an epistemic standard ε that is an introduction rule for the knowledge operator in c*. In CL_{MCB} terms, the standard ε is a subset of the axioms of the knowledge base of c (Δ_c), and to each epistemic context c_ε is associated one and only one epistemic standard. Since it is the epistemic context that determines the meaning of the knowledge operator, then an epistemic context can be envisioned as a c_d, i.e., $\varepsilon \subseteq \Delta_{c_d}$ and more specifically $\varepsilon \subseteq SemanticStates(\Delta_{c_d})$ because ε provides the *indexical content* (variable part) of the meaning of the knowledge operator. In accordance with CL_{MCB}, the complete characterization of an epistemic context depends on a twofold characterization: a characterization of its *epistemic standard* (ε) and (if any) a characterization of its *transposition rules* (τ), which are the rules that govern its relations with other c_ε.

These conceptual choices center the investigation on the conditions for context shifting and, by way of consequence, on the conditions for epistemic standard shifting. This is in line with the contextualist goal of accounting on the one hand for the dynamics observable in our epistemic exchanges, that express the variability of the epistemic standards in use, and on the other hand, for the legitimacy of these variations (i.e., they are not epistemic faults).[4] These variations in the use of epistemic standards show clearly our capacity as epistemic agents to regiment our epistemic practices accordingly to a plurality of norms in function of our epistemic needs.

One immediate consequence of the above definition of c_ε is that it entails a relativization of all contexts, including logical contexts, that is to say logical contexts are local epistemic contexts like any other epistemic contexts. This creates a difficulty of representation in CL_{MCB} since CL_{MCB} has been devised with the explicit goal of making available logical reasoning in local contexts (via *lifting*) by means of a grammar incorporating FOL. The rules PL, UI, MP and UG render accessible the resources of FOL in every local context. However, this structure cannot account entirely for contextualism, because from the contextualist point of view FOL is only one epistemic context among others, and one can imagine that in some rich and complex epistemic situations many logics, stronger or weaker than FOL, may be called upon. Consequently,

[4] And contrary to what Schiffer (1996) suggested, contextualism does not need an error theory to accommodate an indexical interpretation of knowledge attributions.

CL_{MCB} has to be amended in order to reify FOL so as to become an object of the language, which in turn requires the conversion of the rules PL, UI, MP, UG, K, and D into properties of epistemic contexts defined by logical standards.

Before considering some examples of epistemic contexts, I want to underline that the whole idea here is to give some insight into this notion of epistemic context through a (very) programmatic approach, and the proposed formalism will depart slightly from CL_{MCB} in that I make an explicit distinction among axioms between epistemic standards and transposition rules. By definition, an epistemic context will require one and only one epistemic standard, and most of CL_{MCB} grammatical rules (PL, UI, MP, UG, K, D) will be directly incorporated into contextual transposition rules. As a toy example of a set of epistemic contexts, consider the following three partial and plausible definitions of some ordinary (and common) epistemic contexts, $c_{logical}$, $c_{empirical}$ and $c_{perceptual}$:

Axioms of $c_{logical}$ (c_{log})

(ε_{log}.1) $(\forall x)(\phi \supset K(x, \phi))$, where ϕ is an instance of a propositional tautology or of a first-order valid formula

(τ_{log}.1) $ist(c_{log}, \phi \supset \psi(x)) \supset ist(c_{log}, \phi \supset \forall y \phi(y))$, where x is not free in ϕ

(τ_{log}.2) $(\forall x)((ist(c_{log}, ist(c, K(x, \phi))) \wedge ist(c_{log}, ist(c, K(x, \phi \supset \psi)))) \supset (ist(c_{log}, ist(c, K(x, \psi)))))$

(τ_{log}.3) $(\forall x)(ist(c_{log}, ist(c, K(x, \phi \supset \psi))) \supset (ist(c_{log}, ist(c, K(x, \phi))) \supset ist(c_{log}, ist(c, K(x, \psi)))))$

c_{log} corresponds to the classical system of FOL. The axiom ε_{log}.1 is the epistemic standard defining c_{log} and it means that any instance of a propositional tautology or of a valid formula of FOL is sufficient for knowledge.[5] τ_{log}.1, τ_{log}.2, and τ_{log}.3 are respectively the syntactic rules UG, MP, and K of CL_{MCB} expressed in terms of rules of transposition. It is worth noting that τ_{log}.2 guarantees reasoning by *modus ponens* within the scope of the knowledge operator in a given and fixed context, in the very same manner τ_{log}.3 preserves de-

[5] One will recognize in ε_{log}.1 an analogue to the rule of necessitation in modal logic.

ductive closure in a logical context.[6] According to the formulation of $\tau_{log}.3$, the epistemic context c of the antecedent and of the consequent remain fixed. Even though the problem of deductive closure escapes the limits of this paper, I shall observe nonetheless that failures of deductive closure take their origin in a confusion between distinct epistemic contexts, something for which the present proposal can account. One can easily see that any valid pattern of inference can be expressed in the form of a rule of transposition and the set of these rules could be ultimately reduced to a single axiom schema.

Axiom of $c_{empirical}$ (c_{emp})

$(\varepsilon_{emp}.1)\ \ (\forall x)(EmpiricalControl(x,\phi) \supset K(x,\phi))$

$\varepsilon_{emp}.1$ stipulates that the condition to satisfy in order to introduce the knowledge operator in this context is some sort of empirical control made by an agent x towards the state of affairs described by a proposition ϕ. The notion of empirical control in $\varepsilon_{emp}.1$ consists only in a set of procedures providing a sufficient level of discrimination between a state of affairs described by a proposition ϕ and a state of affairs described by a proposition (or several propositions) incompatible with ϕ. In the present illustration, no transposition rule enables one to export empirical knowledge into another c_{ε}.

Axiom of $c_{perceptual}$ (c_{per})

$(\varepsilon_{per}.1)\ \ (\forall xv)((See(x,v) \vee Hear(x,v) \vee Taste(x,v) \vee Smell(x,v) \vee Touch(x,v)) \supset K(x,\phi))$, where ϕ is immediately linked to v

As regards the perceptual standard, things are different since v is not a propositional content but rather a perceptual content. The knowledge operator is introduced only in virtue of a perceptual state (or a percept). The knowledge operator is in this way dependent on our physiological mechanisms and their respective limitations (think of the various perceptual biases identified by cognitive psychology for instance). No transposition rule is available in c_{per}.

The fact that neither c_{emp} nor c_{per} contain a transposition rule is determined exclusively by the definitions of the epistemic standards. A transposition rule makes possible the propagation of knowledge either within a given context or between different contexts. As opposed to the grammatical rules *Enter* and

[6] $\tau_{log}.3$ is comparable to a kind of principle of scope alteration that switches the scope of K (superior level) with the one of $\supset$ (inferior level). Such a permutation is tolerable solely in a logical order.

Exit which are only rules of access to information, the transposition rules act as qualification rules in much the same manner epistemic standards themselves do. The transposition rules of c_{log} ($\tau_{log}.1$, $\tau_{log}.2$, and $\tau_{log}.3$) are intracontextual rules of transposition. For reasons of simplicity, no such rule has been defined in c_{emp} and c_{per}. Furthermore, there is no intercontextual rule of transposition for $\left\{c_{log}, c_{emp}, c_{per}\right\}$. In c_{per}, for instance, the assertability conditions are evidently too weak to satisfy the assertability conditions of c_{log} and c_{emp}. There is no intercontextual rule of transposition between c_{per} and c_{emp}, because the satisfaction of ε_{per} does not imply the satisfaction of ε_{emp} (ε_{per} is simply too weak), and conversely, the satisfaction of ε_{emp} does not entail the satisfaction of ε_{per} (a property, for example, may be tested empirically while not being itself an object of direct perception). This shows clearly the primitive character of the notion of epistemic standard, which dictates the possibility or the non-possibility of transposition rules. As for the question whether a transposition rule can be valid *a priori*, i.e., independently of any epistemic standard, one can easily see its irrelevance within the proposed contextualist framework.

Another noticeable aspect of the previous definitions is that no intracontextual rule of transposition specifies the conditions of transmission of a knowledge item from one epistemic agent to another. One could think, for instance, that if an agent a has run an empirical control with respect to ϕ and $K(a, \phi)$, then an agent b, who knows that a has performed a test, would know by some testimonial relation that ϕ. More formally: if $\vdash_{c_{emp}} K(a, \phi)$ and $\vdash_{c_{emp}} K(b, K(a, \phi))$, then $\vdash_{c_{emp}} K(b, \phi)$. The main difficulty in the formulation $\vdash_{c_{emp}} K(b, K(a, \phi))$ can be straightforwardly isolated. If b knows that $K(a, \phi)$, then it is surely not in virtue of ε_{emp} since b is not the one who has run the test, but in virtue of another epistemic standard, namely $\varepsilon_{testimony}$. The specification of all the transposition rules for testimonial knowledge constitutes a major issue from an epistemological point of view. These rules require a fine-grained analysis that is beyond the limits of the present paper. Given that the proposed treatment aims only at presenting a workable notion of epistemic context, it is preferable on this occasion to avoid the problem of the transmission of knowledge from one agent to another.

Epistemological Theory

It seems that in our ordinary epistemic situations, the perceptual standard, the empirical standard, and the logical standard (all defined above) are represen-

tative of the epistemic resources at our disposal as epistemic agents. But the chief interest in the toy example lies elsewhere. In defining epistemic contexts by means of explicit epistemic standards, one not only gives the knowledge operator its various meanings, but one also describes a structure in which epistemic normativity is spelled out in different terms. Such a conception of epistemic normativity allows for multiple configurations of epistemic contexts, which in turn can be captured by the idea that *an epistemological theory is as a set of* c_ε. The epistemological theory presented above, say Θ, is defined as $\Theta = \left\{c_{log}, c_{emp}, c_{per}\right\}$. An epistemological theory is consequently defined by a specific set of epistemic contexts (or knowledge bases), that is to say a specific set of epistemic standards and transposition rules. The epistemological structure of the theory is given by the transposition rules that govern the inter and intracontextual relations between contexts. This definition provides a new perspective on major debates in contemporary epistemology. Foundationalism, coherentism, reliabilism, and other options based on the JTB model, may be construed as exemplifying different epistemological structures designed to meet different epistemic demands. None of them is the ultimate epistemological theory simply because all of them are instances of particular structural configurations.

The specific structure of an epistemological theory shows the relations between the different assertability conditions of the knowledge operator proper to each context. It could seem that this treatment of epistemic normativity is eluding the crucial problem of the truth conditions of the knowledge operator. Of course, this difficulty has to do with the debate between a realist and an antirealist interpretation of the knowledge operator. One merit of proposed view is its clear response: the truth conditions of K in a given epistemic context are provided by the assertability conditions of K in the given context, so that truth-conduciveness from one context to another follows assertability from one context to another. The purpose of a transposition rule is to authorize the dissemination of assertions in multiple contexts on the basis of one given context. The function of transposition rules though is to be sharply distinguished form the function of the *ist* operator, because the formula in the argument position of the operator is in mention not in use. The *Exit* rule makes explicit the genealogy, so to speak, of the truth of a formula from another context, whereas the *Enter* rule does the inverse, i.e., it encapsulates the truth into the assertability conditions of a context. For a realist, this isomorphic relation between truth conditions and assertability conditions boils down to the elimination of the truth conditions, conceived as contextually independent.

Some realists, e.g., Williamson (1996, 2000), go as far in the opposite direction as making knowledge the norm of assertion. Such a reversal in the assertability conditions does not do justice to the observable variability of epistemic standards in our epistemic practices.

These considerations lead naturally to another important difficulty that a contextualist perspective is facing. Can contextualism account for the implication between knowledge and truth, as the factivity (or veridicality) condition requires it, i.e., $K\phi \supset \phi$? This time the debate takes place between a fallibilist and an infallibilist conception of knowledge.[7] The factivity condition springs from an analysis centered on the necessary conditions for knowledge (analysis *in consequentia*). The framework developed here makes explicit only the sufficient conditions for knowledge (analysis *in antecedentia*); the epistemic standards are nothing else than introduction rules for the knowledge operator, and the antecedent of the epistemic standard may not even contain any epistemic terms, depending on the context. In the proposed view, here lies the main interest of contextualism as it constitutes a general epistemological framework within which epistemic normativity can be analyzed primarily in terms of its function rather than its content. So, in order to make explicit the characterization of some K by means of necessary conditions, the general contextualist framework has to be singularized and that process amounts to the specification of an epistemological theory, as previously defined.

According to the proposed framework, and in conformity with McCarthy and Buvač (1994), the epistemic contexts are conceived independently from the epistemic agents. This only means that the epistemic perspective of a given agent does not alter in any way the facts, or the epistemic states of Δ_{c_ε}. This property of *flatness* makes it easier to isolate the contextual variations at the level of the contexts, in other words at the level of their respective transposition rules. This reification of an epistemic context brings autonomy to the context with respect to the epistemic agents, and this accounts for the constraint that within one given epistemic context all of the epistemic agents are regimented by the very same epistemic standard and submitted to the very same epistemic demands. Certainly one could define an epistemic context with a parameter in relation to the propositional attitudes of the epistemic agents so that a context would vary as a function of the agents. But such a change would represent more than a change of epistemological theory, it would be a more radical change of logic (or grammar) since one would have

[7] In the epistemological theory Θ presented above, ε_{per} shows a high level of fallibility, compared to ε_{emp}, which is moderate, and to ε_{log} which is null.

to give up the $Unif$ axiom of CL_{MCB} in order to render possible alterations of the epistemic states of one context by means of another context. No doubt the rejection of $Unif$ would be relevant in some particular epistemological investigations, but within the limits of the proposed approach that would have the undesirable effect of concealing (at least partially) the dynamics between the epistemic standards.

Conclusion

The notion of context and the contextual logic defined by McCarthy and Buvač prove to be rich in epistemological applications. Since their formal notion of context is devised to resolve the problem sof lexical ambiguity, it furnishes by the same token an adequate framework for the indexical interpretation of the knowledge operator. By extending epistemologically this notion of context, one can provide a basis for epistemological contextualism. An epistemic context can then be defined as a set of one epistemic standard and some transposition rules, and an epistemological theory can be defined as a set of epistemic contexts. From this viewpoint, in which an epistemic standard is conceived as an introduction rule for the knowledge operator in a given context, contextualism appears to be an epistemological framework for epistemic normativity in general, rather than a particular epistemological theory in the strict sense.

3. References

Saša Buvač. Resolving lexical ambiguity using a formal theory of context. In *Semantic Ambiguity and Underspecification*, pages 101–124. CSLI Publications, Stanford, 1996.

Saša Buvač and Ian A. Mason. Propositional logic of context. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 412–419, 1993.

Saša Buvač, Vanja Buvač, and Ian A. Mason. Metamathematics of contexts. *Fundamenta Informaticae*, 23:263–301, 1995.

Stewart Cohen. Knowledge, context, and social standards. *Synthese*, 73:3–26, 1987.

R. Guha and John McCarthy. Varieties of contexts. In Patrick Blackburn, Chiara Ghidini, Roy Turner, and Fausto Giunchiglia, editors, *Modeling and*

Using Context, volume 2680 of *Lecture Notes in Computer Science*, pages 164–177. Springer, Berlin, 2003.

Jaakko Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, 1962.

Jaakko Hintikka. Impossible possible worlds vindicated. *Journal of Philosophical Logic*, 4:475–484, 1975.

E. J. Lemmon and G. P. Henderson. Is there only one correct system of modal logic? *The Aristotelian Society*, 33:23–40, 1959.

David Lewis. Elusive knowledge. *Australasian Journal of Philosophy*, 74:549–567, 1996.

John McCarthy and Saša Buvač. Formalizing context (expanded notes). In Atocha Aliseda, Rob van Glabbeek, and Dag Westerståhl, editors, *Computing Natural Language*, pages 13–50. CSLI Publications, Stanford, 1994.

Stephen Schiffer. Contextualist solutions to scepticism. *Proceedings of the Aristotelian Society*, 96:317–333, 1996.

Timothy Williamson. Knowing and asserting. *The Philosophical Review*, 105: 489–523, 1996.

Timothy Williamson. *Knowledge and Its Limits*. Oxford University Press, Oxford, 2000.

22

Acceptation, cohérence et responsabilité *

HENRI GALINON

1. Introduction

Si un agent rationnel accepte une théorie, il est également rationellement justifié à accepter que cette théorie est cohérente ; c'est la thèse que je me propose de défendre ici (T). Cette thèse ne va pas de soi. D'un côté, d'une façon générale, la cohérence d'une théorie n'est pas une conséquence déductible de cette théorie. Accepter la cohérence d'une théorie c'est donc en général accepter plus que la théorie elle-même. D'un autre côté, nous connaissons de nombreux exemples de théories un temps acceptées et qui se sont révélées incohérentes par la suite. En quel sens pouvait-il être rationnellement justifié d'accepter au départ la cohérence de ces théories ?

Je voudrais présenter ici deux façons de soutenir la thèse (T). La première est ancienne et relativement bien connue : on montre qu'il est possible de *déduire* de A *et de principes généraux relatifs la notion de vérité* l'affirmation que tous les théorèmes de A sont vrais et, de là, que A est cohérente. Si un sujet rationnel accepte les prémisses de cette preuve, il doit donc accepter sa conclusion, et nous avons justifié notre thèse de départ. Je présenterai cette explication plus en détail et en montrerai quelques limites.

Mais il est également possible de justifier la thèse (T), c'est du moins ce que je voudrais soutenir, par une tout autre voie, à l'effet que l'acceptation de la

*Le projet de cette étude s'est nourri des nombreuses réflexions de Pascal Engel sur la notion d'acceptation et la dimension normative de la vérité (par exemple 1998, 2000, 2001), ainsi que de remarques de Jacques Dubucs sur le conventionnalisme arithmétique et le problème de la *stabilité des décisions* dans Dubucs 2003.

cohérence par un agent rationnel qui accepte une théorie donnée est justifiée *par défaut*, sur la base d'un certain nombre de principes qui relèvent purement de la rationalité en première personne. Je propose d'explorer cette voie ici en réfléchissant aux contraintes que fait peser l'hypothèse de rationalité d'un agent sur la logique de ses *décisions épistémiques*. Je montrerai que la plausibilité de la thèse de départ dépend de la plausibilité du principe suivant, pour la défense duquel je présenterai quelques arguments :

> (Principe de Responsabilité) Si un sujet rationnel S accepte un ensemble de propositions X, S doit accepter de surcroît qu'il est justifié à accepter X.

2. Accepter, à la réflexion

L'activité scientifique qui est celle du choix théorique est en droit une activité rationnelle méthodique, critique et réfléchie. Pour cette raison, la relation qu'entretient l'homme de science au produit théorique positif de sa recherche me semble mieux décrite comme une relation d'*acceptation* que comme une simple relation de *croyance.* La croyance, comme attitude sur laquelle nous ne pouvons exercer de contrôle volontaire, doit avoir moins de part à l'élaboration des théories qu'une forme d'action volontaire, comme l'est l'acceptation, si cette élaboration doit être une activité rationnellement contrôlée.

Le terme d'acceptation recouvre dans la littérature un spectre assez large d'attitudes différentes (v. par exemple van Fraassen 1980, Cohen 1989, et Engel 1998 pour une discussion), et il est utile de préciser encore un peu la notion que j'ai en vue. A la différence de la plupart de celles développées en opposition à la notion de croyance, le plus souvent en vue d'essayer de comprendre et de mettre en valeur certains aspects pragmatiques du choix théorique et les différentes normes au travail dans ces choix, la notion d'acceptation qui m'intéresse est plus spécifiquement épistémologique. Accepter au sens où je l'entends ici procède d'une décision réflexivement informée et guidée uniquement par des buts qui sont ordinairement reconnus être ceux de l'activité scientifique, sans préférence idéologique marquée : qu'il s'agisse de la connaissance, de l'explication, de la prédiction, ou pourquoi pas de la simple organisation systématique des données disponibles. Je suppose également que, contrairement à un certain usage du terme, l'acceptation d'une proposition est, comme la croyance, susceptible de degrés, correspondants plus ou moins aux degrés auxquels elle est tenue pour justifiée. Ce précisions faites, j'en viens au problème de l'acceptation de la cohérence d'une théorie.

3. Le problème de la cohérence

Etant donnée une théorie quelconque A (nous supposerons seulement que cette théorie est, ou peut être, formalisée de façon à ce que le problème reçoive une définition précise), la question de savoir si A est ou non une théorie cohérente est une question bien définie : c'est celle de savoir s'il est possible dériver une contradiction des axiomes et des règles admises dans A. Si la théorie A est formalisée, c'est-à-dire si son langage et sa structure sont parfaitement spécifiés par des règles effectives de construction, on peut identifier cette théorie à un objet linguistico-syntaxique dont la description et l'étude peuvent eux-mêmes faire l'objet d'un traitement formel. Ainsi, la question de savoir si une théorie formalisée donnée est ou non cohérente peut être vue comme une question purement mathématique.[1]

En fait, si A est une théorie contenant un appareil minimum de mathématiques et de syntaxe (et A peut alors être un fragment d'arithmétique ou la totalité des hypothèses tant formelles qu'empiriques engagées dans la physique newtonienne), la question de savoir si une théorie donnée est cohérente ou non est une question qui pourra être posée dans le langage de A. Et en particulier, si A est une théorie formalisable (et de grands fragment de l'arithmétique aussi bien que la physique newtonnienne le sont), la question de savoir si A elle-même est cohérente est une question qui se pose dans le langage et avec les concepts de A.

Pourtant, c'est un des plus fameux résultats de la logique mathématique (le second théorème d'incomplétude de Gödel), si A est une théorie cohérente (en plus d'être formalisable et de contenir certains principes de mathématiques et de syntaxe élémentaires), l'énoncé standard du langage de A qui affirme qu'aucune démonstration dans A n'est la démonstration d'une contradiction ne peut pas être postulé au titre d'axiomes *de* A, ni *a fortiori* être

[1] En théorie classique, bayésienne, de la rationalité, il est habituel de faire un certain nombre d'idéalisations et, parmi elles, de supposer qu'un agent rationnel doit accorder à toutes les vérités mathématiques une croyance de degré 1, et de degré 0 aux autres énoncés mathématiques. Mais si la cohérence (ou l'incohérence, selon les cas) d'une théorie est un fait mathématique, alors il est rationnel de croire cohérente une théorie si et seulement si cette théorie est cohérente. Notre recherche sur la dépendance qui doit exister, en matière de comportement rationnel, entre acceptation de A et acceptation de la cohérence de A serait minée d'emblée par cette idéalisation. Par conséquent, nous renoncerons ici à cette idéalisation et adopterons l'hypothèse que les théories mathématiques doivent être traitées sur un pied d'égalité avec les théories empiriques ordinaires (une hypothèse qui est par ailleurs en phase avec certaines recherches d'épistémologie des mathématiques. Voir par exemple Maddy 1990. Voir aussi Detlefsen 1979, note 1, pour une remarque analogue.).

déduit des axiomes de A. Par conséquent, si nous voulons expliquer qu'un agent acceptant une théorie A doit accepter que A est cohérente, nous ne pouvons espérer procéder, même dans les cas *a priori* favorables dans lesquels la cohérence de A est exprimable dans le langage de A, en montrant que la cohérence de A est une conséquence logique de A.[2] L'explication, si elle est possible, doit faire appel à des principes auxiliaires extérieurs à *A stricto sensu*.

4. La cohérence par la vérité

Une façon d'expliquer pourquoi un agent rationnel croyant une théorie A au degré x doit croire A cohérente à un degré supérieur ou égal x, consiste à montrer que l'on peut dériver de A la cohérence de A moyennant l'introduction de principes connus *a priori* concernant la notion de *vérité* pour le langage de A.[3] Ces principes, dégagés par Tarski, et qui ne sont en substance qu'une généralisation des énoncés de la forme :

> « F » est vraie si et seulement si F
> (où F est mis pour n'importe quel énoncé du langage de A)

constituent ensemble ce que j'appellerai la théorie tarskienne de la vérité (pour le langage de A), que je noterai T_V.[4]

Moyennant l'introduction de ces principes aléthiques, la dérivation de la cohérence de A dans $A + T_V$ procède de la façon suivante :

[2] Je fais ici tacitement l'hypothèse le langage dans lequel est formalisé la théorie A est un langage du premier ordre ; par conséquent, quand je parlerai de « conséquence logique », c'est « conséquence logique du premier ordre » qu'il faudra comprendre. Cette hypothèse est importante, mais ce n'est pas le lieu de la discuter ici.

[3] Pour faciliter la présentation, je supposerai que la théorie A contient déjà elle-même les principes syntaxiques et mathématiques permettant de décrire sa propre structure morphologique et déductive, y compris donc un principe schématique d'induction permettant de conduire des démonstrations par récurrence sur la longueur des dérivations (ou simplemen un principe schématique d'induction arithmétique si nous supposons que la syntaxe est codée dans l'arithmétique, à la Gödel).

[4] Ce que l'on appelle « théorie tarskienne de la vérité », ou parfois la "théorie compositionnelle de la vérité", est une théorie dont les axiomes sont les clauses récursives qui sont utilisées dans Tarski 1983 pour construire une définition explicite du prédicat « vrai-dans-L_A » dans une théorie essentiellement plus riche que A (par exemple une théorie d'ordre supérieur à l'ordre de A). Les détails de la théorie ne sont pas importants ici et je préfère rester vague sur les moints techniques qui obscurciraient inutilement les enjeux à ceux qui ne sont pas familiers de ces questions. Pour les détails logiques, parfaitement connus et classiques, voir par exemple Shapiro 1998 ou Feferman 1991.

1. Axiomes de A (prémisse 1)
2. Axiomes de T_V (prémisse 2)
3. Tous les théorèmes de A sont vrais (par déduction à partir de 1 et 2)[5]
4. L'ensemble des énoncés vrais est cohérent (par déduction à partir de 2 et de la théorie de la syntaxe comprise dans A)[6]
5. Donc A est cohérente. (Par 3 et 4)

Puisque la cohérence de A est déductible de A et de la théorie de la vérité pour le langage de A, un agent rationnel qui accepte la théorie A et les principes aléthiques en question, doit accepter que A est cohérente à un degré supérieur ou égal à son degré d'acceptation de leur conjonction, et nous avons une justification de notre thèse de départ.

Mais ce qui est remarquable dans cette preuve, du point de vue qui nous intéresse ici, c'est le détour qu'elle impose par l'affirmation que *la théorie A est vraie* (le point 3 du schéma de preuve ci-dessus). Car ce détour ouvre un espace de discussion possible relativement à la question de savoir si cette explication est toujours adéquate.

Si l'on en croit van Fraassen (van Fraassen 1980), par exemple, accepter une théorie et accepter que cette théorie est vraie sont deux choses différentes, et il n'est pas en général légitime d'identifier l'acceptation de A et l'acceptation de la vérité de A.[7] Bien sûr, nous venons de le rappeler, le passage de l'un à l'autre est logiquement garanti en présence de la théorie tarskienne de la vérité ; mais le point est l'on peut douter que cette théorie ait un sens pour

[5]Pour être précis, d'un point de vue logique, il est crucial ici de permettre au prédicat de vérité d'apparaître dans le schéma d'axiome d'induction de la théorie A pour pouvoir obtenir la conclusion 3 à partir de 1 et 2. Sinon on peut montrer que $T_V + A$ est en fait une extension conservative de A. Autrement dit, dans la terminologie de Feferman, on supposera que A est une théorie *schématique*. Voir par exemple Feferman 1991 sur cette notion et pour les preuves des affirmations précédentes. Il existe un débat philosophique concernant la signification épistémologique de ce genre de preuve sémantique de la cohérence. Pour une discussion, voir par exemple Shapiro 1998, Ketland 1999, Field 1999, et Tennant 2002. Sur toutes ces questions on pourra également consulter l'utile ouvrage Horsten 2011.

[6] Plus précisément la dérivation 3–5 peut être présentée de façon élémentaire comme suit. Tous les théorèmes de A sont vrais. Supposons que l'on puisse montrer, par exemple que $1 + 1 = 2$ est démontrable dans A. Il s'en suit que « $1 + 1 = 2$ » est vrai. Dans T_V on peut alors en déduire que « $1 + 1 \neq 2$ » n'est pas vrai, car T_V démontre que, pour tout énoncé F du langage de A, non-F est vrai si et seulement F n'est pas vrai. Donc « $1 + 1 \neq 2$ » n'est pas un théorème de A. Ce qui est une façon de dire que A est cohérente (tout énoncé est déductible d'une théorie incohérente).

[7] Cette position semble être également celle de Engel (1998).

van Fraassen relativement au langage d'une théorie envers laquelle il est disposé à entretenir une attitude anti-réaliste. Si les énoncés instrumentaux des parties les plus théoriques de la science n'ont pas de signification, ou n'ont qu'une signification incomplète, alors la question de leur vérité se pose pas, et l'application du prédicat de vérité à ces énoncés n'est pas légitime.

Mais même si l'on adopte une attitude réaliste générale, et l'idée que tout le langage de la science est réellement descriptif, la preuve de la cohérence par la vérité apparaît comme un détour étonnant. Ce que nous essayons de justifier est le caractère *rationnel* d'une certaine décision (la décision d'accepter cohérence) dans un certain contexte épistémique ; (T) est un principe qui a trait à la *structure* de l'ensemble des décisions prises par un agent rationnel, indépendamment de la valeur de vérité des hypothèses qu'il accepte; par conséquent on ne voit pas bien pourquoi l'acceptation par l'agent du caractère vérace des propositions qu'il accepte devrait jouer un rôle essentiel dans la justification de son acceptation de leur cohérence : il y a là une forme d'impureté de la justification qui nuit à la manifestation de l'ordre des raisons. Ce que l'on voudrait, c'est en donner une justification qui ne fasse appel qu'a des considérations relatives à la nature de l'action rationnelle et à la structure des états épistémiques de l'agent.

5. Un *Dutch book* gödelien

Pour montrer qu'un agent acceptant une théorie A tout en acceptant à un moindre degré que A est cohérente est irrationnel, l'idée de montrer qu'il est vulnérable à un *dutch book* se présente d'elle-même.[8] Supposons, pour fixer les idées, que nous croyions la théorie A au degré 1, mais croyions en la cohérence de A à un degré strictement inférieur à 1, disons $0,5$. Un bookmaker hollandais, appelons-le Kurt, nous propose d'acheter $0,4$ un pari qui paye 1 si A n'est pas cohérente, rien sinon. Etant donné nos croyances, ce pari est acceptable, et même avantageux. Maintenant supposons que A est cohérente : alors le pari est perdu, nous perdons $0,4$ et Kurt gagne $0,4$. Si maintenant A n'est pas cohérente : alors nous gagnons $0,6$ sur ce pari, mais puisque A n'est pas cohérente et que nous acceptons au degré 1 tous les théorèmes de A, nous sommes vulnérable à un *dutch book* (et même une infinité) dont l'issue est une perte certaine de 1 pour nous. Au bout du compte, nous perdrions donc $0,4$ et Kurt gagnerait à nouveau $0,4$. Par conséquent, si notre degré de croyance

[8] Van Fraassen 1984 développe une stratégie semblable pour défendre un autre type de principe de réflexif.

dans les axiomes de A est 1, nous ne devrions pas accepter de payer pour un pari sur l'incohérence de A.

Cette esquisse d'argument semble montrer qu'il est irrationnel d'accepter la cohérence de A à un degré moindre de celui auquel nous acceptons A, en un sens inspiré des théories bayesiennes de la rationalité, et sans qu'il y ait contradiction formelle entre A et la proposition que A n'est pas cohérente. Mais avons-nous réellement montré sans recours à la notion de vérité qu'un sujet acceptant une théorie A doit également accepter la cohérence de A ? En réalité, pas tout à fait. En effet, il se peut que le joueur qui parie à la fois sur A et sur l'incohérence de Asoit irrationnel, quoique la conjonction de A et de la proposition que A est incohérente ne soit pas logiquement contradictoire ; le problème est que pour le reconnaître il semble qu'il faille reconduire le détour par la vérité. Car le raisonnement que nous avons tenu pour prouver la ruine certaine du sujet met en jeu, de façon cachée, le concept de vérité : car comment savons-nous que si A n'est pas cohérente, alors Karl perdra de l'argent, sinon parce que nous avons dérivé *logiquement de la vérité de A la cohérence de A* ? Si nous devions écrire dans le détail la façon dont nous avons calculé les gains et les pertes associés au contrat proposés par Kurt, nous nous apercevrions que nous avons précisément fait le genre de raisonnement présenté dans la section précédente, c'est-à-dire que nous avons fait un détour par la notion de vérité pour rendre compte du fait que la cohérence de A est une conséquence de A et d'un petit nombre de principes analytiques concernant la notion de vérité.

6. Réflexion épistémique

La preuve de la cohérence par la vérité faisait un détour par ce que l'on pourrait appeler le Principe de réflexion aléthique sur A :

Principe de Réflexion aléthique : A est vraie

Une fois ce principe justifié, il suffisait de prouver dans un second temps que l'ensemble des énoncés vrais est cohérent, ce dont une analyse conceptuelle de la notion de vérité nous assure, pour inférer la cohérence de A. Mais la cohérence n'est pas une propriété que possèderait exclusivement les ensemble d'énoncés vrais. Je voudrais à présent soutenir que la seconde partie du raisonnement à l'œuvre dans la "preuve par la vérité" peut être reproduite en remplaçant le principe de réflexion aléthique sur A par un principe de réflexion *épistémique* sur A :

Principe de Réflexion épistémique : Je suis justifié à accepter A

En effet, que signifie l'affirmation que je suis justifié à accepter la proposition ou la théorie A, ou comme je dirai aussi de façon synonyme, que A est acceptable (par moi, maintenant) ? Cela signifie que mon acceptation de A satisfait à une certaine norme, que j'affirme qu'il m'est permis, au regard d'une certain code tacite d'éthique épistémique, d'accepter A.

Bien entendu, la question de savoir ce que doit contenir un tel code éthique est aussi difficile que la question de la nature de la justification elle-même, comme en témoigne l'abondante littérature épistémologique sur cette question.[9] Faut-il inscrire dans ce code l'injonction de Descartes de n'accepter que ces énoncés dont la vérité est claire est distincte ? Faut-il suivre la règle de Clifford et proportionner toujours et partout notre acceptation aux évidences disponibles ? On peut en douter. Mais d'autres règles pour la direction de l'esprit sont sans nul doute moins problématiques. On peut penser qu'il n'est pas permis d'accepter une hypothèse en présence seulement d'indices de sa fausseté ; ou qu'il n'est permis d'accepter une observation que si nous avons vérifié que les conditions de cette observation remplissaient un certain nombre de critères variées (des conditions d'éclairage à la reproductibilité de l'observation). C'est sans doute seulement lorsque de telles conditions sont réunies que je suis justifié à accepter une hypothèse, et seulement lorsque je me suis assuré de la conformité de mes actions à cette éthique épistémologique que je peux me reconnaître justifié à accepter ce que j'accepte.

Maintenant, ce qui semble ne faire aucun doute dans l'analyse des conditions structurelles sous lesquelles un sujet est justifié à accepter l'ensemble des propositions qu'il accepte, c'est l'idée que cet ensemble doit au minimum être cohérent. C'est ce principe qui permet de rendre compte du fait que nous ne sommes pas prêts à accepter une théorie que *nous tenons* pour incohérente. Le problème n'est pas seulement qu'une théorie incohérente doive être fausse (car après tout, à nouveau, cette idée n'a qu'une application limitée pour un instrumentaliste). Le problème est plutôt qu'une théorie incohérente est inutile. Puisqu'il faut accepter les conséquences logiques de ce que l'on accepte, accepter une théorie incohérente reviendrait à tout accepter, c'est-à-dire tout et son contraire. Or le point même de l'élaboration d'une théorie, *in fine*, est, *a minima*, la *discrimination* de certains énoncés, ceux que l'on accepte, de ceux que l'on accepte pas, dans une organisation systématique et compacte. Si tout est acceptable, alors c'est l'objet même de cette activité qui disparaît. Par conséquent, il est hautement plausible que, de même qu'une analyse conceptuelle

[9] Je renvoie à Alston 1988 pour une entrée dans cette litterature.

de la notion de vérité révèle que l'ensemble des énoncés vrais est cohérent, de même une analyse conceptuelle de la notion de justification doit révéler que l'ensemble des énoncés qu'un agent est justifié à accepter doit être tenu pour cohérent. Par conséquent, si un sujet juge A acceptable, alors il doit juger A cohérente. Il doit donc accepter le principe suivant: si A est acceptable, alors A est cohérente (les théories incohérentes ne sont pas acceptables).

Remarquons maintenant que si ce que nous venons de dire est correct, la dérivation de la cohérence de *A à partir* du principe de Réflexion épistémique est quasiment immédiate :

1. A est acceptable (Principe de réflexion épistémique)
2. Si A est acceptable, alors A est cohérente. (réflexion sur les normes d'acceptabilité)
3. Donc A est cohérente. (par 1, 2)

Il reste donc à examiner si le principe de réflexion épistémique peut lui-même être justifié, et comment.

7. Le principe de responsabilité en première personne

Nous avons une dérivation de la cohérence de A à partir du principe de réflexion épistémique. Cette dérivation ne souffre pas du défaut dont souffrait la preuve par la vérité : elle ne fait appel qu'à des concepts épistémologiques et sa validité devrait convaincre tant les philosophes enclins à une forme d'instrumentalisme ou d'anti-réalisme vis-à-vis de A que les philosophes soupçonneux à l'égard de la notion de vérité. Mais cette dérivation est encore loin de constituer en elle-même une explication de notre thèse de départ (T). Ce qu'il s'agissait d'expliquer, c'est qu'un agent rationnel acceptant une théorie A doit accepter la cohérence de A. Or il y a un fossé conceptuel apparemment infranchissable entre l'*acceptation* (de A) par un agent et l'*acceptabilité* de A, entre le fait qu'un agent accepte une théorie et le fait que cet agent soit justifié à accepter cette théorie.

Le principe qui permet de faire le pont entre la petite dérivation de la section précédente et la thèse (T) est le suivant, que j'appellerai le Principe de Responsabilité :

> (Principe de Responsabilité) Si un agent rationnel S accepte un ensemble de propositions X, S doit accepter « X est acceptable ».

Si ce principe est correct, en effet, nous avons l'explication cherchée :

1. S accepte A (notre hypothèse de départ)

2. Donc S doit accepter « A est acceptable » (par 1 et Responsabilité)

3. Or S doit juger que si X est acceptable, alors X est cohérent (réflexion sur les normes d'acceptabilité/justification)

4. Donc S doit accepter « A est cohérent » (2 et 3)

La question est donc savoir si le principe de Responsabilité est correct.

Pour comprendre ce qui est en jeu, il est important de noter que le principe suivant, avec son implication matérielle, est évidemment *faux* :

J'accepte $X \rightarrow X$ est acceptable

Il peut être *vrai* qu'un agent rationnel accepte de fait la théorie A, sans que pour autant A satisfasse aux critères d'acceptabilité. C'est une situation banale dans laquelle l'agent s'est simplement trompé et une illustration parmi d'autre du fossé qui existe entre ce qui est et ce qui doit être. Comment alors le principe de responsabilité peut-il être un contrainte rationnelle ? Une façon de le voir est de considérer le cas d'un agent rationnel qui accepterait :

(*) La terre tourne autour du soleil, mais je ne suis pas justifié à accepter que la terre tourne autour du soleil.

Les conditions de vérité de cet énoncé ne sont pas problématiques, pas plus que, pour prendre un exemple célèbre entre tous, la négation du *cogito* cartésien (« Je ne suis pas ») n'est une contradiction logique en elle-même. Le caractère paradoxal de ces affirmations n'est pas à chercher dans le contenu sémantique de ce qui est affirmé. Le paradoxe est pragmatique, au sens où c'est un paradoxe de l'*action* rationnelle ; autrement dit ces énoncés ne sont pas paradoxaux, c'est leur affirmation, ou leur acceptation qui l'est.

De même, un sujet qui accepterait "A, mais A n'est pas acceptable" serait dans une situation quelque peu paradoxale. Quel est, dans le cas qui nous occupe, la source du paradoxe ? Pourquoi un agent acceptant une théorie doit-il accepter de surcroît que cette théorie est acceptable ? Nous avons dit que l'acceptation au sens où nous employons ce terme est une action délibérée, et que nous avions plus particulièrement en vue l'acceptation dans un contexte scientifique, réflexif et critique. Dans ces conditions l'acceptation d'une hypothèse ou d'une théorie est *lumineuse*, au sens où, si nous l'acceptons nous

savons que nous l'acceptons. Dès lors, je propose que la clef du paradoxe, et du même coup la justification du principe de responsabilité, est à chercher dans une la réflexion sur la relation que doit entretenir un agent rationnel au contenu de ce qu'il accepte. Cette relation ne peut pas être simplement conçue sur le modèle observationnel, celui d'un agent constatant simplement qu'il accepte une hypothèse ou une théorie donnée. Au contraire, la rationalité d'un agent commande que la nature de l'articulation de ses jugements de premier ordre et de ses jugements à propos de ses propres jugements incorpore essentiellement le fait que les uns comme les autres sont *ses* pensées. On a pas la même relation épistémique, en termes de droits comme en termes de devoirs, avec le contenu de ses propres pensées et avec le contenu des pensées d'autrui, quand bien mêmes ces contenus seraient identiques d'un point de vue sémantique. L'idée qu'il existe un lien essentiel entre la rationalité et la nature de notre rapport à nos propres pensées, n'est pas une idée nouvelle. C'est au contraire un thème classique des études philosophiques sur la connaissance de soi. Tyler Burge (1996) écrit par exemple :

> «Trouver de façon justifiée ses propres raisons invalides ou ses pensées injustifiées, est normalement *en soi* une raison paradigmatique, du point de vue des pensées examinées (ainsi que dans la perspective de l'examen), de les altérer [...]. L'examen des raisons qui est partie intégrante du raisonnement critique inclut l'examen et les attitudes examinées en un unique point de vue. Le modèle observationnel simple traite l'examen et le système examiné comme dissociés d'une façon qui est incompatible avec les normes de l'examen critique. Il fait du système examiné un objet d'investigation, mais non une partie du point de vue de l'investigation. [...] Nous sommes épistémiquement responsables seulement parce que nous sommes capables d'examiner nos pensées et nos raisonnements.[...] Notre responsabilité lorsque nous réfléchissons sur nos pensées s'étend immédiatement à l'ensemble du point de vue. » (Burge 1996, p.110-111).

Si Burge a raison alors, plus généralement, il faut conclure que pour toute norme N d'acceptation de p telle que l'échec à la satisfaire constituerait une raison de ne pas accepter p, si un sujet rationel accepte P, il est rationnellemet engagé à accepter que P est N. De ce point de vue, la cohérence n'est qu'un cas particulier, et pour ainsi dire minimal, d'une telle norme.

Ce même principe de responsabilité rend compte de la rationalité, pour un sujet engagé dans une certaine pratique de preuve (celles associées à l'arithmétique

du premier ordre par exemple), d'accepter non seulement ce qu'il a prouvé (les théorèmes de l'arithmétique), mais encore les principes qui explicitent le fait que si un éononcé ou un ensemble d'énoncés ont été prouvés alors ils satisfont la norme qui est le *point* même de notre engagement dans cette pratique discursive. Les logiciens, en paticulier Solomon Feferman (1962, 1991), ou John Myhill (1960), qui se sont intéréssé aux "principes de réflexion", du type "Si p est prouvable alors p", ont reconnu depuis longtemps que ces principes s'imposent rationnellement à quelqu'un qui est engagé dans la pratique de la démonstration par certain moyens de preuve, mais ils ont surtout cherché a étudié la force logique de ces principes – ce qui s'en déduit – et se sont peu intéressés à leur justification. On peut voir l'appel au principe de responsabilité comme un premier pas dans cette direction.

8. Conclusion

Discuter plus à fond le principe de Responsabilité nous engagerait trop loin. Il me suffit ici d'avoir montré qu'un tel principe pouvait avoir un rôle à jouer dans une explication du caractère justifié *par défaut*, c'est-à-dire en l'absence d'indice positifs de leur vérité, de l'acceptation de certaines hypothèses, et en particulier de la cohérence d'une théorie que nous acceptons. a justification de la cohérence par la vérité montrait ceci : si nous avons un indice de la vérité A, nous avons indirectement un indice de la cohérence de A. En effet, il existe une inférence déductive, une suite d'opérations préservant la vérité, de A à la cohérence de A, moyennant les principes auxiliaires et vrais *a priori* gouvernants la notion de vérité. Il est simplement contradictoire, au sens logique usuelle, d'accepter qu'une théorie est vraie sans accepter qu'elle est cohérente. La justification de l'acceptation de la cohérence par le principe de responsabilité est d'une nature fondamentalement différente. Ce n'est pas une *preuve* de la cohérence : un indice que j'accepte A *n*'est *pas*, sans hypothèses substantielles supplémentaires, un *indice*, même indirect, du fait que A est acceptable et de la cohérence de A. Cette justification du caractère rationnel de l'acceptation de la cohérence d'une théorie que nous acceptons (aussi longtemps que nous l'acceptons) est donc une forme de justification *par défaut*, qui vaut en l'absence d'indice de la vérité de la cohérence. C'est aussi une justification *défaisable*, au sens où elle perd toute force si apparaissent des indices positifs de l'inacceptabilité de A (en particulier si nous découvrons une contradiction dans A, nous ne sommes plus justifiés à accepter que A est cohérente : mais nous ne le sommes plus non plus à accepter A). Elle n'en est

pas moins rationnelle.

L'idée qu'il est rationnel de tenir certaines propositions pour justifiées par défaut a été défendue dans la littérature par Crispin Wright (v. Wright 2004) dans un effort pour tirer des leçons épistémologiques du scepticisme radical. Les propositions que Wright vise à justifier de la sorte sont ce qu'il appelle « les pierres de touche » de toute entreprise cognitive, ces hypothèses sans lesquelles nous ne pourrions regarder aucune de nos méthodes de justifications pour correctes (les lois de la logique, le fait que nous ne sommes pas le jouet d'un malin génie, etc.). Je suggère que nous pensions au fait de la cohérence de ce que nous acceptons comme faisant partie de ces pierres de touche. De même qu'il semble vain de chercher une justification positive ultime à toute connaissance - mais qu'il y a un sens auquel nous sommes justifiés a priori, par défaut, à faire certaines hypothèses qui fondent la possibilité même de l'enquête - , de même, si ultimement nous ne pouvons espérer prouver la cohérence de nos théories, il est néanmoins permi de tenir notre acceptation de cette cohérence pour justifiée.[10] La présente approche fonde la possibilité d'une telle justification de notre acceptation de la cohérence sur les exigences spéciales de la rationalité en première personne, en ceci que c'est cette perspective qui donne du sens à l'idée de responsabilité : parce que *je* décide d'accepter une théorie donnée, certaines décisions supplémentaires s'imposent *à moi*.

9. Références

ALSTON, P., 1988, "The deontological conception of justification", *Philosophical Perspectives*, 2.

BURGE, T., 1996, « Our entitlement to self-knowledge », *The Journal of Philosophy*, 85, 11, 649-63

COHEN, J., 1989, « Belief and acceptance », *Mind*, 98, 391, 367-389

DETLEFSEN, M., 1979, "On Interpreting Gödel's Second Theorem", *Journal of Philosophical logic*, 8, 3, 297-313.

[10]Si ce que nous disons est correct, il y a donc là une façon - certes à la marge des intérêts habituels des philosophes des mathématiques, mais néanmoins intéressante - de rendre compte de la spécificité épistémologique des énoncés de Gödélien de cohérence parmi l'ensemble des énoncés laissé indécidés par une théorie mathématique comme la théorie des ensemble de Zermelo-Fraenckel. Le façon dont Gödel lui-même comprenait cette spécificité est présentée dans Gödel 1964. Nous remettons à une autre occasion la comparaison approfondie de ces deux façons d'envisager les choses.

DUBUCS, J., 2003, « Carnap, Gödel et la nécessité mathématique », in Lepage F. et Rivenc F. éd. *Carnap aujourd'hui*, Paris, Vrin-Bellarmin.

ENGEL, P., 1998, « Believing, holding true and accepting », *Philosophical explorations*, 1, 2, 140-151

ENGEL, P. 2000, Believing and Accepting. Springer.

ENGEL, P. 2001, "Is Truth a Norm ?" Dans : Interpreting Davidson. Ed. par Petr Kotatko, Peter Pagin et Gabriel Segal. CSLI, p. 37–51.

FEFERMAN, S, 1962, "Transfinite recursive progressions of axiomatic theories" in *Journal of Symbolic Logic* 27, p. 259–316.

FEFERMAN, S., 1991, « Reflection on Incompleteness », *Journal of Symbolic Logic*, 1-48

FIELD, H., 1999, « Deflating the conservativeness argument », *Journal of Philosophy* 96, 533-540.

GÖDEL, K., 1964, "What is Cantor's continuum problem ?" *in* PAUL BENACERRAF et HILARY PUTNAM (éd.) *Philosophy of Mathematics : Selected Reading (2nd ed.)* ed, Cambridge University Press, p. 470–485.

HORSTEN, L., 2011, « The Tarskian Turn », Oxford University Press.

KETLAND, J., 1999, « Deflationism and Tarski's Paradise », *Mind*, 108, 69-94.

MADDY, P., 1990, *Realism in mathematics*, Oxford University Press.

MYHILL, John (1960). "Some remarks on the notion of proof" in *Journal of Philosophy* 57.14, p. 461–471.

SHAPIRO, S., 1998, «Proof and Truth : Through Thick and Thin » , *Journal of Philosophy*, 95, 10, 493-521.

TARSKI, A., 1983, *Logic, Semantics, Metamathématics*, Hackett pub.

TENNANT, N., 2002, « Deflationism and the Gödel-Phenomena », *Mind*, 111.

VAN FRAASSEN, B., 1980, *The Scientific Image*, Oxford University Press.

VAN FRAASSEN, B., 1984, « Belief and the Will », *The Journal of philosophy*, 81, 5, 235-256

WRIGHT, C., 2004, « Warrant for nothing (and foundations for free) ? », *Philosophical Studies*, 106, 1-2, 41-85.

23

Duperie de soi, croyance et acceptation

VASCO CORREIA

Résumé La distinction entre la notion de croyance et celle d'acceptation permet à Engel et Cohen d'élucider le phénomène de duperie de soi sans se heurter aux paradoxes qui menacent la conception de Davidson (paradoxe doxastique, paradoxe de la stratégie et paradoxe de l'homoncule). Dans cet article, je montre que la solution de Cohen n'y parvient pas tout à fait, notamment parce qu'elle laisse sans réponse la question de savoir comment une croyance indésirable pourrait être intentionnellement « supprimée » par le sujet. La conception dispositionnelle-fonctionnaliste proposée par Engel, en revanche, a le mérite de rendre compte du caractère tacite, voire non-conscient des croyances neutralisées dans le processus de duperie de soi, ainsi que du mécanisme causal qui le sous-tend. J'émets cependant quelques réserves quant à l'hypothèse que la neutralisation de la croyance indésirable que *p* s'accompagne de l'« acceptation » de la croyance contraire que non-*p*. Mon hypothèse est que le sujet qui se dupe lui-même ne fait pas qu'*accepter* le faux ou l'invraisemblable, mais y *croit* tout-à-fait.

Mots clé acceptation, croyance, duperie de soi, homoncularisme, paradoxe doxastique, paradoxe de la stratégie, soupçon, volitionisme.

1. Introduction

Si le phénomène de « duperie de soi » (*self-deception*) continue de fasciner les philosophes et les psychologues, c'est peut-être parce qu'il renferme une énigme aux allures de paradoxe : d'un côté, celui qui se dupe lui-même doit être un tant soit peu conscient de la vérité, car on ne peut *duper* stricto sensu que si l'on connaît la vérité ; mais, d'un autre côté, cette conscience préalable de la vérité semble compromettre l'adhésion au mensonge en question, car on ne peut être *dupe* que si l'on ignore le vrai. Ainsi que le souligne Sartre, le problème tient ici à ce que le sujet doit savoir en tant que trompeur la vérité qui lui est masquée en tant qu'il est trompé[1]. D'après Davidson, c'est justement parce que le *self-deceiver* entretient au départ la croyance qu'il cherche ensuite à nier que le phénomène de duperie de soi ne saurait se confondre avec la simple « prise des désirs pour des réalités » (*wishful thinking*) : « pour être dupe de soi-même, on doit avoir su la vérité à un moment donné, ou, pour être plus précis, on doit avoir cru quelque chose de contraire à la croyance engendrée par la duperie »[2]. Ainsi, par exemple, si Carlos croit par duperie de soi qu'il réussira son examen de permis de conduire en dépit de tous indices contraires, c'est vraisemblablement parce qu'il lui est arrivé au préalable de croire, sur la base de ces indices, qu'il ne le réussirait pas.

L'hypothèse de Davidson est que la croyance vraie initiale constitue l'élément déclencheur du processus de duperie de soi : « la pensée que *p*, ou la pensée qu'il serait rationnel de croire que *p*, constitue un motif qui fait agir *A* de manière à ce qu'il cause en lui la croyance que la négation de *p* est vraie »[3]. C'est par exemple à cause de la croyance préalable qu'il ne réussira pas son examen, et plus précisément encore à cause du caractère déplaisant de cette croyance, que Carlos sent le besoin d'induire en lui-même la croyance inverse. Bien plus, ajoute Davidson, la croyance contraire n'est pas seulement requise au départ, comme élément déclencheur, elle l'est aussi pendant le processus de duperie de soi et autant de temps que se maintiendra dans l'esprit de l'agent la croyance fausse qui la contredit : « la croyance que *p* non seulement cause une croyance en la négation de *p*, mais aussi l'étaye ». Du coup, conclut Davidson, la croyance vraie initiale n'est jamais tout à fait supprimée par la croyance fausse que la duperie engendre, puisqu'elle est précisément (et paradoxalement) le motif et la raison d'être de cette dernière :

[1] Cf. Sartre, *L'être et le néant* (1943), p. 84.
[2] Davidson, « Who is fooled ? » (1998), p. 4.
[3] Davidson, « Duperie et division » (1995), p. 56-57.

> « Quand la réalité (ou la mémoire) continue à menacer la croyance que le sujet s'induit lui-même à avoir quand il se dupe lui-même, il faut une motivation continue pour maintenir en place la pensée réconfortante. Si ceci est correct, il s'ensuit que celui qui se dupe lui-même ne peut pas se permettre d'oublier le facteur qui a en premier lieu provoqué son comportement de duperie de soi : la prépondérance des données allant à l'encontre de la croyance induite »[4].

En d'autres mots, cela veut dire que le processus de duperie de soi empêche l'agent de tenir pour vraie la croyance *p* à laquelle il croyait initialement, mais n'évacue pas pour autant cette croyance de son esprit. Au contraire, celle-ci est requise en tant que cause qui sous-tend la croyance fausse qui finit par prendre le dessus, ce qui prouve qu'elle continue d'exercer son efficacité sur le plan de la causalité mentale. « Finalement, ajoute Davidson, et c'est ce qui fait de l'action de se duper soi-même quelque chose de problématique, l'état qui motive la duperie de soi et l'état qu'elle produit *coexistent* »[5]. Le problème de cette coexistence, toutefois, est que les deux états concernés sont des croyances contradictoires. La question qui se pose inévitablement est alors : peut-on croire une chose est son contraire en même temps ?

2. Les paradoxes de la partition de l'esprit

La réponse de Davidson intervient à la fin de l'article « Paradoxes de l'irrationalité », ainsi que de « Duperie et division ». Il s'agit du postulat que l'on connaît sous la désignation de « partition de l'esprit » et que certains critiques préfèrent appeler « homoncularisme ». D'après cette hypothèse, le paradoxe des croyances contradictoires se dissout si l'on admet que notre esprit se trouve divisé en des sous-parties relativement autonomes. La duperie de soi serait, au fond, la duperie qu'*une partie* de l'esprit exerce sur une autre partie de l'esprit. Or, à supposer qu'il existe des « frontières » entre ces diverses parties de l'esprit, il est aisé de concevoir « qu'une de ces frontières [puisse] passer quelque part entre deux croyances manifestement conflictuelles »[6], les maintenant à distance l'une de l'autre et empêchant ainsi le sujet de s'apercevoir de leur contradiction. L'essentiel, observe Davidson, est que chacune

[4] *Id.*, p. 59.
[5] *Id.*, p. 57.
[6] *Id.*, p. 60.

de ces croyances demeure cohérente avec le sous-système auquel elle appartient. Mais le philosophe rejette la nécessité de postuler le caractère inconscient de l'un de ces sous-systèmes, bien qu'il reconnaisse tirer son hypothèse de la partition de l'esprit de la psychanalyse de Freud[7] : « Je ne vois pas de bonne raison de supposer que l'un des territoires doive être fermé à la conscience »[8]. En réalité, pour que le sujet néglige la contradiction entre ses deux croyances, il n'est pas nécessaire que l'une seulement soit accessible à la conscience, il suffit que « l'agent ne p[uisse] pas avoir une vision générale de l'ensemble de ses croyances sans effacer les frontières entre les territoires »[9].

Le problème, toutefois, est que l'hypothèse divisionniste entraîne à son tour des paradoxes non moins redoutables[10]. Le premier, que l'on nomme parfois le "paradoxe du refoulement", a été mis en lumière par Sartre et ne concerne que les conceptions divisionnistes qui font appel à la notion d'inconscient (Freud, Audi, Pears, Gardner)[11]. Il se laisse formuler comme suit. En principe, le sous-système inconscient de l'esprit s'efforce de duper intentionnellement le sous-système conscient dans le but d'épargner à ce dernier quelque croyance trop pénible, ou inversement dans le but lui procurer quelque croyance agréable. Seulement, le sous-système inconscient serait incapable de faire le tri entre, d'une part, les contenus qui seraient susceptibles de réjouir le sujet et, d'autre part, ceux qui lui seraient insupportables, s'il ne connaissait pas d'emblée la teneur de ses contenus. Mais il faudrait alors qu'il soit *conscient* de ces contenus tout en demeurant une structure *inconsciente*, autrement dit il faudrait qu'il se comporte comme une sorte de « conscience inconsciente », ce qui est absurde.

Un deuxième paradoxe concerne en revanche tout l'ensemble des conceptions divisionnistes. Il s'agit d'une difficulté signalée par Wittgenstein que Kenny propose d'appeler le "sophisme de l'homoncule" (*homunculus fallacy*)[12]. D'un côté, pour que l'esprit soit capable de se duper lui-même de façon intentionnelle, il faut admettre sa subdivision en parties distinctes, de telle sorte que la partie trompée ne se doute ni du savoir ni des intentions perfides de la partie trompeuse. Mais, d'un autre côté, pour ce faire on est obligé d'attribuer à chaque partie de l'esprit des attitudes propositionnelles (désirs, intentions, croyances, etc.) que l'on réserve en principe aux *in*-dividus à part

[7] Cf. Davidson, « Paradoxes de l'irrationalité » (1982), p. 40.
[8] Davidson, « Duperie et division » (1986), p. 60-61.
[9] *Id.*, p. 61.
[10] Cf. Correia, « Une conception émotionnaliste de la *self-deception* » (2007), p. 34 sq.
[11] Cf. Sartre (1943), p. 88.
[12] Kenny (1961), p. 125.

entière. Les parties de l'esprit apparaissent alors comme des homoncules qui se comportent comme autant de personnes à l'intérieur de la personne. Mais à ce moment nous n'avons plus affaire à une duperie de *soi-même par soi-même*.

3. Croyance et acceptation

Il existe cependant une manière alternative d'expliquer la coexistence en l'esprit de contenus propositionnels contradictoires. Cette solution consiste à suggérer qu'à chacun de ces contenus correspond une attitude de nature distincte. Au lieu de supposer que ce sont deux croyances qui se contredisent dans l'esprit du sujet, il s'agit d'envisager la possibilité que l'opposition ait lieu entre une croyance et une attitude propositionnelle d'un autre type. Une implication évidente de cette supposition est qu'on n'a plus besoin de faire face au paradoxe des croyances contradictoires, ce qui permet en même temps de supprimer l'exigence d'un esprit divisé.

Dans sa forme la plus sophistiquée, cette solution passe par la distinction que font certains auteurs entre la notion de *croyance* et celle d'*acceptation*[13]. Il faut préciser tout d'abord qu'il existe plusieurs versions de cette distinction[14]. Leur dénominateur commun est cependant l'idée qu'il faut distinguer deux façons d'accorder sa créance à un contenu propositionnel. La première est ce qu'on appelle à proprement parler *croire* et correspond à la façon passive et involontaire dont on acquiert en général des croyances. Il nous arrive par exemple de croire qu'il pleut dehors parce qu'on entend le bruissement des gouttelettes sur les toits, sans pour autant avoir eu l'occasion d'y songer

[13] Cf. J. Perry, « Belief and Acceptance », in P. Frech, T. Uehling and S. Weinsein (éds.), *Midwest Studies in Philosophy*, 5, 1980, p. 533-542 ; R. Stalnaker, *Inquiry*, M.I.T. Press, Cambridge Mass, 1984 ; Audi, « Self-Deception and Practical Reasoning », *Canadian Journal of Philosophy*, 19, 1989, p. 246-266 ; K. Leher, *Metamind*, Oxford, Clarendon Press, 1990 ; M. Bratman, « Practical Reasoning and Acceptance in a Context », *Mind*, 101, p. 1-15 ; J. Cohen, *An Essay on Belief And Accepting*, N.Y., Oxford U.P., 1992 ; P. Engel, « Believing, Holding True and Accepting », *Philosophical Explanations*, I, 2, 1998, p. 140-151 et surtout Engel (éd.) *Believing and Accepting*, Kluwer Academic Publishers, Dordrecht, 2000.

[14] P. Engel propose une lecture critique de ces différentes versions, et développe la sienne, dans toute une série d'articles : « Croyance, jugement et *self-deception* », *L'inactuel*, 3, 1995, p. 105-122 ; « Croyances collectives et acceptations collectives », in R. Boudon, A. Bouvier, & F. Chazel, (éds.), *Cognition et sciences sociales*, Paris, PUF, 1996 ; « Dispositions à agir et volonté de croire », in J.Proust & H. Grivois, (éds.), *Subjectivité et conscience d'agir*, Paris , PUF, 1997, p. 115-137 ; « Believing, Holding True and Accepting », *Philosophical Explanations*, I, 2, 1998, p. 140-151 ; « Dispositional Belief, Assent and Acceptance », *Dialectica*, 53, 3-4, 1999, p. 211-226. Voir aussi sa préface à l'ouvrage Engel (éd.) *Believing and Accepting*, Kluwer Academic Publishers, Dordrecht, 2000.

de façon consciente. On désigne ce type de croyances « dispositionnelles », au sens où elles peuvent demeurer tacites (ou non conscientes) et pourtant ne pas manquer de se manifester lorsque l'occasion se présente. Si par exemple il nous arrive ensuite de sortir, nous n'oublierons pas de prendre un parapluie, ce qui semble attester de la présence en notre esprit de la croyance pertinente.

Il semble cependant permis d'envisager une deuxième façon d'embrasser un certain contenu, consistant cette fois à y adhérer activement par la voie d'un décret ou d'un acte de volonté. C'est cette attitude à la fois réfléchie et volontaire qu'il est convenu de nommer *acceptation*. Comme le note Cohen, cette notion est assez proche de celle de « jugement » que Descartes employait à la fois dans le *Discours* et dans les *Méditations* pour se référer au pouvoir absolu qu'aurait notre volonté d'affirmer ou de nier les choses que notre entendement conçoit[15]. Plus exactement encore, cette notion semble se rapprocher de ce que Locke entendait par « assentiment », c'est-à-dire un « jugement [que] l'on fait sur les vérités exprimées en mots »[16], car à l'inverse de la croyance, qui peut demeurer implicite, l'acceptation entraîne « le réquisit conceptuel *a priori* que ce qui est envisagé se trouve associé à quelque type de formulation linguistique, même lorsque celle-ci n'est pas exprimée à haute voix »[17]. Or, sur la base de cette distinction, la revendication fondamentale est que l'acceptation est indépendante de la croyance, autrement dit que *pour accepter point n'est besoin de croire*. Cohen illustre cette thèse en évoquant l'exemple d'un avocat qui *croit* que son client est coupable mais *accepte* de le considérer comme étant innocent afin de mieux remplir son devoir professionnel. Cela est possible, précise-t-il, parce que « les raisons pour accepter que *p* n'ont pas toujours besoin d'être épistémiques : elles peuvent être éthiques ou de précaution »[18]. Puisque les acceptations ont un caractère volontaire, rien n'empêche en effet qu'elles s'accomplissent au détriment des données dont l'agent dispose.

Cohen et Engel mettent en évidence la fécondité dont semble faire preuve une telle distinction lorsqu'on l'applique à la compréhension des cas de duperie de soi[19]. Au lieu d'entretenir deux croyances contradictoires, comme le voulait Davidson, l'agent qui est dupe de lui-même entretient deux attitudes distinctes à l'égard de chacune des propositions en cause. Prenant comme exemple le cas du mari qui ne voit pas que sa femme le trompe en dépit de

[15] Cf. Cohen, « Belief and Acceptance » (1989), p. 370.

[16] Locke, *Essai sur l'entendement humain* (2002) p. 241.

[17] Cohen, *An Essay on Belief And Accepting* (1992), p. 12.

[18] Cohen, « Belief and Acceptance », (1989), p. 369.

[19] Voir le chapitre 5 de Cohen, *An Essay on Belief And Accepting* (1992) ; et aussi Engel, « Croyance, jugement et *self-deception* » (1995), p. 105-122.

tous les indices alarmants, Engel explique que ce mari « *accepte* consciemment (réflexivement) que *non p* (que sa femme ne le trompe pas) », alors qu'« [e]n même temps il *croit* que *p* (que sa femme le trompe) »[20]. C'est cependant l'acceptation qui l'emporte ; d'une part parce que ce type d'état est forcément réfléchi et conscient, et d'autre part parce « [l]e second état, sans être inconscient, est, d'une manière ou d'une autre, mis à l'écart, ou désactivé »[21]. En conformité avec cette distinction, on pourrait définir la duperie de soi dans les termes suivants :

> *S* se dupe lui-même relativement à *p* si et seulement si :
>
> (1) *S croit* que *p*
>
> (2) *S accepte* que non-*p*

L'intérêt de cette distinction est qu'elle permet d'expliquer le conflit de l'agent avec lui même dans le processus de duperie de soi tout en évitant l'écueil des croyances contradictoires (paradoxe doxastique). En effet, il n'y a de paradoxe que si l'on considère que le sujet qui croit que *p* à un moment donné parvient, d'une façon ou d'une autre, à induire en son esprit la croyance que non-*p*. Mais si l'on suppose que le conflit en question se produit, non pas entre deux croyances, mais entre une croyance et un acte d'acceptation, on ne pourra plus parler à ce moment de contradiction proprement dite, puisque ce n'est pas sous la même attitude que l'individu embrasse les contenus opposés. L'idée est donc qu'il existe deux manières d'assentir à ce que revendique un certain contenu propositionnel : soit en y adhérant de façon à la fois passive et involontaire, comme lorsqu'on acquiert la croyance qu'il pleut simplement parce qu'on voit la pluie tomber – et c'est cela qu'on appelle *croire* à proprement parler ; soit, au rebours, en y accordant sa créance de façon à la fois active et volontaire, comme par exemple lorsque quelqu'un accepte de considérer que la personne dont il est amoureux ne l'aime pas en retour (alors qu'il n'en sait rien) simplement pour éviter une nouvelle déception amoureuse – et c'est cela qu'il convient d'appeler une *acceptation*.

Nous verrons que cette distinction ne va pas sans soulever quelques difficultés, que Engel décèle et développe lui-même dans sa préface à *Believing and Accepting* et sur lesquelles nous aurons l'occasion de nous arrêter plus en avant[22]. Avant d'y venir, cependant, il convient d'examiner en détail les solutions proposées par Cohen et Engel.

---

[20] Engel, « Croyance, jugement et *self-deception* » (1995), p. 120.
[21] *Ibid.*
[22] Cf. Engel (éd.), *Believing and Accepting* (2000), p. 11 sq.

4. Cohen : la « désactivation » intentionnelle des croyances

La distinction que propose Cohen entre la notion de « croyance » et celle d'« acceptation » recoupe au fond la distinction classique (bien que pas toujours explicite) que font fait nombre de philosophes entre la notion croyance et celle de « jugement ». Cohen souligne lui-même que ce qu'il entend par « acceptation » se rapproche beaucoup de ce que Descartes entendait par « jugement », en ce sens notamment qu'il s'agit d'une attitude cognitive qui dépend d'un décret de la volonté[23]. L'élève de La Flèche prétendait qu'il incombe à notre volonté d'accorder ou pas son assentiment aux idées que notre entendement ne fait que concevoir. Cohen, de façon similaire, soutient que « l'acceptation se produit à volonté parce que, au fond, elle réalise un choix »[24]. En ce sens, donc, Cohen s'aligne sur la position « volitioniste » qui veut que notre volonté ait un certain pouvoir sur nos jugements.

Mais d'un autre côté, Cohen se rallie très ouvertement au point de vue « évidentialiste » (ou anti-volitioniste) en ce qui concerne la notion de *croyance*. A l'instar de Hume, cette fois, Cohen soutient qu'« en principe la croyance ne se forme pas à volonté parce qu'elle est causée dans chaque type de cas par quelque chose d'indépendant de la volonté immédiate du sujet »[25]. Impossible donc de décider de croire qu'il fera beau demain, ou que l'avion dans lequel on monte est parfaitement fiable – ce genre de chose, dit-il, on ne peut que décider de l'*accepter*. Cohen semble donc éviter à la fois le volitionisme et l'évidentialisme stricts, mais en toute rigueur il faudrait dire plutôt que sa solution a justement le mérite de surmonter le débat posé en ces termes antagonistes, faisant voir que derrière les différents types d'assentiment il existe peut-être des attitudes mentales corrélatives d'une nature distincte. Reste à savoir ce qui permet de caractériser les notions de croyance et d'acceptation en tant qu'attitudes *sui generis* irréductibles l'une à l'autre.

« La croyance que *p* », précise-t-il, « est une disposition à sentir qu'il est vrai que *p* »[26], disposition qui peut d'ailleurs envelopper plusieurs formes de sentiment : le sentiment de *conviction* que *p* est vrai, lorsque les données suggèrent clairement que *p*, le sentiment de *surprise* que *p*, lorsqu'au contraire les données semblaient suggérer l'invraisemblance de *p*, le sentiment de *réjouis-*

[23]Cf. Cohen, « Belief and Acceptance » (1989), p. 368 et 370. Engel rapproche pour sa part la notion d'acceptation de la notion stoïcienne de *syncatathesis*, qui était comprise également comme un jugement soumis au contrôle de la volonté (cf. Engel, éd., *Believing and Accepting*, *op. cit.*, p. 4).

[24] Cohen (1989), *op. cit.*, p. 370.

[25] *Ibid.*

[26] Cohen, *Id.*, p. 368.

sance envers *p*, et ainsi de suite. Là encore, la conception de la croyance de Cohen se rapproche indéniablement de celle de Hume, pour qui également « la croyance consiste uniquement à *éprouver un certain sentiment*, consiste en quelque chose qui ne dépend pas de la volonté mais naît forcément de certaines causes et de certains principes déterminés dont nous ne sommes pas maîtres »[27]. Pour Cohen, en tout cas, la croyance apparaît ainsi comme une disposition involontaire à sentir qu'une chose est vraie, une disposition qui n'a même pas besoin de s'exprimer en mots pour exister en tant que telle.

L'acceptation que *p*, par contraste, correspond à l'acte mental volontaire par lequel le sujet s'engage à « adopter la politique de juger, soutenir ou postuler que *p* – c'est-à-dire à inclure cette proposition ou règle dans ses prémisses lorsqu'on décide ce qu'il faut faire ou penser dans un contexte particulier, qu'il sente ou pas qu'il est vrai que *p* »[28]. L'acceptation n'est donc pas affaire de sentiment mais d'engagement et action, et c'est la raison pour laquelle Cohen affirme que nous sommes uniquement responsables de ce que nous acceptons, et nullement de ce que nous croyons[29]. Aussi, on voit mal comment un tel compromis pourrait se passer d'une formulation langagière susceptible en quelque sorte de sceller l'acte d'engagement. Contrairement à la croyance (qui peut demeurer inexprimée), l'acceptation requiert donc « que ce qui est envisagé se trouve associé à quelque type de formulation linguistique, même lorsque celle-ci n'est pas exprimée à haute voix »[30]. Il est vrai cependant que le plus souvent la croyance et l'acceptation vont de pair, car les raisons que nous avons d'accepter une idée sont aussi, en général, de bonnes raisons de croire qu'elle est vraie. Cohen illustre cet aspect en faisant allusion au prosélytisme de Pascal relativement à la doctrine catholique : ce qui amène le sujet à accepter dans un premier temps la vérité révélée devrait aussi, en principe, l'amener dans un second temps à croire véritablement en cette réalité[31].

Au chapitre V de *Believing and Accepting*, Cohen tente de montrer comment le pair conceptuel croyance-acceptation permet d'élucider le phénomène de duperie de soi. Selon lui, la clé pour comprendre la « structure conceptuelle » de la duperie de soi consiste à envisager ce phénomène comme un cas particu-

[27] Hume, *Traité de la nature humaine*, trad. fr. P. Baranger et P. Saltel, Gallimard, Paris, Livre III, Appendice, p. 372.

[28] Cohen, *An Essay on Belief And Accepting* (1992), p. 4.

[29] Cf. Cohen, *Id.*, p. 23 : « Nous pouvons donc conclure, à strictement parler, que les personnes sont responsables et doivent répondre de ce qu'elles acceptent ou négligent d'accepter, et pas pour ce qu'elles croient ou négligent de croire ».

[30] Cohen, *Id.*, p. 12.

[31] Cf. Cohen, *id.*, p. 18.

lier du décalage entre la croyance et l'acceptation à l'égard d'une même réalité. « Concrètement, écrit-il, la pensée supprimée est typiquement une croyance que non-*p*, alors que la personne se dupe elle-même en trouvant des raisons pour accepter que *p* (ou vice versa). Une même personne peut donc avoir, simultanément, des états mentaux opposés à propos de la proposition que *p* »[32]. Prenons un autre exemple de Cohen, celui de Paul, un homme qui croit que son amoureuse acceptera sa proposition de mariage, à l'encontre de ce qu'indiquent clairement les données dont il dispose. Au fond de lui, Paul doit tout de même savoir (ou croire) que sa proposition de mariage sera refusée, mais il préfère accepter de prendre pour vraie la possibilité contraire, le cas de figure où la femme qu'il aime lui dira « oui », simplement parce que « c'est cela qu'il aimerait qu'il advienne »[33].

L'avantage de cette nuance, ainsi que nous l'avions remarqué, est qu'il n'est plus nécessaire de supposer l'existence en l'esprit de deux croyances contradictoires. Engel insiste aussi sur ce point : « il n'y a pas de contradiction entre deux croyances incompatibles : il y a bien deux contenus de croyances incompatibles, mais ils ne sont pas les objets de deux états identiques, des croyances : l'un est l'objet d'une croyance, l'autre d'une acceptation »[34]. Du même coup, il devient inutile de supposer que l'esprit est divisé et que différentes parties de l'esprit peuvent croire des choses contraires. On échappe donc au paradoxe de l'homoncule, comme le note cette fois Cohen : « Nul besoin donc de supposer qu'il existe deux mannequins qui interagissent l'un avec l'autre. Le sujet de tous les prédicats mentaux est une seule personne »[35]. Enfin, ajoute ce dernier, cela permet de comprendre pour quelle raison les enfants en bas âge et les animaux sont incapables de se duper eux-mêmes : « ils sont seulement capables de croire, mais pas d'accepter »[36], ce qu'il les empêche de réussir à accepter le contraire de ce qu'ils croient.

Reste à savoir si cette analyse résiste à ce que Mele appelle le « paradoxe de la stratégie » :

> « En principe, *A* ne peut pas employer une stratégie de duperie à l'encontre de *B* si *B* connaît l'intention et le plan de *A*. Cela semble plausible également lorsque *A* et *B* sont la même personne.

[32] Cohen, *An Essay on Belief And Accepting* (1992), p. 142.
[33] *Id.*, p. 133.
[34] Engel, « Croyance, jugement et *self-deception* », *L'inactuel*, 3, 1995, p. 116. Cf. Aussi Cohen, *Id.*, p. 149.
[35] Cohen, *id.*, p. 142.
[36] *Ibid.*

> Une connaissance éventuelle par celui qui se dupe lui-même de sa propre intention et de sa propre stratégie semblerait en principe rendre celles-ci inefficaces »[37].

En réalité, même s'il n'est pas logiquement paradoxal d'accepter le contraire de ce que l'on croit, cela demeure tout de même quelque peu étrange. En principe, quelqu'un qui croit que son amoureuse refusera toute éventuelle demande en mariage ne cherchera pas à accepter comme vrai le cas de figure opposé, et à agir en conformité avec cette acceptation, quand bien même il le souhaiterait très intensément. La question est donc de savoir s'il est psychologiquement (et non plus logiquement) possible d'accepter sans réserves une réalité illusoire, tout en sachant qu'il s'agit d'une simple illusion.

Cohen semble prévoir la difficulté, puisqu'il stipule que dans les cas de duperie de soi, à la différence de ce qui se passe dans d'autres cas de décalage entre croyance et acceptation, l'agent qui accepte que *p* ne peut pas en même temps avoir conscience du fait qu'il croit que non-*p* : « la personne qui se dupe elle-même accepte consciemment que *p*, mais ne croit pas *consciemment* que non-*p* »[38]. Mais comment est-ce possible ? Sommes-nous reconduits à une hypothèse de type freudien, avec une instance inconsciente qui connaît la vérité et qui dupe sournoisement l'instance consciente ?

Cohen s'empresse de préciser qu'il n'en est rien, qu'il ne s'agit pas de renvoyer la croyance que non-*p* à un quelconque sous-système séparé, mais tout simplement de la sortir de son esprit (*put it out of the mind*) : « D'une manière ou d'une autre, [la personne qui se dupe elle-même] doit sortir la croyance que non-*p* de son esprit »[39]. Certes, concède l'auteur, nous ne pouvons pas nous débarrasser d'une croyance à dessein, pas plus que nous ne pouvons en acquérir une à dessein, étant donné que les croyances sont involontaires. Mais il ne s'agit pas de dire que le sujet « perd » sa croyance, seulement qu'il la « supprime » ou la « désactive » :

> « Ce qui se passe est que la croyance se trouve d'une manière ou d'une autre supprimée [*suppressed*], et pas éliminée. Elle demeure en l'esprit, mais pas dans la conscience. Et cela est possible parce que la croyance est une disposition – plus exactement, une disposition à sentir qu'il est vrai que ceci ou que cela – et les dispositions

[37] Mele, « Two Paradoxes of Self-Deception » (1998), p. 38.
[38] Cohen (1992), *op. cit.*, p. 142.
[39] *Ibid.*

> peuvent continuer à exister pendant qu'elles se trouvent désactivées [*unactivated*], exactement de la même manière qu'un morceau de caoutchouc demeure flexible même quand il n'est pas plié »[40].

Cohen invoque donc la dimension *dispositionnelle* des croyances pour expliquer le fait qu'elles puissent se trouver mises à l'écart de la conscience du sujet sans pour autant disparaître mystérieusement dans le tréfonds d'un inconscient[41]. Le problème, cependant, est que Cohen prétend de surcroît que le sujet a un contrôle *direct* sur l'activation et la désactivation de ses croyances, ce qui nous ramène au problème du paradoxe de la stratégie (même s'il est vrai que cela nous éloigne en définitive des écueils de l'hypothèse freudienne). En effet, comme Cohen le précise dans le passage qui suit, la désactivation des croyances indésirables peut se réaliser de façon *intentionnelle*, grâce à une aptitude de contrôle qui se passe de l'entremise d'un quelconque artifice indirect : « Ainsi, pour autant qu'une personne puisse contrôler l'activation ou la désactivation d'une croyance-disposition, elle peut être dite capable de maintenir la croyance intentionnellement ou non intentionnellement hors de son esprit – c'est-à-dire hors de sa conscience – tout en conservant la croyance elle-même »[42].

Or, bien que Cohen ne prétende pas, ainsi que le font les partisans du « volitionisme direct », que l'on peut choisir à volonté le contenu de ses croyances[43], mais seulement que l'on peut choisir d'avoir présent à l'esprit ou d'« oublier » temporairement ce que l'on croit déjà, il n'en reste pas moins que sa thèse présente précisément le même type d'insuffisance qui caractérise le volitionisme direct. De fait, il semble très douteux que l'on puisse à tout moment décider de ne plus songer à une certaine croyance (ou inversement, d'en révoquer une autre) à son gré, par un simple décret de la volonté, comme s'il y avait des interrupteurs dans l'esprit susceptibles de couper le circuit à certaines attitudes et de plonger dans l'obscurité leurs contenus respectifs. Nous en avons pour preuve toutes les pensées et mauvais souvenirs qui nous hantent inlassablement sans que nous parvenions à nous en défaire. Si Raskolnikov, le héros de *Crime et châtiment*, avait pu « désactiver » la croyance qu'il avait commis un terrible meurtre, il n'aurait sans doute pas ruminé son remords aussi longtemps et fini par prendre lui-même l'initiative d'avouer son forfait.

---

[40] *Id.*, p. 143.

[41] Pour un examen approfondi de la conception dispositionnelle des croyances, voir Engel, « Dispositions à agir et volonté de croire » (1997), p. 115-137.

[42] Cohen, *ibid.*

[43] Pour Cohen, seules nos *acceptations* sont volontaires en ce sens (et nullement nos croyances).

D'autre part, demeure aussi sans réponse la question de savoir comment nous pourrions faire pour « désactiver » intentionnellement une certaine croyance, sachant que la mémoire n'est pas toujours aux ordres de la volonté. Certes, on peut s'efforcer de chasser une idée de l'esprit en faisant précisément le contraire de ce qu'on aurait fait pour l'évoquer : se concentrer sur des objets qui n'ont aucun rapport avec cette idée, s'adonner à une activité très absorbante, et ainsi de suite. Mais cette forme de contrôle est très peu fiable, en particulier lorsque la croyance qu'on cherche à oublier n'est pas indifférente (ou « froide ») du point de vue affectif, mais se trouve associée à une certaine charge émotionnelle, comme il arrive souvent dans les cas de duperie de soi. Raskolnikov n'aurait pas pu décider d'oublier son crime aussi aisément qu'il aurait pu, par exemple, décider d'oublier l'endroit où il avait caché son butin.

En outre, on peut remarquer que cette théorie de la désactivation des croyances s'accorde mal avec la conception de la notion de croyance en termes de « *feeling-disposition* ». En effet, Cohen définit la croyance que *p* comme correspondant au fait de « sentir qu'il est vrai que *p* » (*to feel it true that* p). La croyance est donc une disposition, mais uniquement une *disposition à sentir*[44], et non point une *disposition à l'action* ou une habitude, comme le prétend une tradition qui remonte à Peirce et à Ramsey : « la croyance est une disposition à sentir certaines choses, pas une disposition à parler ou agir de certaines manières »[45]. Le problème est que, contrairement aux croyances, qui peuvent être conservées en mémoire même quand on n'y pense pas, il semblerait que les sentiments aient besoin d'être sentis ou éprouvés pour exister réellement. C'est la raison pour laquelle Freud hésitait lui-même à parler de *sentiments inconscients*, se résignant à reconnaître qu'« il est de l'essence d'un sentiment d'être perçu, donc d'être connu de la conscience »[46]. Il est vrai que Cohen prend le soin de distinguer les sentiments-de-croyance, qui sont orientés vers la perception de la vérité et de la fausseté (*credal feelings*), des sentiments qui ont trait à l'expérience de ce qui est Bien ou Mal (*affective mental feelings*)[47], qui correspondent aux émotions au sens classique. Mais il n'élucide pas la question du statut du sentiment-de-croyance à l'état « désactivé ». Certes, d'aucuns soutiennent que certains sentiments peuvent être dispositionnels au même titre que les croyances (une culpabilité récurrente, par exemple), et peut-être faudrait-il inclure les « *credal feelings* » dont parle Cohen parmi ces sentiments. Mais la question de la « désactivation » des contenus indésirables devient

[44] Cf. Cohen, *id.*, p. 5.
[45] *Id.*, p. 21.
[46] Cf. Freud, « l'inconscient » (1915), p. 82.
[47] Cf. Cohen, *id.*, p. 11.

alors d'autant plus délicate, étant donné que notre volonté a encore moins d'emprise sur ce que nous sentons que sur ce que nous croyons. Sans compter que cette manière d'envisager les croyances comme étant des sortes de sentiments se heurte à l'objection mise en évidence par Ramsey, qui attire notre attention sur le fait que « les croyances auxquelles on tient le plus fortement ne sont souvent accompagnées d'aucun sentiment que ce soit »[48].

5. Engel : la conception dispositionnelle-fonctionnaliste

L'explication de la duperie de soi que propose Engel semble assez proche, à première vue, de ce que suggère Cohen : à l'instar de ce dernier, Engel considère que l'agent qui se dupe lui-même possède une croyance et une acceptation opposées à propos d'un même contenu propositionnel. « Celui qui s'aveugle volontairement [...] *accepte* consciemment (réflexivement) que *non p* », alors qu'« [e]n même temps il *croit* que *p* »[49]. Seulement, à la différence de l'analyse de Cohen, celle d'Engel semble à même de surmonter les difficultés que nous venons d'aborder, dans la mesure où elle se fonde sur une conception foncièrement distincte non seulement de la notion de « croyance », mais aussi du statut du mécanisme de « désactivation » (ou suppression) qui est à l'œuvre dans la duperie de soi.

En ce qui concerne la notion de croyance, tout d'abord, Engel avoue « ne pas comprendre ce que [l'expression "sentir que *p* est vrai"] signifie au juste »[50] et signale à son tour la tension qui résulte du fait de vouloir définir la croyance en termes de sentiment tout en revendiquant l'aptitude dispositionnelle de cette notion à demeurer tacite : « [Cohen] dit que la croyance n'est pas une disposition ou une habitude, mais un état mental ou sentiment, lequel – à la différence d'une disposition – est en principe conscient »[51]. Mais Engel ne prétend pas non plus que les croyances sont uniquement des dispositions à agir, comme le préconisent Peirce et Ramsey, pour lesquels la croyance est une espèce d'habitude[52]. En effet, la conception purement dispositionnelle a justement beaucoup mal à rendre compte du caractère tacite ou implicite qui semble caractériser certaines croyances, puisqu'elle revendique qu'on ne doit

[48] Ramsey, « Vérité et probabilité » (2003), p. 163.

[49] Engel, « Croyance, jugement et *self-deception* » (1995), p. 120.

[50] Engel, « Believing, Holding True and Accepting » (1998), p. 145.

[51] *Ibid.*

[52] Cf. Peirce, « Comment se fixe la croyance », (1877), p. 276 ; et aussi Ramsey, « Vérité et probabilité » (1926), p. 164.

reconnaître une croyance qu'à ses fruits comportementaux. Mais surtout, souligne Engel, définir la croyance comme une disposition à l'action empêche de voir en pleine lumière le lien qui semble exister entre les croyances et d'autres états mentaux, notamment les désirs, qui jouent un rôle tout aussi déterminant dans la causation de l'action[53].

Il propose donc, non pas d'abandonner la conception dispositionnelle, mais plus exactement de la remanier de façon à y intégrer ces aspects-clés de la notion de croyance. La nouvelle théorie sera le fruit de l'articulation de cette conception dispositionnelle de la croyance avec, d'un autre côté, la conception dite « fonctionnaliste » des *états mentaux* que partagent des philosophes contemporains comme Fodor, Putnam, Lewis et Dennett[54]. Selon cette conception, les états mentaux ne se définissent ni comme les attributs d'une substance pensante (dualisme), ni comme de simples états du cerveau (physicalisme), ni comme des dispositions purement comportementales (béhaviourisme), mais comme des états qui jouent un rôle causal (ou fonctionnel) au sein d'un système, et qui entretiennent des relations de type causal non seulement entre eux, mais aussi avec les « entrées d'information » (*sensory input*) et les « sorties comportementales » (*behavioural output*) de ce système[55].

De ce mariage naît ce que Engel appelle la théorie « dispositionnelle-fonctionnaliste » de la croyance, laquelle « définit comme croyance *tout état qui occupe un rôle fonctionnel déterminé dans la production de certaines actions et de certains états mentaux* »[56]. Loin d'apparaître comme de simples dispositions à l'action, les croyances peuvent alors manifester, en amont de leurs effets comportementaux, toute la complexité de leurs rapports avec les autres états mentaux susceptibles de causer une action – et en particulier avec les désirs. Après tout, la croyance qu'il pleut dehors ne suffit pas, à elle seule, pour expliquer notre action de prendre un parapluie en sortant : encore faut-il supposer que nous ayons le désir de rester secs. Et réciproquement, si nous sortons dans la rue munis d'un parapluie, notre action de prendre un parapluie ne peut pas s'expliquer uniquement à la lumière de notre désir de rester secs, puisqu'en l'absence de la croyance corrélative qu'il pleut dehors - ainsi que de tout un réseau d'autres croyances plus ou moins implicites (que la pluie mouille, que ce parapluie nous appartient, etc.) – ce désir aurait été pour ainsi dire « aveugle »

[53] Cf. Engel, « Les croyances » (1995), tome II, p. 29 sq.

[54] Pour un examen détaillé de la conception fonctionnaliste dans chacune de ses versions, voir P. Engel, *Introduction à la philosophie de l'esprit* (1994), p. 30-34 et aussi le ch. 8.

[55] C'est à peu près la definition que donne W. G. Lycan dans « Functionalism » (1995), p. 317-323.

[56] Engel, *id.*, p. 31.

aux contraintes de la réalité. Il faut donc définir la croyance à la fois en fonction du désir et en rapport avec l'action : « C'est cette idée qu'incorpore la conception selon laquelle croyances et désirs sont des états *fonctionnels* : une croyance conduit à des actions en fonction de désirs, et un désir conduit à des actions en fonction de croyances, et croyances et désirs sont fonctions les uns des autres »[57].

Concernant le sujet qui nous occupe, cette conception a d'abord l'avantage d'expliquer pour quelle raison les croyances peuvent demeurer implicites (ou tacites) sans perdre pour autant leur statut de croyances à part entière. Cohen rejetait cette possibilité – à cause vraisemblablement de sa définition de la croyance en termes de sentiment –, estimant que « la croyance purement implicite que *p* n'est pas plus une sorte de croyance que *p* qu'un soldat de plomb n'est une sorte de soldat »[58]. D'où le statut ambigu des croyances « désactivées » telles qu'il les décrivait. Mais Engel peut faire valoir, pour sa part, que les croyances implicites demeurent tout de même des croyances proprement dites, dans la mesure où les croyances ne sont pas des sentiments qui aient besoin d'être *éprouvés* par une sensation consciente pour être réels :

> « Les sensations et les expériences sont des états conscients : je ne peux pas, du moins dans les cas normaux, avoir une sensation ou une expérience sans être conscient de cette sensation ou de cette expérience, alors que je peux croire *que Paris est plus grand que Caen* ou *que les amanites sont vénéneuses* sans y penser explicitement ou être conscient de ces croyances »[59].

Mais surtout, la conception dispositionnelle-fonctionnaliste a le mérite de rendre intelligible le processus d'« activation » et de « désactivation » des croyances que la réflexion de Cohen laissait inexpliqué. Ce dernier prétendait en effet que « nous contrôlons très souvent l'activation ou la désactivation de nos croyances »[60] de façon pleinement intentionnelle, non seulement dans le contexte de la duperie de soi, mais plus généralement lorsque nous décidons de ne plus penser à telle idée ou, inversement, lorsque nous décidons de rappeler telle autre idée à la conscience. A ses yeux, donc, la volonté aurait non seulement le pouvoir de déterminer le contenu de nos acceptations, mais aussi de rendre actives ou inactives nos croyances (faute de pouvoir déterminer également le contenu de ces dernières). Nous avons cependant constaté que cette

[57] Engel, « Dispositions à agir et volonté de croire » (1997), p. 119.
[58] Cohen, *An Essay on Belief And Accepting* (1992), p. 32.
[59] Engel, « Les croyances » (1996), p. 23.
[60] Cohen, *id.*, p. 143.

suggestion repose en fait sur une pétition de principe, d'une part parce que Cohen n'avance aucun argument capable de justifier une telle aptitude, et d'autre part parce que les exemples qu'il évoque pour l'illustrer ne sont guère concluants : ce n'est pas parce qu'il m'arrive parfois de réussir à « sortir de ma tête » une croyance indésirable que j'ai pour autant le pouvoir de « désactiver » à mon gré n'importe laquelle de mes croyances.

Or, à la différence de Cohen, Engel semble reconnaître que le processus moyennant lequel l'individu qui se dupe lui-même *cesse de croire* consciemment le contraire de ce qu'il a l'intention consciente d'accepter n'est pas un processus intentionnel et délibéré, mais au contraire un processus « sub-intentionnel » :

> « J'admets que le processus qui conduit à la subdivision entre ce que le sujet croit et ce qu'il accepte est *sub-intentionnel* ou *cognitif*. Il ne se laisse pas décrire à la manière ordinaire dont nous décrivons des processus causaux naturels (par exemple neuronaux), mais il est cependant causal »[61].

Si notre lecture est correcte, cela veut dire que ce n'est pas l'intention ou un acte de volonté qui de façon pleinement consciente met à l'écart (ou désactive) la croyance indésirable. Il s'agit plutôt d'un processus causal dont le sujet n'est pas forcément conscient – bien qu'il ne s'agisse pas non plus d'un processus inconscient au sens freudien – et qui n'est pas forcément déclenché par des raisons de croire tout à fait réfléchies. La solution d'Engel semble se rapprocher sur ce point de celle de Davidson, qui insiste également sur la nécessité de reconnaître la dimension causale sous-jacente au processus de duperie de soi, sous peine de tomber dans une explication purement intellectualiste de l'irrationalité (qui aurait la propriété paradoxale de supprimer son objet en le rendant parfaitement intelligible).

Rappelons en effet que, pour Davidson, ce qui cause l'irrationalité cognitive est une cause qui n'est pas une raison de croire légitime. Bien qu'il faille sans doute que l'agent irrationnel ait l'intention explicite d'adopter la conviction que non-*p* à l'encontre de ce qui lui semble vrai – et en ce sens Cohen a raison de dire que le processus est intentionnel - il faut aussi reconnaître que d'autres états mentaux peuvent contribuer de manière à la fois non intentionnelle et non consciente à son acceptation de non-*p* – et en ce sens Cohen a tort de laisser entendre que le processus est *purement* intentionnel. En particulier, Davidson et Engel soulignent le rôle causal que joue la croyance initiale que

[61] Engel, « Croyance, jugement et *self-deception* » (1995), p. 121.

p dans l'avènement du désir de non-*p* qui, à sont tour, contribuera en grande partie à la formation de l'intention consciente de se persuader que non-*p*. C'est d'ailleurs précisément pour cette raison que cette croyance de départ doit demeurer dans l'esprit du sujet sous une forme ou une autre, puisqu'en son absence le support causal de la duperie de soi disparaîtrait. « La croyance que *p*, écrit Davidson, non seulement cause une croyance en la négation de *p*, mais aussi l'étaye »[62].

Mais on comprend du même coup pour quelle raison cette solution demeure irréductible à l'hypothèse freudienne : l'état mental de départ « n'est pas inconscient, explique Engel, car s'il était totalement non accessible au sujet, on ne pourrait pas dire que la croyance que *p* a provoqué le désir que non-*p* »[63]. En réalité, il n'y a un sens à désirer croire que non-*p* (par exemple, que je ne suis pas en train de perdre mes cheveux) que si je me doute un tant soit peu du fait que *p* (que je suis en train de perdre mes cheveux), puisqu'en principe on ne désire pas l'avènement de quelque chose qui est déjà une réalité. En outre, il faut noter qu'en toute rigueur Freud ne parle jamais de « croyances inconscientes », bien que des explications de type freudien n'hésitent pas à le faire (Audi, Fingarette, Pears, Gardner, etc.). Pour Freud, le véritable contenu refoulé n'est pas une croyance mais ce qu'il appelle le « représentant de la pulsion » (*Triebrepräsentanz*), qui n'a pas la propriété d'être intentionnel car il ne représente pas la pulsion au sens où un portrait représente une personne, mais au sens où un élu représente le peuple, c'est-à-dire au sens où il se trouve investi de certaines propriétés de la pulsion (quantum d'affect ou « énergie d'investissement »), étant donné que c'est à lui que la pulsion vient « se fixer » à la suite du refoulement[64]. Or, Engel rejette précisément ce « modèle freudien [qui] postule, pour la *self-deception*, quelque chose comme un système de causes inconscientes qui n'est à aucun moment marqué du sceau de l'intentionnalité »[65].

Enfin, la désactivation de la croyance désagréable s'opère d'après Engel en fonction des interactions causales qui se produisent entre les états mentaux pris individuellement, et non à cause de l'activité refoulante de quelque sous-système autonome et indépendant de la conscience, « car il n'y a pas d'instance qui enterre la croyance dans un inconscient »[66]. L'auteur n'écarte pas la possibilité d'un clivage au sein de l'esprit, mais précise qu'un tel clivage

[62] Davidson, « Duperie et division » (1986), p. 57.
[63] Engel, « Croyance, jugement et *self-deception* » (1995), p. 120.
[64] Cf. Freud, « Le refoulement » (1905), p. 54 sq.
[65] Engel, *id.*, p. 121.
[66] *Id.*, p. 119.

n'aurait de sens que si les états non conscients demeuraient potentiellement accessibles à la conscience. D'après lui, un tel modèle de la division de l'esprit se rapprocherait peut-être du clivage que décèle Freud, non plus entre les constituants de l'esprit (Moi, Ça, Surmoi), mais au sein du Moi lui-même (*Ichspaltung*). En réalité, à la différence de ce qui se passait avec les modèles topiques, qui étaient fondés sur la notion de refoulement, selon cette nouvelle explication des conflits psychiques (qui ne fut avancée par Freud qu'assez tardivement) il n'y a pas une mise à l'écart de certains contenus au profit de revendications plus impératives ou menaçantes, mais une *coexistence de deux attitudes antagonistes* qui peuvent « persiste[r] côte à côte tout au long de la vie sans s'influencer mutuellement »[67] : « l'une qui tient compte de la réalité, l'attitude normale, l'autre qui, sous l'influence des pulsions, détache le moi de la réalité »[68]. Selon Freud cette explication permet d'élucider notamment ce qui se passe dans la tête des individus fétichistes et psychotiques, chez lesquels le principe de plaisir prévaut à tel point sur le principe de réalité qu'ils sont capables de nier les choses les plus indéniables et d'affirmer les choses les plus invraisemblables. Mais rien n'empêche que les croyances plus conformes à la réalité, qui sont en quelque sorte réprimées, redeviennent conscientes au cours par exemple d'une thérapie et que l'individu retrouve une vision plus lucide de la réalité.

Or, comme Engel le fait remarquer, ce nouveau modèle du conflit psychique – que Freud n'a pourtant pas suffisamment développé et que ses critiques ont souvent négligé – présente plusieurs avantages par rapport au modèle du refoulement et du retour du refoulé. Premièrement, on ne se réfère qu'au Moi et à lui seul, ce qui permet de faire l'économie de l'hypothèse divisionniste dans sa version forte. Deuxièmement, les contenus réprimés ne sont pas des « forces pulsionnelles » aveugles, mais des états intentionnels doués d'un contenu propositionnel (chez le fétichiste, par exemple, c'est la croyance que les femmes n'ont pas de pénis qui, selon Freud, se trouve réprimée). Et troisièmement, les états en question ne sont pas « refoulés » et inaccessibles, mais seulement tacites et pouvant donc redevenir conscients à tout moment. « On peut être ici en désaccord avec le *contenu* de l'explication freudienne, écrit Engel, mais on doit être d'accord avec sa *forme* : la distinction, à l'intérieur du moi, de deux types de processus, certes inconscients, mais qui se manifestent au niveau du moi, donc d'états qui ne sont pas *sub-intentionnels*, mais inten-

[67] Freud, *Abrégé de psychanalyse* (1950), p. 79.
[68] *Id.*, p. 78.

tionnels »[69].

6. Acceptation ou simulation ?

Seulement, ces précisions ne concernent que le premier volet de la solution d'Engel, puisque d'après lui il ne suffit pas que la croyance indésirable soit désactivée : encore faut-il que la croyance contraire soit *acceptée* par un acte conscient et réfléchi. Tel est le réquisit « positif » du processus de duperie de soi tel que Engel le décrit. Sur ce point, son analyse semble se rallier à celle de Cohen, hormis qu'à ses yeux « l'acceptation est beaucoup plus intimement liée à la croyance que Cohen et d'autres ne le pensent »[70], en particulier dans la mesure où la croyance figure parmi les principales causes (ou raisons) des actes d'acceptation, ensemble avec les désirs et les autres motifs.

Or, c'est justement cette exigence « positive » que Cohen et Engel assignent au processus de duperie de soi qui nous semble poser problème, et ceci pour plusieurs raisons. En premier lieu, on peut se demander s'il est réellement nécessaire de supposer un tel acte d'acceptation pour parvenir à se duper soi-même. Est-ce que la simple désactivation de la croyance indésirable, avec tout ce que cela implique au niveau de l'interprétation de la réalité, ne suffit pas pour expliquer les cas ordinaires de duperie de soi ? Si, par exemple, un agent en vient à croire, en dépit de tous les indices contraires, qu'il n'est pas en train de devenir chauve, est-ce qu'il faudrait *en plus* qu'il accepte par un acte conscient et réfléchi qu'il n'est pas chauve ? Ne suffirait-il pas tout simplement qu'il cesse de croire qu'il perd ses cheveux ? A quoi bon supposer de surcroît un acte d'acceptation alors que, *ex hypothesi*, l'état mental qui lui cause tant de désagrément se trouve déjà évacué de sa conscience ? Du point de vue de l'équilibre psychique qu'évoque Engel[71], il semblerait que le conflit se trouve résolu dès le stade de neutralisation de la croyance désagréable, moyennant une stratégie de défense à la fois plus simple et moins coûteuse. En ce sens, K. Bach a en partie raison de suggérer que « la duperie de soi ne concerne pas tant ce qu'on croit de façon positive mais ce qu'on évite de penser »[72]. C'est

[69] Engel, *id.*, p. 121.

[70] Engel, « Believing, Holding True and Accepting » (1998), p. 148.

[71] Cf. Engel, « Croyance, jugement et *self-deception* » (1995), p. 122 : « Mais le comportement [de l'akratique et du *self-deceiver*] est aussi une stratégie de défense, visant à équilibrer leur psychisme, et à résoudre le conflit qu'ils éprouvent ».

[72] K. Bach, « Thinking and Believing in Self-Deception » (1997), p. 105. L'auteur dénombre au moins trois stratégies auxquelles l'agent peut recourir afin d'éviter de penser à une idée : la rationalisation, le détournement de son attention et l'encombrement de sa mémoire avec pensées

aussi ce que suggèrent les interprétations de type freudien, pour lesquelles le processus de refoulement (ou un mécanisme de défense semblable), qui ne fait que tenir des contenus à l'écart de la conscience, suffit à lui seul pour expliquer l'irrationalité.

Engel et Cohen pourraient cependant rétorquer que dans d'autres cas non moins ordinaires de duperie de soi, c'est au contraire l'effort de croire ce qui semble faux qui prévaut sur l'effort corrélatif de cesser de croire ce qui semble vrai. C'est le cas, par exemple, de l'étudiant qui se persuade qu'il aura une bourse d'études en dépit de tous les signes défavorables : il n'a pas à s'efforcer de refouler la croyance selon laquelle il n'aurait aucune chance d'avoir cette bourse, pour la simple et bonne raison qu'il n'a même pas cru un seul instant à ce scénario pessimiste (pas même avant la duperie). En réalité, la duperie de soi n'implique *pas nécessairement* que le sujet se persuade de quelque chose *à l'encontre de ce qu'il croyait auparavant*, bien que cela puisse arriver dans une partie des cas, mais seulement qu'il s'en persuade *à l'encontre de ce que suggèrent les données dont il est conscient*[73]. On peut donc se duper soi-même sans avoir à refouler ou à « désactiver » quoi que ce soit, bien que dans ce cas de figure la duperie de soi s'apparente beaucoup à la simple prise des désirs pour des réalités. Mais à ce moment, la question qui se pose est de savoir si l'agent qui veut embrasser la croyance invraisemblable se contente simplement d'*accepter* la réalité de l'état de choses en question, ou bien s'il y *croit* réellement.

Cela nous amène à formuler une deuxième objection, qui concerne cette fois le statut même de la notion d'acceptation. D'après les partisans de la distinction entre croyance et acceptation, la différence fondamentale entre ces deux attitudes est que l'une est involontaire (on ne saurait croire à volonté) tandis que l'autre est volontaire (on peut accepter ce qu'on veut, pourvu qu'on ait des raisons pratiques de le faire). Une première conséquence de ce cette asymétrie est que l'acceptation ne peut pas consister à donner créance à la réalité de l'état de choses dont il s'agit, autrement dit à le *tenir pour vrai* en toute sincérité, comme il arrive avec la notion de croyance (quelle qu'en soit la définition). En effet, si l'acceptation était un acte de *tenir pour vrai* – ce que Kant appelle justement *das Fürwahrhalten*[74] –, on ne pourrait pas affirmer son caractère volontaire sans se heurter aux objections dirimantes dont sont la cible les défenseurs du volitionnisme doxastique.

incompatibles avec cette idée.

[73] Cf. Correia, *La duperie de soi et le problème de l'irrationalité* (2010), p. 133.

[74] Cf. Kant, *CRP*, II, ch. 2, 3e section : « De l'opinion, du savoir et de la foi », p. 683 (III, 533).

Mais si accepter que *p* ne revient pas à accepter *la vérité* de *p*, si l'on peut accepter que *p* même quand on croit tout le contraire de *p*, en quoi consiste au juste cette attitude ? Que signifie accepter une proposition ? Engel et Cohen, avec d'autres auteurs, précisent qu'il s'agit d'un acte volontaire d'« engagement » (*commitment*) envers une certaine réalité *p*, que l'on réalise en vertu d'un certain nombre de raisons qui ne sont pas d'ordre épistémique mais pratique. C'est ainsi, par exemple, que l'avocat est à même d'*accepter* l'innocence de son client pour des raisons strictement utilitaires, alors que son évaluation épistémique de la situation le porte à *croire* que son client est coupable[75]. Mais il semble légitime de se demander alors si l'acceptation d'une certaine réalité est autre chose que le simple fait d'*agir comme si cela était vrai*. Après tout, l'avocat en question n'a pas besoin d'adhérer au contraire de ce qu'il croit pour maximiser le succès de sa défense : il lui suffit de *faire semblant* de croire le contraire de ce qu'il croit. Mais à ce moment, la seule attitude qu'il y ait dans son esprit concernant les agissements de son client est bel et bien la croyance qu'il est coupable, le reste de ses attitudes relevant plutôt de son comportement (ton du plaidoyer, affectation de la voix et des gestes, etc.). Il croit que son client est coupable mais agit comme s'il ne l'était pas parce que son métier n'est pas d'établir la vérité mais de défendre la version de la vérité proposée par son client. Nul besoin, dès lors, de supposer que l'avocat entretient de surcroît une deuxième attitude à l'égard de ce client, qui serait l'acceptation de son innocence, puisque son comportement est déjà parfaitement intelligible à la lumière de ses croyances et ses désirs seulement.

Sur ce point, cependant, les positions d'Engel et de Cohen sont quelque peu divergentes. Cohen refuse catégoriquement cette lecture : « l'acceptation de *p* n'est pas identique au fait de parler et d'agir comme si *p* était vrai »[76]. Engel, pour sa part, semble reconnaître (avec R. Stalnaker et F. Recanati[77]) que la notion d'acceptation est au moins en partie réductible au phénomène de simulation de la croyance : « Accepter que P, affirme-t-il, c'est au moins se comporter *comme si* on croyait que P, et être prêt à maintenir que P, ou à le prendre pour acquis, même quand les données ne sont pas favorables, ou sont contraires »[78]. Il peut alors faire voir de façon cohérente comment, par

[75] Cohen, « Belief and Acceptance » (1989), p. 369.

[76] Cohen, *Believing and Accepting* (1992), p. 14.

[77] Cf. R. Stalnaker, *Inquiry* (1994), p. 80 : « accepter une croyance, c'est agir, à certains égards, comme si l'on y croyait » ; et F. Recanati, « The Simulation of Belief » (2000), p. 284 sq. Pour Recanati, l'acceptation qui n'est pas accompagnée de croyance (*acceptance-without-belief*) est un cas de *simulation*.

[78] Engel, « Dispositions à agir et volonté de croire » (1997), p. 126.

exemple, un professeur est à même de donner une bonne note à un élève, et donc d'accepter que sa copie est bonne, notamment dans le but de l'encourager, quand en réalité il croit que cette copie ne mérite pas une note aussi bonne. Dans ce type de cas, en effet, il semble que l'acte d'acceptation du professeur consiste simplement en une stratégie de simulation destinée à engendrer de la confiance chez l'élève et à l'encourager pour la suite. Ce professeur n'accepte donc pas que la copie est bonne au sens fort que préconise Cohen, il n'admet pas cela *in foro interno*, pour la simple raison qu'il n'en a pas besoin : l'unique raison (pratique) de son acceptation est le désir d'encourager l'élève, et pour ce faire il lui suffit de renvoyer à ce dernier les signes d'une réelle approbation. Les raisons de son acceptation n'exigent pas qu'il s'engage à agir *en toute circonstance* en conformité avec cette possibilité, mais seulement qu'il le fasse dans un contexte précis (devant l'élève en question), et par conséquent rien ne l'empêche éventuellement de faire part de sa stratégie à l'un de ses collègues en dehors des cours, dans un contexte où il n'est plus du tout question pour lui d'accepter que la copie de son élève mérite la note qu'il lui a donné.

Un autre exemple que donne Engel de l'« agir comme si » sous-jacent à la notion d'acceptation est celui du mathématicien qui cherche à démontrer un énoncé par la technique de réduction à l'absurde : pour les besoins de la démonstration, il acceptera à titre de supposition hypothétique que la négation de cet énoncé est vraie, sachant que si cette supposition débouche sur une absurdité, la vérité de l'énoncé opposé se trouvera par là-même établie. Or, Cohen considérait de son côté que l'attitude du mathématicien ne constitue pas un acte d'acceptation véritable, mais seulement ce qu'il appelle un acte d'« assomption » (*assumption*), sous prétexte que « l'acceptation que *p* n'est en principe justifiable que si l'on a des raisons en sa faveur », tandis que l'assomption consisterait à admettre une certaine réalité simplement « en l'absence de raisons qui s'y opposent »[79]. (Et pourtant, pourrait-on objecter au passage, le mathématicien semble avoir autant de raisons valables de supposer que la négation de son énoncé est vraie que n'en a, par exemple, l'avocat pour accepter que son client est innocent). Quoi qu'il en soit, Engel admet au contraire que l'acte d'accepter que *p* puisse consister simplement à partir du principe que *p* est vrai (*take* p *for granted*) ou, ce qui revient au même, à agir comme si l'on croyait que *p* est vrai, quelles que soient par ailleurs les raisons que l'on ait de faire cette supposition. Comme il le dit lui-même : « La plupart des cas [d'acceptation] impliquent ce que nous pouvons appeler des *faire semblants de croire* (*pretendings to believe*), ou des croyances *simulées*, en général

[79] Cohen, *Believving and Accepting* (1992), p. 13.

dans des contextes sociaux, et la plupart du temps elles impliquent un certain conflit entre les normes épistémiques et les normes pratiques de la formation de la croyance »[80].

Or, cette conception plus souple de la notion d'acceptation permet certes d'éviter les ambiguïtés que nous avions décelées plus haut dans la conception de Cohen. Mais d'un autre côté, à supposer que l'acte d'acceptation soit simplement un engagement à agir comme si une certaine possibilité était vraie – ou plus simplement encore, à adopter une certaine ligne de conduite indépendamment de tout souci de vérité -, on peut se demander si cette attitude s'applique réellement au cas de duperie de soi. En effet, il semble douteux que les sujets qui se dupent eux-mêmes se contentent d'accepter une réalité invraisemblable, faisant seulement semblant d'y croire : au contraire, ils semblent croire pleinement à cette réalité. Si l'on songe par exemple au cas de la femme trompée, qui refuse d'admettre l'infidélité de son mari en dépit de tous les signes alarmants (rouge à lèvres sur le col de sa chemise, traces persistantes d'un parfum féminin, retards nombreux et injustifiés, etc.), on se rend compte que sa croyance en la fidélité de son mari n'a rien à voir avec la « croyance simulée » du mathématicien à l'égard de l'hypothèse qu'il admet seulement dans le but de prouver son absurdité, ni même avec l'assomption, par l'avocat, de l'innocence de son client.

L'avocat et le mathématicien, eux, savent pertinemment lorsqu'ils font leur supposition que ce qu'ils supposent est faux, et ils sont aussi parfaitement conscients de la stratégie qui justifie leur acte de faire semblant de croire. Mais la femme trompée n'a pas choisi de façon froide et calculée de croire à la fidélité de son mari, sous prétexte que cela vaudrait mieux pour son confort psychologique, car nous avons vu qu'il est impossible de se persuader de la vérité d'une hypothèse tout en étant conscient à la fois de sa fausseté et de la stratégie employée pour l'induire en l'esprit (paradoxe de la stratégie)[81]. Certes, il se peut qu'elle fasse seulement semblant de croire à la fidélité de son mari, par exemple parce qu'elle dépend financièrement de lui ou qu'il lui manque le courage de le quitter. Mais dans ce cas nous n'avons pas affaire à

[80] Engel, « Dispositional Belief, Assent and Acceptance » (1999), p. 221.

[81] C'est pourtant la voie qu'emprunte la réflexion de Cohen, qui prône une désactivation *intentionnelle* de la croyance vraie de départ. Mais comme le signale Engel, cette démarche est paradoxale : « selon la solution proposée, [l'agent] a "supprimé" ou désactivé, ou non activé la pensée que *p*. Mais pourquoi a-t-il fait cela ? Est-ce parce qu'il ne désirait pas l'activer ? Mais si c'est le cas, alors il faut que l'agent ait eu, devant son esprit, pour ainsi dire, la croyance que *p*, qu'il l'ait entretenue consciemment, contrairement à l'hypothèse envisagée, selon laquelle la croyance que *p* est restée *non consciente* » (Engel, « Croyance, jugement et *self-deception* », 1995, p. 117).

un cas authentique de duperie de soi. Pour que la duperie soit authentique, en effet, il faut que la personne adhère sincèrement à l'illusion qu'elle s'est fabriquée ; il faut que son attitude soit autre chose qu'un simple engagement à « agir comme si » ou à « faire semblant de croire ». D'un autre côté, on pourrait supposer que la femme en question, à la différence du mathématicien et de l'avocat, ignore la vérité de *p* (que son mari la trompe) au moment où elle accepte que non-*p*, ce qui s'accorde avec l'idée que la croyance que *p* est désactivée au cours du processus de duperie de soi. Mais à ce moment on ne comprend pas quelle raison a-t-elle d'accepter que non-*p*, étant donné que la croyance ou le soupçon que son mari la trompe était la seule raison qu'elle avait pour faire l'effort d'accepter le contraire.

Enfin, on pourrait émettre également quelques réserves à propos de l'argument que mettent en avant Cohen, Engel et aussi Recanati pour démontrer que la notion d'acceptation est irréductible à celle de croyance[82]. D'après eux, nous serions incapables d'apprendre certaines tâches et certaines règles si nous n'avions pas la possibilité d'accepter certains contenus propositionnels sans pour autant croire qu'ils sont vrais. « Par exemple, dit Engel, ma méthode d'apprentissage du Mandarin me demande de tenir pour vrai *"Wo hui shuo Fawen"*, même si je ne découvrirai que plus tard ce que cela signifie »[83]. Ce cas de figure illustre parfaitement l'ensemble de situations dans lesquelles nous devons admettre ce qu'on nous explique (a) sans savoir si c'est vrai et (b) sans même comprendre la *signification* de ce qu'on nous explique. Il faut en effet distinguer ces deux aspects, car en l'occurrence le professeur de Mandarin n'exige pas forcément que les élèves admettent (a) la vérité de proposition *"Wo hui shuo Fawen"*, mais peut-être seulement qu'ils admettent (b) que cette proposition signifie quelque chose en Mandarin. Mais supposons qu'il demande les deux choses à la fois, comme il arrive par exemple lorsqu'un professeur de mathématique écrit au tableau la formule d'un théorème à apprendre et que ses élèves le copient minutieusement dans leurs cahiers, admettant que ce théorème est doué de sens et aussi qu'il est vrai. Le plus souvent, les élèves ne comprennent pas encore le théorème au moment où ils l'acceptent, ce qui semble aller dans le sens de la thèse proposée : puisqu'ils acceptent que *p* avant même de croire que *p*, c'est qu'on peut accepter sans même croire.

Néanmoins, rien n'empêche de considérer que cette prétendue attitude d'acceptation soit en réalité une croyance ordinaire, dans la mesure où l'autorité du professeur en tant que telle est une raison de croire parfaitement légi-

[82] Cf. Cohen (1992), p. 19 ; Engel (1998), p. 141 ; et Recanati (2000), p. 272.
[83] Engel (1998), p. 141.

time. De même que nous croyons au résultat du calcul arithmétique qu'affiche la machine à calculer, parce que nous faisons confiance à la machine, aussi les élèves peuvent-ils croire ce que dit le professeur simplement parce qu'ils considèrent d'entrée de jeu qu'ils peuvent lui faire confiance. Ils peuvent se dire en toute simplicité : « Puisque le professeur l'affirme et que le manuel le confirme, c'est que cela doit être vrai », et en venir donc à croire pleinement à la véracité de l'énoncé, même si le sens de cet énoncé leur échappe (et *a fortiori* les raisons d'ordre épistémique qui permettent de le justifier). L'autorité de l'éducateur et celle du manuel suffisent pour susciter l'adhésion sans réserve des élèves au théorème qui leur a été soumis. Certes, un élève particulièrement scrupuleux (ou sceptique) pourrait se garder systématiquement de tenir une affirmation pour vraie avant d'en avoir saisi pleinement la signification et d'en avoir trouvé la justification, mais la plupart des élèves semble *de facto* admettre la vérité de ce qui est écrit dans les manuels d'étude avant même de l'avoir examiné. Songeons par exemple au plus connu des théorèmes, le théorème de Pythagore : n'importe quel élève dira avec franchise qu'il est vrai, même s'il ne sait pas encore (ou ne sait plus) en quoi il consiste, parce qu'il se fie à ce qu'en disent les livres, et à l'assentiment des adultes, et au fait qu'il n'a jamais entendu quelqu'un de sensé mettre en cause ce théorème. Ce n'est donc pas qu'ils agissent comme s'ils y croyaient, se contentant par exemple d'*accepter* une règle pour les besoins du calcul (bien que cela soit faisable aussi), c'est plutôt qu'ils y *croient* tout à fait, c'est-à-dire qu'ils estiment que le théorème est vrai sans aucune réserve.

Il faut concéder néanmoins qu'il ne va pas de soi que l'on puisse seulement croire que *p* quand on ne comprend même pas ce que *p* signifie ; mais, d'une part, cette difficulté concerne tout autant la notion d'acceptation et, d'autre part, on pourrait considérer qu'il existe non pas différentes attitudes *sui generis* à propos d'un même contenu (la croyance, l'acceptation, et pourquoi pas aussi l'assomption, la présupposition, etc.), mais plutôt *différentes modalités d'une même attitude* - en l'occurrence différents *types de croyance*, comme par exemple : (a) une croyance qui n'est pas accompagnée d'une compréhension de son objet – ce que Recanati propose d'appeler « quasi-croyance »[84] ; (b) la croyance ordinaire, qui serait comprise mais pas justifiée ; et enfin (c) la croyance accompagnée de justification - que Platon identifie à la « connaissance » (*episteme*) elle-même, lorsque de surcroît elle vraie[85]. Et l'on pourrait

[84] Cela dit, pour Recanati la croyance ainsi que la quasi-croyance demeurent des cas particuliers de l'acceptation Cf. Recanati (2000), p. 283.

[85] Cf. Platon, *Théétète*, 201c-210a.

même affiner cette distinction en suggérant, dans la lignée de Ramsey, qu'à chaque type de croyance puisse correspondre un degré de conviction variable, allant de la croyance qui est proche du doute à la plus ferme des certitudes[86]. Pour certains auteurs, d'ailleurs, c'est précisément ce type de nuance qui explique la possibilité de croire deux choses contradictoires lorsqu'on se dupe soi-même : le paradoxe doxastique serait ainsi évité parce que la croyance finale que non-*p* est « pleine » (*full-blown belief*), tandis que la croyance originelle que *p* n'est que « partielle » (*partial belief*)[87].

Quoi qu'il en soit, il semble possible de ramener à la notion de croyance les principaux attraits de la notion d'acceptation, et aussi, semble-t-il, d'élucider les paradoxes de l'irrationalité sans avoir à assigner aux agents une attitude propositionnelle essentiellement distincte de celle de croyance. Les difficultés que nous avons signalées ne suffisent sans doute pas pour conclure que la notion d'acceptation est elle-même inacceptable, même si elle semble plus pertinente et plus féconde sur le plan normatif que sur le plan descriptif, mais elles semblent suggérer en tout cas que cette notion n'est pas non plus incontournable dans le cadre de l'explication de l'irrationalité cognitive.

7. Conclusion

Contrairement à Davidson, Cohen et Engel parviennent à expliquer le phénomène de duperie de soi tout en évitant aussi bien le paradoxe doxastique que les paradoxes de la division de l'esprit. Toujours au rebours de Davidson, leurs solutions ont aussi le mérite de rendre compte du processus de « désactivation » de la croyance au moyen duquel le contenu indésirable se trouve mis à l'écart de la conscience du sujet. Nonobstant une indéniable similitude formelle, toutefois, leurs caractérisations de ce processus se sont avérées somme toute assez distinctes. Au regard de Cohen, nous avons le pouvoir de contrôler la présence ou l'absence des croyances dans le champ de la conscience, de les convoquer ou de les congédier à notre gré, faute de pouvoir déterminer volontairement leur contenu. Mais bien cela puisse effectivement arriver dans certains cas, Cohen ne réussit pas à montrer comment un tel contrôle pourrait s'exercer sur la totalité de nos croyances, et en particulier sur le type

[86] Cf. Ramsey, « Vérité et probabilité » (1926). Cela dit, dans cet article de 1926 Ramsey préconisait une méthode de quantification des degrés de croyance fondée sur la psychologie qui s'est avérée chimérique, ainsi qu'il l'admettra quelques années plus tard (cf. Ramsey, « Probabilité et croyance partielle », 1929, p. 189-190).

[87] Voir en particulier K. Gibbins : « Partial Belief as a Solution to the Logical Problem of Holding Simultaneous, Contrary Beliefs in Self-Deception Research » (1997), p. 115-116.

de croyances dont il est question dans la duperie de soi, c'est-à-dire sur des croyances qui se trouvent en général associées à des désirs ou des émotions fortes et dont l'« oubli » volontaire s'avère souvent impossible. Nous avons évoqué l'exemple de Raskolnikov hanté par son crime et ruminant son remords, mais il aurait peut-être suffit de songer aux cas tout aussi malheureux des personnes qui boivent « pour oublier » sans jamais y parvenir. De toute évidence, nous ne pouvons pas supprimer les croyances indésirables simplement parce que nous le souhaitons. De surcroît, Cohen définit la croyance en termes de sentiment (*feeling*), ce qui rend encore plus invraisemblable la possibilité d'activer et désactiver les croyances à volonté, non seulement parce que la production et la suppression des sentiments semble échapper au contrôle de la volonté, mais aussi parce que l'on tient en général les sentiments pour des états occurrents, tandis que les croyances dont parle Cohen ne sauraient demeurer tacites si elles n'étaient pas avant tout des dispositions.

Engel parvient pour sa part à surmonter ces difficultés en proposant (1) une conception causale et « sub-intentionnelle » du processus de suppression de la croyance indésirable et, d'autre part, (2) une conception dispositionnelle-fonctionnaliste de la croyance en général. La dernière de ces hypothèses lui permet de rendre compte du caractère tacite des croyances supprimées, que Cohen laissait inexpliqué. Dans le sillage de ce que revendiquaient Peirce et Ramsey, Engel envisage les croyances comme des dispositions pouvant ou pas devenir occurrentes, et pouvant ou pas devenir conscientes. Quant à la première hypothèse, elle permet de comprendre comment le processus de désactivation doxastique est possible : ce n'est pas de façon intentionnelle et consciente que la croyance indésirable se trouve évacuée de la conscience, comme le prétendait Cohen, mais en vertu d'un processus causal sub-intentionnel dont le sujet n'est pas maître et qui pourrait même échapper à sa conscience.

Seulement, il nous semble que ce type d'explication suffit à elle seule pour rendre compte de la duperie de soi, sans qu'il faille supposer de surcroît que les agents *acceptent* intentionnellement le contraire de ce qu'ils cessent de croire consciemment. C'est sur ce point que notre analyse diffère de celle d'Engel. D'autant que la notion d'acceptation soulève plusieurs difficultés, comme il semble le reconnaître lui-même. D'une part, on ne voit pas quelle raison aurait un agent de prendre la décision d'accepter que non-*p* alors que de toute manière il ne croit plus que *p*, étant donné que cette dernière croyance a été supprimée (et avec elle l'anxiété ou le malaise qu'elle pouvait engendrer). D'autre part, la notion d'acceptation semble réductible à celle de simulation de la croyance. Mais si l'agent qui accepte que *p* ne fait qu'*agir comme s'il croyait* que *p*, alors on ne peut pas dire qu'on ait affaire à un cas authentique

de duperie de soi, qui exigerait que l'agent adhère sincèrement et sans réserves à la croyance que *p*. Et enfin, affirmer que la duperie de soi résulte du concours d'attitudes propositionnelles de genres distincts, et qui obéissent à des logiques distinctes, risque de nous ramener derechef aux difficultés de la solution homoncullariste. C'est Engel lui-même qui attire notre attention sur cette difficulté : « le fait de faire la distinction entre croyance et acceptation, l'une volontaire, l'autre involontaire, ne revient-il pas, malgré tout, à postuler un double centre d'action ou de cognition dans le sujet, quelque chose comme deux systèmes, l'un responsable des croyances, l'autre responsable des acceptations ? [...] Cela ne réintroduit-il pas implicitement l'homoncularisme ? »[88].

Or, il semble possible d'éviter tous ces écueils en mettant de côté l'aspect intentionnaliste de la solution d'Engel, qui tient uniquement à la notion d'acceptation, pour n'en garder que l'aspect purement cognitif. Après tout, si l'on parvient à expliquer la désactivation d'une croyance *p* à la lumière d'un processus causal purement sub-intentionnel, comme le préconise Engel, il y a des chances que l'on parvienne, en outre, à trouver une explication semblable pour la formation de la croyance inverse que non-*p*. Au lieu de supposer que celle-ci doit être acceptée par un acte intentionnel et réfléchi, contrastant avec la façon sub-intentionnelle dont son inverse se trouve supprimée, on supposera que c'est un seul et même principe dynamique qui explique à la fois la désactivation de la croyance de départ et l'induction de la croyance finale. Une hypothèse plausible consiste à suggérer que ce principe réside dans l'influence sub-intentionnelle de nos émotions et de nos désirs sur les processus cognitifs à l'œuvre dans la formation de nos croyances.

8. Références

Bach, K. (1997) « Thinking and Believing in Self-Deception », *Behavioral and Brain Sciences*, 20, p. 105.

Cohen, J. L. (1989) « Belief and Acceptance », *Mind*, 1989, pp. 367-389.

Cohen, J. L. (1992) *An Essay on Belief And Accepting*, N.Y., Oxford, Oxford University Press.

Correia, V. (2007) « Une conception émotionnaliste de la *self-deception* », *Teorema*, vol. XXVI, 3, p. 31-43.

Correia, V. (2010), *La duperie de soit et le problème de l'irrationalité*, Saarsbruck, Berlin, Editions Universitaires Européennes.

[88] Engel (1995), p. 117.

Davidson, D. (1986) « Duperie et division », trad. fr. P. Engel in *Paradoxes de l'irrationalité*, l'Éclat, Combas, 1991, p. 45-61.

Davidson, D. (1998) « Who is Fooled ? », in J. P. Dupuy (éd.), *Self-Deception and Paradoxes of Rationality*, CSLI Publications, Stanford, California, p.1-19.

Davidson, D. (1982) « Paradoxes de l'irrationalité », trad. fr. P. Engel in Davidson, *Paradoxes de l'irrationalité*, Combas, Paris, 1991, p. 21-43.

Kenny, A. (1984) « The Homunculus Fallacy », in *The Legacy of Wittgenstein*, Blackwell, Oxford, chap. 9, p. 125-136.

Engel, P. (1995) « Croyance, jugement et *self-deception* », *L'inactuel*, 3, p. 105-122.

Engel, P. (1998) « Believing, Holding True and Accepting », *Philosophical Explanations*, I, 2, p. 140-151.

Engel, P. (1996) « Croyances collectives et acceptations collectives », in R. Boudon, A. Bouvier, & F. Chazel, (éds.), *Cognition et sciences sociales*, Paris, PUF, p. 155-173.

Engel, P. (1997) « Dispositions à agir et volonté de croire », J. Proust et H. Grivois (éds.), *Subjectivíté et conscience d'agir*, Paris, PUF, p. 115-138.

Engel, P. (1998) « Believing, Holding True and Accepting », *Philosophical Explanations*, I, 2, p. 140-151.

Engel, P. (1999) « Dispositional Belief, Assent and Acceptance », *Dialectica*, 53, 3-4, p. 211-226.

Engel, P. (2000) *Believing and Accepting*, Dordrecht, Kluwer Academic Publishers.

Freud, S. (1915) « Le refoulement », trad. fr. J. Laplanche et J.-B. Pontalis in *Métapsychologie*, Paris, Editions Gallimard, 1968, p. 45-63.

Freud, S. (1915) « L'inconscient », trad. fr. J. Laplanche et J.-B. Pontalis in *Métapsychologie*, Paris, Editions Gallimard, 1968, p. 65-120.

Freud, S. (1938) *Abrégé de psychanalyse*, Paris, PUF, 1950.

Gibbins, K. (1997) « Partial Belief as a Solution to the Logical Problem of Holding Simultaneous, Contrary Beliefs in Self-Deception Research », *Behavioral and Brain Sciences*, 20, p. 115-116.

Lycan, W. G. « Functionalism », in S. Guttenplan, *A Companion to the Philosophy of Mind*, Blackwell, Oxford, 1995, p. 317-323.

Mele, A. (1998), « Two Paradoxes of Self-Deception », in J.P. Dupuy, *Self-Deception and Paradoxes of Rationality*, CSLI Publications, Stanford, California, p. 37-58.

Peirce, C. S. (1877) « Comment se fixe la croyance », trad. fr. j. Chenu in *Textes Anticartésiens*, Aubier Montaigne, Paris, p. 267-286.

Ramsey, F. (1926) « Vérité et probabilité », trad. fr. sous la dir. de P. Engel et M. Marion, *Ramsey : Logique, philosophie et probabilités*, Vrin, Paris, 2003. P. 153-188.

Ramsey, F. (1929) « Probabilité et croyance partielle », trad. fr. sous la dir. de P. Engel et M. Marion, *op.cit.*, p. 189-190.

Recanati, F. « The Simulation of Belief », in P. Engel (éd.), *Believing and Accepting*, 2000, p. 267-282.

Sartre, J. P. (1943) *L'être et le néant*, Paris, Seuil.

Stalnaker, R. (1994) *Inquiry*, Cambridge Mass., M.I.T. Press.

24

Between Knowing how and Knowing that *

CARLO PENCO

There is something I don't understand about the discussion on "knowing how" and "knowing that". Is it a real alternative, or is it a question on how to use the term "to know"? The recent solution by Williamson-Stanley 2000 ("knowing how" is reducible to "knowing that") implies a distinction between two kinds of "knowing that": a normal "knowing that" and a "knowing that" with *practical modes of presentation (MOP)*. Does the second take the place of the old "knowing how"? Is that a real advantage? What could we gain from abandoning the old distinction of Ryle's Anti-Intellectualism and accepting the new distinction of Intellectualism?

Old Distinction	*New Distinction*
(Ryle)	(Stanley-Williamson)
knowing that	knowing that
knowing how	knowing that with practical MOPs

*Thanks to Cristina Amoretti for her criticism on a first version of the paper.

Is that all? Well, no. While anti-intellectualists tend to identify knowing how with having a certain ability or being able to do something, we are suggested by intellectualists to distinguish knowledge from ability to do things; one may be able to do certain actions without knowing how to do them. Really? If there is a basic distinction between "knowing a way to do things with practical MOPS" and "ability", why not to say that Ryle was confused in overlapping the conception of "knowing how" and the conception of "being able to"?

With these worries, I decided to re-read Engel 2007 to find suggestions on this issue (and to find a justification of my participating to the volume in his honor). And I have found further worries. In this paper I will therefore present some problems raised by reading Engel on "Taking seriously Knowledge as a Mental State".

1. Engel and Williamson: knowledge as a mental state is not a natural kind

Engel accepts Williamson's definition of knowledge. As Sellars claimed the priority of "seeing "on "seeing as", Williamson claims the priority of "knowing" on "believing". Sellars 1956 argued that Descartes – giving preeminence to ideas or impressions in the mind – put things in reverse order. From a stereotypical "Cartesian" point of view, that something *looks* green is the primary datum from which to start; then we may reach certainly and knowledge when we are justified to say that something *is* green. Contrary to this view – as Brandom 1977 says in his commentary to "Empiricism and the Philosophy of Mind" – "'Looks' talk does not form an autonomous stratum of the language – it is not a language-game one could play though one played no other. One must already be able to use 'is-*F*' talk in order to master 'looks-*F*' talk, which turns out to be parasitic on it. In this precise practical sense, is-F is *conceptually* (Sellars often says 'logically') *prior* to *looks-F*." An argument similar to the one suggested by Sellar for the priority of *is-F* to *looks-F* could be developed for the conceptual priority of knowing over believing. Mimicking Sellars' argument in short we may say that belief could not be a language game on its own unless we presuppose knowing, given that to believe can be interpreted as to be uncertain about our knowledge. I believe because I am not sure to know.[1]

[1] This is just my intuitive formulation of the problem. Williamson's main argument is different, but one of the arguments given in Williamson (2000: 69-70)on the primeness of knowlege is based on the primeness of "seeing"; the topic is discussed by Engel (2007: 54-55) to show that

Both knowing and seeing share the semantic property of being factive. To claim that an attitude A is factive is to claim that if one As that P, therefore P. An internalist may accept that knowledge is factive: S knows that he is thinking, therefore he is thinking; but, from an internalist viewpoint, this factive aspect of the attitude of knowing is bound to be conceived inside the mind. To claim that knowledge as a mental state is factive – in respect to all things said to be known – one has to renounce to the Cartesian idea of "private" or "internalist" mental state, because no internal, private idea may have an external fact as a consequence. A mental state factive[2] in the externalist sense may give a different explanation: "being factive, knowledge is a mental state that is *essentially* factive. And being external, knowledge is a condition which is such that one can possess it without ... knowing that one has it": following Williamson, Engel rejects the KK principle (if you know that *p*, then you know that you know that *p*) normally accepted by internalists, and claims that knowledge does not imply that one knows that one knows (55-6). This point may be an essential point of some arguments against the distinction between knowing that and knowing how.

Assuming Williamson's view of knowledge, Engel needs to better define what is no so clearly defined in Williamson's view: to what extend and in which sense knowing is a "primitive" state of mind, a factive mental state from an externalist viewpoint. Assuming that knowing is a mental state seems to imply a naturalization of knowledge, but it is not necessarily so. Actually Williamson's theory of knowledge might be interpreted in two different ways:

> (i) as a normative characterization of knowledge at the conceptual level, as a critique to the traditional definition of knowledge as JTB and a new alternative definition.

---

states like seeing and knowing are not a "mere conjunction of something purely internal and something purely external", avoiding therefore a useless opposition between internal and external elements of knowledge.

[2] Engel claims that, being factive, knowing cannot be properly considered a "propositional attitude" because, differently from belief or desire, the state of knowledge cannot be "separated" from the content of knowledge. At first impression this sounds a little strange: "knowing that" is typically "knowing that *p*": what is "*p*"? A proposition. What is "to know"? an attitude. I may have an attitude towards a proposition such that my attitude is inextricably connected to the content of the proposition, but still having an attitude towards it; maybe Engel wants to differentiate the attitude of certainty that accompanies knowledge and knowledge itself which is a kind of state connected to the content of a proposition, but not a proper "attitude". However, if I say "x does not know that p" I have a proposition that can be separated from x's state of knowledge. Can we define "not knowing" an attitude?

> (ii) as a claim on the nature of knowledge; although it may be presented as a metaphysical research and definition, speaking of a mental state implies that knowledge must be studied also with empirical means from psychology to biology.

Not disregarding the second interpretation, Engel maintains that we need to verify *how* Williamson's theory can be confronted with different naturalistic enterprises, to check its compatibility with them. Among the different theories of naturalization of knowledge Engel discusses Kornblith 2002 – according to whom knowledge is a natural kind – aiming to show that Kornblith's account is incompatible with the main features of Williamson's "knowledge" defined as follows:

> (i) K is a genuine mental state
>
> (ii) K is factive
>
> (iii) K is not transparent
>
> (iv) K is primitive
>
> (v) K plays an essential role in the explanation of belief, assertion and action.

There are different kinds of mental states, but we may make the hypothesis that there are kinds of mental states developed by evolution and characterized therefore as natural kinds; according to Kornblith knowledge is a set of cognitive capacities that *underlies* true beliefs and allow member of a species "successfully to negotiate their environment" (Kornblith 2002: 56). While individual behavior may be explained referring to beliefs and desires, successful behavior of a species is explained by the adaptation of these more fundamental cognitive capacities: "if we wish to explain why it is that members of a species have survived, we need to appeal to the causal role of the animal's knowledge of their environment in producing behavior which allows them to succeed in fulfilling their biological needs" (62).

Given that knowledge, in Kornblith's perspective, is a factive mental state, intrinsically associated with the success of interaction with environment at the level of species, and not transparent – given that it belongs also to non human animals – it appears as if it fits most of Williamson's main features of knowledge listed above. Yet, Engel reacts against this possible agreement between these two paradigms, claiming that Kornblith's knowledge cannot satisfy conditions (i)-(v). Engel's discussion is not linked to each condition,

and is touching different general perspectives; it seems to me however that Engel's two main arguments are the ones concerning the inability of Kornblith to differentiate knowledge and belief, and to differentiate human and not human knowledge.

The first argument is that that if we accept that knowledge is "based upon a set of information processing capacities of a general kind which deserve the name of 'natural kind"', then "it is unclear that this can allow us to characterize the mental state in which knowledge consists." (Engel 2007: 63). In fact, also true beliefs are based upon the same set of information processing capacities, and we would have no way to make the difference between knowledge and true beliefs. Besides, defining knowledge in terms of reliability of the same information processing on which beliefs are based, you should not only distinguish knowledge and true belies, but also clarify *why* knowledge is better than true beliefs – something that cannot be explained given the difficulty to distinguish knowledge and true beliefs.

Engel does not make any reference to another possible line of criticism which runs against Kornblith's view in a more direct way: information processing based on perception certainly helped our species to adapt to our environment, but it did so producing from time to time some cognitive information processing enduring in time and typically considered "knowledge" through ages that are false and yet have been fundamental for our survival: just think of the cognitive (perceptual) processing of the rising of the sun as giving information that our star rotates around the earth. We cannot therefore identify all basic information processing or mechanism of survival with knowledge. We should therefore separate knowledge as factive mental state and information processing as mental or biological mechanism apt for survival. In the analogy between Williamson's and Kornblith's knowledge, at least in this case, what is really put in doubt is therefore (ii): success in interaction with the environment does not seem to amount to the truth of the matter. From the fact that S perceives that the sun rotates around the earth, it does not follow that the sun rotates around the earth, although this cognitive processing of our basic perceptual information has been useful to interact with the environment.

The second argument used by Engel is the radical difference in mental representations between humans and not human animals. Kornblith's knowledge as a natural kind is common to humans and non human animals. This "knowledge" would be characterized as a genuine mental state, primitive and not transparent. We have already cast a doubt on the "factive" condition. Engel, although does not say it explicitly, suggests that the Kornblith's "knowl-

edge" forgets the radical difference between human and not human animals given by the advent of language. This advent implies a differentiation in animal and human representation, and the more we consider this difference "the more difficult is to accept that there is a single state of knowledge underlying all instances across all species".[3] I wonder whether there is a more direct way to see the impossibility to use Kornblith's knowledge as an account for Williamson's knowledge: probably it is enough to remark that the information processing common to humans and not human animals cannot perform condition (v) in Williamson's definition as given above, that is "to play an essential role in the explanation of belief, assertion and action". In fact, even if we reject Davidson's view on the impossibility to attribute "beliefs" to animals, it is very difficult to say that animals make assertions. Therefore, what Williamson calls "knowledge" cannot be attributed to non human animals given that it cannot play any role in explaining activities non human animals are not supposed to possess.[4]

2. "Core knowledge" as a challenge to the Intellectualist Theory

One of the main tenets in Engel's viewpoint is that knowledge is specialised and domain specific. He uses also this aspect to criticize Kornblith's view, but probably this feature could be easily taken into account in Kornblith's treatment of knowledge as a set of information processing : a cluster of properties given by the information processing developed as instrument of survival in the environment may be well organized into different specialized kinds. We have however seen at least two features that make Kornblith's perspective incompatible with Williamson's. What about Engel's suggestions? Engel suggests that we can make a good comparison, finding similarities between

[3] Engel 2007: 63. Engel suggests also that taking knowledge as a natural kind for all species in the animal realm faces the difficulty of the generality problem, that has been overcome by Williamson's theory. If Kornblith's "knowledge" is defined by a fixed set of features, we may not be able to assess all the features which make a belief in a given environment reliable as knowledge. But here again I miss the point; in fact Kornblith's knowledge is primitive and does not require to be organized as reliable belief. We are back to the previous point of differentiating knowledge and true belief with some reliability condition. It is not clear whether Kornblith's account is bound to answer to this requirement.

[4] A more complex answer should rely on the role of knowledge to explain action if we give their proper role to objective reasons for acting, abandoning the typical belief/desire account (in an anti-naturalist stance as the one taken by Hornsby 2007).

Williamson's proposal and what is called "core knowledge" by cognitive scientists. We have a basic set of "domain specific capacities" studied by cognitive science as "knowledge" shared by infants and primates: capacity to represent different sort of things (agent, numbers,[5] objects ,places, sounds); capacity to answers questions linked to different tasks (who did it?where? how? what does that do?); capacities that are relatively "encapsulated" and automatic and fast. Engel insists that these capacities extend across species (Spelke and Hausman 2004), but the fact that much is in common between human infants and other animals "does not show that knowledge is a natural kind underlined by a single type of process, for the large variety of process and systems which are at play, and the divergence between animals and humans". (Engel 2005: 65). The point made by Engel is that knowledge cannot be considered a natural kind because there is a such a large varieties in core knowledge that prevents it to be considered a unique natural kind.

Yet, according to Engel, "core knowledge" seems to be fit for most of Williamson's characterization of knowledge: it is a factive, externally based, primary and not transparent state of mind, and it is also linked to action. There is a missing aspect Engel does not explicitly states: core knowledge might not be enough to explain belief, assertion and action, given that sometimes belief, assertion and action require explanation involving complex inferences that core knowledge might be unable to do. This missing aspect makes core knowledge something only partially compatible with a general definition of knowledge, but at the same time makes core knowledge a different proposal from the more radical naturalistic view held by Kornblith. The difference with Kornblith's view is that core knowledge "by no means implies that all sorts of knowledge can be based on it, or even reduced to it; on the contrary a lot of knowledge is built out of the basic system of core knowledge, recombined and advanced." (Engel 2007: 66).

Let us stop here for a moment. Core knowledge does not seem to be propositional knowledge; at first sight it seems that core knowledge is what is considered "knowing *how*", or, better, knowing who did it, where, and how. What does it mean that "a lot of knowledge" is built "out" of the basic system of core knowledge? It seems that it means that, once infants have mastered these kinds of capacities, they are ready to develop, inside human communities, the ability to speak a language and therefore the capacity of knowing *that*. At first

[5] It is common to speak of animals recognizing numbers. As far as I have seen, experiments prove that animals choose what is more rewarding: they are able to differentiate groups that we count with different numbers, not that they use numbers; if they are able to make one-one mapping, they do not have the concept "one" :)

sight this should appear more compatible with anti-intellectualists than with intellectualists: core knowledge seems to be exactly the "Knowing how" of which Ryle was speaking about. Besides, can we reduce these capacities to "knowing that"? This is highly implausible, unless we use our linguistic abilities to describe "beliefs" and "actions" of non human animals with a highly intellectualistic attitude. Still, it is easy to accept the idea that a dog *knows that* his master is arriving, although the dog cannot know that his master will come back again next Thursday.[6] But to claim that we, humans, can describe and explain animal beliefs and actions as "knowing that" does not amount to attribute them a fully propositional knowledge.

Notwithstanding all these reservations, developmental psychologists insist that core knowledge is a kind of *theoretical* knowledge. Engel insists on this fact claiming that, if it is correct to define core knowledge a kind of theoretical knowledge, then Ryle's claim that there is a specific form of knowledge, that is "knowing how", which is distinct from propositional knowledge fails, given that we have a core knowledge which is partly propositional and partly practical, but cannot be reduced to a mere "knowing how" as a set of abilities or dispositions.

It seems therefore that Engel takes stance in favour of the Intellectualist account of knowledge, and counting himself on the same side against the Anti-Intellectualist view of knowing how as a fundamental mode of knowledge. However he does so in a very prudential way: in fact to say that core knowledge is "theoretical knowledge" is not to say that it is fully propositional; on the contrary the status of core knowledge is an "intermediary status between perceptual and inferential knowledge" and is seems "neither fully theoretical or propositional nor fully practical" (p.68). The conclusion is that the notion of core knowledge provides a ground to "reject the division between knowing how and knowing that".

Rejecting the division between knowing how and knowing that, Engel renounces to take a definite stance in the debate about the reduction discussed by Stanley-Williamson. This is confirmed by his reservations towards the reduction: "Williamson and Stanley show *at least* that is not obvious that the distinction between knowing how and knowing that is so clear cut, and that

[6] "A dog believes his master is at the door. But can he also believe his master will come the day after to-morrow?—And what can he not do here?—How do I do it?—How am I supposed to answer this? Can only those hope who can talk? (Only those who have mastered the use of a language. That is to say, the phenomena of hope are modes of this complicated form of life. (If a concept refers to a character of human handwriting, it has no application to beings that do not write)." Wittgenstein 1953, II,i.

a lot of knowing how involves knowing that and propositional knowledge" (68). This sounds as a moderate approval, that does not grant the more radical conclusion Stanley Williamson has thought to have proved. Besides Engel seems to interpret their result as a further justification of his own idea that the division between "knowing how" and "knowing that" is to be rejected; but if this is the conclusion, then it amounts to take a stance *against* the reduction of knowing how to knowing that. In fact, eventually, Engel is dissatisfied of the arguments in favour of the reduction: "Stanley and Williamson's arguments are unconvincing insofar they are purely linguistic, and it is not clear to me that a purely linguistic argument can show that knowing how is a form of knowing that" (*ibid*).

Summarizing Engel's position w.r.t. the debate, on the one hand, we have a positive account about conditions of knowledge shared by our "core knowledge", on the other a programmatic view of the kind of discussion to be done to solve the contrast between linguistic behaviorists and linguistic intellectualists. The positive account about conditions of knowledge accepts Williamson's view: if Williamson's view is correct, and as the studies on the core knowledge suggest, all knowledge – "from children's basic capacities to our scientific knowledge" – has the properties of being factive, non transparent, externally individuated and prime. I may also add that condition (v), that is knowledge as what is relevant in the explanation of assertion, does not belong to core knowledge and we still have the difficulty to make a clear distinction between what can be considered "full" knowledge with conditions (i)-(v) and aspects of knowledge that could be shared also with non human animals. The criticism to Kornblith about the inability to distinguish cognitive processing of non human animals and cognitive processing of linguistic groups seems to be an obstacle also to the foundation of core knowledge, until we can find a way to explain the links between core knowledge and linguistic knowledge (up to scientific knowledge)

The programmatic view is "to bring together empirical findings in psychology and the general conceptual features of knowledge." (Engel 2007: 69). Engel, attempting to bring together empirical findings and theory of knowledge, has left us, notwithstanding an appreciation of the main features of Williamson' theory of knowledge, with a strong doubt on the reduction of knowing how to knowing that. In the last paragraph I will try to explore the doubts left open to further research.

3. On the reduction of knowing how to knowing that

The debate on the reduction of "knowing how" to "knowing that" has actually taken the direction indicated by Engel, with two worries: (1) can a reductio based on linguistic data be applicable to a mental state which is not always conceived as linked to the use of language? (2) can the solution given by Stanley-Williamson answer to recent findings in pychological research?

Actually there is nothing wrong in using linguistic data; actually we use the term "Knowledge" mainly in knowledge attribution, therefore our way to use the word "to know" is of fundamental importance. The linguistic fact that knowing *where* to F or *why* to F or *when* to F or *how* to can be *defined* in terms of propositional knowledge seems unobjectionable. The "reduction theory" may express all those knowledge quantifying over places, reasons, places and ways (under some practical modes of presentation).[7] The linguistic evidence is apparent; any time we speak of knowing *where* or *when* or *how* we may "translate" this knowledge in term of "knowing that *x* is the *place* where...", or "knowing that *x* is the *time* when .." or "knowing that *x* is a *way of doing* things".

But there is a worry that remains unanswered: what about chicken sexers or, to give some more sophisticated examples, what about wine tasters or Balsamic Vinager tasters? Expert Balsamic vinager testers in Modena may say with a reliable degree of approximation in which kind of wood barrel and in which year that vinager has been staying in that barrel, going back to 10 years of different kinds of woods. However when asked to explain the way the do it, they seem unable to do it; they possess a practical ability to recognize good vinager, like chicken sexers have a practical ability to recognise the sex of chicken. But they cannot express this ability as " we know that *x* is a way to recognize chicken's sex" or "that *x* is a way to recognize the seasoning of the vinager". It seems that there is no recognizable "*x*", and yet we find it difficult not to attribute them some kind of knowledge.

The main worries the reductio has to face is – as it is clear in these cases – the over-linguistification of knowledge attribution, the relation between the theory and normal use of knowledge attribution and the relation with psychological data. A common feature of all these worries is the (old) problem of different intuitions we have in front of the same linguistic data.

One of the main criticism comes from a cognitive scientist, Alva Noë, and starts from a different interpretation of linguistic data: among the examples

[7] The first paragraph of Stanley 2011 is a synthetic and clear presentation of this strategy.

Stanley and Williamson offer in their favour Alva Noë reports the following: (i) If Hannah digests food, she does not know how to digest food, and (ii) If Hannah wins a fair lottery, she still does not know how to win the lottery. Are these cases where Hannah does something and yet we cannot say that she knows how to do it, as W-S claim? This is not so, as Noë 2005 remarks: digesting and winning a lottery are not intentional actions, but something that happens to an individual. We have therefore to reject the cases. However we might say that Hannah knows how to buy or eat food with a fork, and knows how to buy a ticket for the lottery; well these knowings *how* can be easily translated into propositional knowledge.

We are touching here the main core of the contrast on how to interpret the normal use of the term "knowledge". Stanley and Williamson maintain that their reduction does not imply that to engage into an action one must contemplate a proposition. To be ale to F implies knowing how to F and this implies knowing that x is a way to F under a practical mode of presentation, *without* be compelled to accompany this knowledge with a contemplation of a proposition. Noë remarks that there is no explanation of justification of this point, but a quotation from Ginet who claims that we may engage in actions (for instance opening a window) without entertaining the corresponding proposition. But this would amount to claim that "we do not have conscious experience of formulating propositions every time we act in ways we give expression to our propositional knowledge" (Noë 205:281). Saying that, Noë wants to distinguish between (i) which kinds of actions *constitutes* propositional knowledge and (ii) on the basis of which actions we *attribute* propositional knowledge. Certainly we attribute propositional knowledge to people on the ground of their answering our request ("please, may you open the window?") and there is no need to attribute to the performer a contemplation of a proposition such "I am opening the window" while doing it. Nobody requires that for an attribution of knowledge. But Ryle would say that knowing how to open the window is *constituted* by the individual ability to perform the action and not by any proposition.

Eventually we have reached the central point of the "reductio": the real target, as Noë (2005: 282) claims, is Ryle's identification of "knowledge how" with the possession of abilities or dispositions. Noë 2005 (284-5) devotes some effort to explain that abilities are embodied and situated, and their capacity to detect significance, where there would be otherwise none, makes it reasonable to consider practical abilities as a kind of knowledge. Accepting the reductionist theory would makes a mystery on "why embodiement and situation should or could be as important as they are". He insists in claiming that at

a some point "it must be possible to give possession-condition for concepts in non-conceptual, and so non-propositional terms. For example, my grasp on [sic] the concept *red* probably does not consist in my knowledge of propositions about redness (...) My grasp of red consists, it is more likely, in my dispositions to apply *red* to an object when it exhibits a certain quality" (285) I wonder which is the difference between humans and non human animals on this level; parrots can be taught to have dispositions to apply *red*, and even to say "red", in front of objects with a certain quality.[8] Do they have the same concept of "red" as we humans have?

Let us skip over this dubious identification of "grasping a concept" and "having a responsive disposition", and let us go to the main challenge to Intellectualists presented by Alva Noë. The challenge is the following: can the intellectualist show that "having the ability to do something does not consist in knowing how to do it"? If possession of abilities is a matter of knowledge-how, then, Noë argues, we have a conclusion opposite to intellectualists: "All knowledge that depends and must be analysed in terms of more basic knowledge how." (286)

Stanley 2011 gives same arguments against the idea of identifying "being able to do it" and "knowing how to do it": they are different matters: "being able to do something and knowing how to do it are certainly not the same"; besides, knowing how to do something, although often is *de facto* connected with ability to explain how to do something, in principle is not directly connected with knowing how to explain. What is required is only to express such propositional knowledge; but propositional knowledge does not require a over idealization and linguistification of knowing how: "the 8 year old Mozart can assert the proposition that constitutes his knowledge how to compose a symphony; he can just say, while composing it, the German translation of "this is how I can do it". (p. 10). Propositional knowledge does not need to be expressed in purely descriptive terms; it may be given also with demonstrative and indexicals. The existence of a radical distinction between being able to and knowing how to implies that possessing an ability or a disposition does not amount to knowing. You may have an ability without knowledge. Stanley reports some complicated examples to show that intentionally performing an action does not amount of *knowing* how to perform it.

[8] I am referring to the well known argument by Brandom. Sellars' ideas are developed by Brandom, who insists on the difference between "responsive classification" (parrots can do it) and "conceptual classification" (Brandom 1994: 88-89; 122)

If we accept this line of thought, Alva Noë conclusion does not follow. First: the reduction of "knowing how to F" to "knowing that *x* is a way to F" (under a practical mode of presentation) does not exclude that "knowing that x is a way to F" is always grounded on practical abilities; if we want to distinguish practical abilities from proper knowledge we may well accept that practical abilities are "situated" and "embodied"; they are the ground on which we may be said to know how to do things, that is to know that there is a way to do things. We may have the abilities to do something and we cannot know how to do it; for instance if we acquired the abilities by chance; but also the contrary happens: we may know how to F without having the abilities that we might have lost with age – I know that :).

Summarizing the basic point: possession of abilities and dispositions is a prerequisite of knowledge, like the possession of a reliable visual system that detects a particular spectrum of wave lengths is a prerequisite of the concept RED; but the possession of the concept RED is not identified in the responsive disposition, because grasping the concept requires our mastering the use of the concept at least in the network of the logic of colors (red is a color that is different from blue). In the same way knowing how to do it is not indentified with the abilities we have in doing it.

If we agree that knowing how is *definable* in term of propositional knowledge (plus a practical MOP) we might also agree that "it would be odd to maintain that ascription of knowledge-how are less than fully propositional" (Stanley 2011, p.7); however– in the new framework – there is a very delicate point: the point is that ascription of knowing how are *not less*, but *more* than fully propositional; that is knowledge how is knowledge that *x* is a way to F... *plus* the requirement to have a practical mode of presentation of *x*. But this is another way to make a difference where we were brought to believe there is none (maybe also for this it may be called a "Pyrric victory").[9] If every knowledge has special modes of presentation, but only a particular knowledge has a practical mode of presentation, why don't you call this kind of knowledge "knowing how"? It would amount to a redefinition of knowing how after

[9]The term is used by Brown 2013 in the context of comparing propositional knowledge as a more general concept in respect of declarative and procedural knowledge. I don't discuss the topic, because the old contrast between procedural and declarative, born in computer science, seems to me a source of confusion. On the one hand the two ways can be translatable one into the other; from this point of view there is no much difference between a procedural or declarative explanation about where Central Park is: "it is in New York near 5thavenue" (declarative) or at "go to New York and take 5th avenue and you find it". On the other hand a procedure (with respect to a function) may be a perfect explanans of a practical mode of presentation, as Pavese (forth.) has suggested.

having destroyed the old fashioned view according to which there is a sharp division among the two kinds of knowledge. As Engel suggested, the distinction is fuzzy: we may always *express* every kinds of knowledge attribution as knowing *that*, and some kinds of knowledge attribution as requiring a practical mode of presentation: these special kinds of knowledge are the heir of what had be once called "knowing how". This would be the most clear way to cut the grass under the resurgence of Ryle's Anti-Intellectualism. But...

...but a question of terminology is still open: what to do if scientists use the term "knowledge" for abilities and capacities to detect differences in the (natural or artificial) environment? Philosophers might fight against an improper use of the term, and insist on the difference between knowledge and responsive disposition; the diffusion of "knowledge terminology" in ethology, psychology and cognitive science makes philosophers uncomfortable; however what is relevant is to keep clear conceptual distinctions, and philosophers might do their job also accepting different uses of the term "knowledge".[10] Actually, nobody in cognitive science or in cognitive psychology uses the term "thought" as Frege used it, and we still are able to entertain discussions between philosophers and cognitive scientists.

4. References

Brandom R. 1994, *Making it Explicit*, Cambridge (Mass): Harvard U.P.

Brandom R. 1977, "Study Guide" in Sellars 1956, printing 1977.

Brown, J. 2013. "Knowing How: Linguistics and Cognitive Science", *Analysis*, 73/2: 220-227.

Engel, P. 2007, "Taking Seriously Knowlede as a Mental State". In *Explaining the Mental,* edited by M. Beaney, C. Penco, M. Vignolo, Newcastle: Cambridge Scholar Publishing: 50-71.

[10] "When philosophers use a word—"knowledge", "being", "object", "I", "proposition", "name"—and try to grasp the essence of the thing, one must always ask oneself: is the word ever actually used in this way in the language-game which is its original home?" (Wittgenstein 1953 §116). We have here a case where there is a new home for "knowledge"; the new home is the setting of scientific research. Here we may decide that there is a new concept with a new origin, or the concept "Knowledge" is the one we wanted to describe following its introduction in our community, and is changing under our eyes. But we might also take Carnap's attitude and call "knowledge1" philosophers' knowledge and "knowledge2" psychologists knowledge. Actually somebody has already done that differentiating "knowledge" and "cognition".

Kornblith, H. 2002. *Knowledge and Its Place in Nature*. Oxford, Oxford University Press.

Hauser, M and Spelke, E. 2004. "Evolutionary and Development Foundations of Human Knowledge". In *The Cognitive Neurosciences III*; edited by M. Gazzaniga, Cambridge (Mass): MIT Press.

Hornsby J. 2007 "Knowledge, Belief and Reasons for Acting". In *Explaining the Mental*, edited by M. Beaney, C. Penco, M. VIgnolo, Newcastle: Cambridge Scholar Publishing: 88-105.

Noë, A. 2005. "Against Intellectualism", *Analysis*, 65/288: 278-290.

Pavese, C. *forth* ."Practical Senses" (manuscript)

Sellars, W. "Empiricism and the Philosophy of Mind", Minnesota Studies in the Philosophy of Science, edited by H. Feigl and M. Schriven; later publisched as a book with the same title, Cambridge (Mass): Harvad Unviersity Press, 1977.

Stanley, J. and Williamson, T. 2000. "Knowing How", *Journal of Philosophy*, 98: 411-444.

Stanley, J. 2011. "Knowing How", *Nous*, 45/2: 207-238.

Williamson, T. 2000. *Knowledge and Its Limits*. Oxford, Oxford University Press.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

25

Knowledge First — a German Folly?

KEVIN MULLIGAN

1. Introduction

Timothy Williamson's *Knowledge and its Limits* broke sharply with received analytic wisdom according to which knowledge-that is a species of belief. It is rather, he argued, a relation in which we stand to true propositions or facts[1].

The Oxonian dimension of the history of the knowledge first approach in epistemology, in particular in the writings of Cook Wilson and Prichard, has been magisterially expounded by Mathieu Marion[2]. The Germanophone dimension of the history of this approach is less well-known. It comprises two parts. The first goes from Jakob Friedrich Fries in the first half of the nineteenth century to Leonard Nelson, an unusually lucid and argumentative neo-Kantian. The second begins with Husserl's *Logische Untersuchungen* in 1900-01 and is developed by several of Husserl's students and disciples, the early, realist phenomenologists - Reinach, Scheler, and von Hildebrand. There is a connexion between the two German strands in our story. Reinach and Nelson, as well as Husserl, were colleagues in Göttingen before the Great War.

[1] Williamson 2000. For the view that there are mental acts and states which are relations, in the proper (non-Brentanian) sense of the word, to objects and facts, in the proper (non-Fregean) sense of the word, cf. Smith 1984, Mulligan & Smith 1986.

[2] Marion 2000 ; cf. Marion 2002, 2003.

Ramsey described the Oxford view that, as Prichard put it, „Knowledge is *sui generis*, and, as such, cannot be explained", as „the Oxford Folly"[3]. If it is a folly, it is also a German Folly. Indeed if some seeds of the view are to be found in Husserl, it is an Austro-German Folly. In §§2-5 I sketch, disentangle and evaluate the views of the early, realist phenomenologists about the priority of knowledge. In the final section (§6), I compare some aspects of Göttingen and Oxford views about the primacy of knowledge.

1.

2. Apprending (*Erkennen*) & Knowledge

The language of epistemology sometimes reflects and is perhaps even a prisoner of the language in which it is written. An epistemic verb which will be important in what follows is „erkennen" , which has sometimes been translated as „cognize", sometimes means „recognize" and which I shall translate here as „apprehend".

The expression „theory of knowledge" translates into German as „Erkenntnislehre" or „Erkenntnistheorie" and into French as „théorie de la connaissance" (and even as „gnoséologie"). Although „connaissance" undoubtedly often refers to the knowledge-that which people possess or have, „connaître » is a verb which takes a nominal complement and, in such cases, is translated into English by « know », « is acquainted with » or „ken" with a nominal complement. „Erkenntnis", too, often refers to the knowledge-that which people have or possess. But the verb „erkennen" has often been used by epistemologists to refer not to the knowledge that *p* which someone has or possesses but to the episode of apprehending or coming to know that *p*, for example, to discovery. Another construction, employed for example by Husserl, is « x apprehends Hans as Hans ». This type of apprehension plays an important rôle in Husserl's development of the view, to be found later in Russell and Evans, that referring with the help of proper names presupposes knowledge of what is referred to.

[3] Prichard 1909 124 ; Ramsey 1991 81. Both remarks are quoted by Marion (2000 310), who plausibly takes Ramsey's remark to be about the sort of view expressed by Prichard. But it is possible that by « the Oxford Folly » Ramsey intends to refer to what he calls « an Oxford error » : « They suppose knowledge can be guaranteed » (Ramsey 1991 82). Marion also quotes Cook Wilson on Nelson on knowledge. The Oxford philosopher's very oxonian reaction to the possibility that he might have German predecessors betrays his apparent ignorance of the fact that Nelson's view of knowledge was a development of the much earlier work by Fries.

A theory, analytic description or philosophy of knowledge which aims at completeness would doubtless do well to distinguish systematically between epistemic episodes and non-episodes (states, dispositions) and also between epistemic intentionality which is thatish and epistemic intentionality which is non-thatish. That is, between knowledge that *p*, coming to know or apprehending that *p*, knowing or being acquainted with *x* and coming to know or making the acquaintance of *x*.

Thus in 1933 the phenomenologist Spiegelberg distinguished two meanings of « Erkenntnis ». The term may signify, he says, a „true judgement, the truth of which is evident" (wahres Urteil, dessen Wahrheit ersichtlich ist). It may also signify „the cognitive act in which an object or state of affairs itself is ´grasped` (der kognitive Akt, in dem ein Gegenstand oder ein Sachverhalt selbst "erfaßt" wird)"[4]. Current versions of the knowledge first approach typically concentrate on knowledge that *p* rather than on apprehending that *p*, the phenomenon which is central in the German versions of the knowledge first approach. This makes for an important difference between the two approaches. But it is interesting to note that early Oxford versions of the knowledge first approach do refer frequently to „apprehension" and even to „acts of knowing" (cf. §6).

Husserl often asserts that knowledge is justified true belief. But he does not understand this claim in the way many anglophone epistemologists understood it during the second half of the twentieth century[5]. His claim does indeed concern knowledge-that (*wissen, dass*). But in what he calls the strict sense of „knowledge" the relevant type of justification or ground is not defeasible. One knows that *p* in the strict sense only if one has perceived that p and such perceiving is not itself any sort of belief or judging[6]. Husserl seems not to have been bothered by the consequence that, on his view of strict knowledge, even taking into account the variety of perception and intuition he allows for, there is not very much of it. He does also allow for a lax sense of „knowledge" where defeasible but undefeated justification plays a role. Indeed Husserl and Meinong seem to have introduced the very idea of defeasible justification, inductive and non-inductive, into twentieth century epistemology[7]

[4] Spiegelberg 1933 111.

[5] Cf. Mulligan 2006.

[6] Nelson (1908 71) notes that because Husserl's concept of justification (*Evidenz*) entails truth he does not belong to the school which takes *Evidenz* to be a criterion of truth, the school criticized by Nelson and to which Meinong belongs.

[7] Meinong's term is *Vermutungsevidenz*, the evidence for conjectures. On Meinong on defeasible justification, cf Teroni 2005. On the history of appeals to defeasible justification in recent

Husserl's view that there is a direct, non-doxastic non-judgmental perception or intuition that p seems to have been the starting point for work by his pupils which puts knowledge first in ways not dreamt of or not countenanced by Husserl.

3. Knowledge First according to the Early Phenomenologists.

One of the last formulations of the knowledge first approach by a phenomenologist is due to Friedrich Bassenge. There is, he thinks, an intimate relation between statements and knowledge. In the normal case, he says, statements express knowledge:

> State of affairs, knowledge, statement — these are the three basic phenomena ... The normal route to a statement is: (1) a state of affairs obtains ; (2) the state of affairs becomes apprehended and as a result known (*gewusst*) ; (3) the known state of affairs is stated (i.e in the typical case communicated to someone else). A materialist theory[8] of knowledge and logic must, it seems to me, take this typical sequence as its starting point and not place the deficient modes in the foreground[9].

It is not clear what „normal" means here; it is presumably not a statistical notion ; we shall meet the notion again in §5. The expression „deficient modes" is a piece of Heideggerian jargon which Bassenge, no Heideggerian, highjacks for his own purposes. (Being alone is a deficient mode of togetherness, says the author of *Sein und Zeit*, who identifies many more such modes). What are the deficient modes Bassenge has in mind ? They include error or merely apparent knowledge and, it seems, judgement and belief:

> The theory of knowledge and logic of past centuries did not put in the foreground the concept of knowledge (*Wissen*) but rather

anglophone epistemology, cf Dutant 2010, ch.1.

[8] A materialist phenomenologist ? Bassenge, an anti-fascist, spent the last part of his life in the DDR. Bassenge (1955) is part of an extensive discussion between DDR logicians and philosophers of logic in which Bassenge does his best to persuade his colleagues of the view defended by Reinach before the Great War to the effect that the logic of propositions is in the first place the logic of states of affairs and only secondarily the logic of propositions. There are even earlier formulations of such a view in Husserl and it is, of course, also reminiscent of some of Frege's views in 1879.

[9] Bassenge 1955 486.

> the concept of judgement. A judgement in this sense is a belief (*Meinen*, believing) whereby one abstracts from whether the belief has just been gained (knowledge, *Erkenntnis*) or is habitual (knowledge, *Wissen*), or remains unexpressed, or is being expressed (as so to speak the inner aspect of the statement) or, above all, from whether the state of affairs believed [or meant] obtains (true judgement) or not (false judgement).[10]

According to Williamson's *bon mot*, mere believing or opining is „botched knowing" . According to Bassenge and other phenomenologists, a great deal of epistemology has been based on the asumption that the central concept of epistemology is in fact that of a deficient mode of knowledge. This way of conceptualising epistemology, says Bassenge,

> was necessary for a philosophy whose starting point was what is subjectively meant or believed and whose main problem was whether it is at all possible to transcend this subjective starting point in the direction of objectivity or not[11].

Dietrich von Hildebrand, another active enemy of Hitler and a militant Catholic, provides what is perhaps the fullest account of the variety of knowledge in the phenomenological tradition and also defends the knowledge first view. He makes all the distinctions introduced above in §2, between coming to know that *p*, knowledge that *p*, acquaintance and coming to be acquainted with objects. He also argues at some length that the two non-episodic phenomena, knowledge-that and acquaintance, may be either merely potential or « supra-actual » , that is, more than mere dispositions. None of these phenomena, he thinks, can be understood in terms of judgment, assertion, belief or conviction[12] :

> Apprehending is one of those ultimate phenomena which cannot be reduced to anything else, which we therefore cannot „define"... [13]

> Conviction [is an] epiphenomenon and the fruit of apprehending[14]

[10] Bassenge 1955 486.

[11] Bassenge 1955 486.

[12] Hildebrand 1950 was written in the 1930's but could not be published. Parts of it are translated in Hildebrand 1960.

[13] Hildebrand 1950 5.

[14] „Die Überzeugung als Epiphänomen und Frucht des Erkennens" (Hildebrand 1976 24).

> Knowing in the wider sense is presupposed by judgement and differs from it[15].
>
> Judging or asserting in the narrow sense forms in a certain way the classical end-point, which does not belong to apprehending itself but is rather founded on the latter as something quite new. . . .I speak; knowledge about the obtaining of a state of affairs is presupposed[16].

Von Hildebrand's account of apprehending that *p* contains the claim that such apprehension is meaning-free, concept-free, and free of predication:

> In apprehending the medium of units of meaning is absent. If I perceive a red (*ein Rot*), the meaning unit "red" is not involved. If I apprehend that the sun is shining, this state of affairs stands immediately before me, without it being the case that I have to go through the proposition (*Satz*): the sun is shining, of which I can predicate truth and falsity[17].

The view that there is visual perception of things, persons, events and monadic qualities which need involve no conceptualisation has been familiar in analytic philosophy ever since Fred Dretske's pioneering investigations of what he, like Husserl, calls "simple seeing". It has even at times been quite popular, and has sometimes been combined with the view that such seeing involves content, a way in which what is seen is seen, which is non-conceptual. But the further claim that not only simple seeing of things, organisms and qualities but also perceptual apprehension that *p* may be concept-free and so non-doxastic and non—judgmental, has, as far as I can see, never enjoyed the same degree of popularity.

This claim that perceptual apprehension that *p* is concept-free is is in fact ambiguous and, on one reading, may be held to be less controversial than on the other reading. The ambiguity was first pointed out by Scheler. To say that perception that *p* or perceptual apprehension that *p* is concept-free could just amount to the claim that it involves no subsumption under concepts, no predication. But it could also be read as claiming that such apprehension involves no mastery of concepts, that, for example, a subject who does not possess the concept *F* can nevertheless perceive or apprehend that *a* is *F*. Scheler considers two views, each of which he rejects:

[15] Hildebrand 1950 8

[16] Hildebrand 1976 23.

[17] Hildebrand 1950 8.

> It is asserted, first: in the content of natural perception nothing like a "meaning" occurs. The only thing I can perceive is a determinate, optical or other sensory content, e.g., the side-view of a house, these forms, lines, colours, surfaces, various of which can succeed one another in such a way that connections of anticipation and of memories between these views come about because of experience and training...
>
> Others say: no ! Perception contains more than this. It contains a *judgement*: one apprehends what is seen "as" a house, or "as" something which falls under the "general meaning" "house"...What an astonishing construction ! We continuously perceive a thousand things - but without a trace of such judging and asserting...
>
> The first theory "sensualises"...the meaning or better the meaning content which lies in natural perception. The second theory "logicises" natural perception and imputes to it something which it certainly does not contain[18].

The correct view, he seems to think, is that perceptual apprehension is shaped by the meanings the perceiver masters but does not necessarily involve any "judgement or subsumption of what is seen" under meanings[19]. From the fact that a perceiver does not actually subsume what is seen under concepts or meanings it does not follow that the perceiver does not master certain concepts. We very often perceive and perceptually apprehend that *p*, suggests Scheler, without any subsumption under concepts going on but we would not perceive or apprehend in the way we do if we did not master certain concepts. In such cases, he claims, our relation to meanings is like our relation to the rules we follow, for example the rules we follow in inferring, as opposed to the premises from which we infer.

Von Hildebrand's main reason for thinking that knowledge, of whatever variety, is not any sort of belief or conviction, is a claim made first of all by Reinach, perhaps the greatest of all the phenomenologists. Belief and conviction, like emotions, are, Reinach argues, *attitudes* (*Stellungnahmen*). These may vary in degrees and usually come in one of two polarly opposed kinds – belief and disbelief, positive and negative conviction, joy and sadness. Attitudes, on this view, are not always propositional : admiration, for example, is a non-thatish attitude. But knowledge is not any sort of attitude – it does

[18] Scheler 1957 472-3.
[19] Scheler 1955 360

not admit of degrees and has no polar opposite. Attitudes are reactions to what is known. Knowledge is no reaction. So knowledge wears the trousers. Beliefs and convictions are reactions to what we know. Von Hildebrand develops many of these claims in some detail in many different publications. His starting point is Reinach's 1911 account :

> There is an opposition running through this...class [conviction, striving, expecting] between positivity and negativity. We not only strive positively after something but may also struggle against it. In both cases [*Streben* and *Widerstreben*] we have a striving, but the two are, so to speak, of opposite sign. Now we find exactly the same in the case of conviction. So far we have naturally concentrated upon positive conviction; there is however, standing in opposition to this, a negative conviction, having a fully equal status....Both positive and negative convictions ...are... attitudes. The moment of conviction is common to the two, just as the moment of striving is common to positive striving for and to striving against something. It is this moment which separates the two types of conviction from other intellectual attitudes, e.g. from conjecture or doubt[20].

Reinach thinks that a conviction just is a belief-that. This is, I think, a mistake ; a conviction is a belief which has a high degree of certainty[21]. But beliefs, like convictions, are attitudes. What is the relation between beliefs and convictions, on the one hand, and knowledge ? Reinach's unhusserlian answer is as follows :

> Let us suppose that someone asserts that a flower is red, and that in order to convince ourselves of this we go to the place where the flower is to be found, and see that it is yellow. Thus we have approached the flower with the question whether it is truly red. Now with respect to this state of affairs there grows in us a negative conviction, a 'disbelief` that the flower is red. Both positive and negative convictions may relate to one and the same state of affairs;...[22]

[20] Reinach 1989 109.
[21] Cf. Mulligan 2013.
[22] Reinach 1989 109.

These beliefs or convictions are what he calls *Erkenntnisüberzeugungen*, convictions which are *based on* knowledge:

> I apprehend the being red of the rose; in this apprehension the state of affairs is presented to me, and *on the basis of the apprehension there develops in me the conviction of, or belief in, that state of affairs*. Conviction is, in this case, founded in apprehension; the former is the position which I take up, my receipt, so to speak, for that which apprehension offers to me[23].

The difference between apprehending and conviction, he claims, lies not only in the fact that the former unlike the latter allows of no degrees but also in the fact that the former is punctual and the latter a state, the sort of thing which endures (however long it exists; endurance is the mode of being of states as of things). Reinach's claims about the apprehension of the being red of the rose, like von Hildebrand's formulations, are ambiguous between the two views distinguished by Scheler. On one view, apprehension is no type of belief or conviction, involves no subsumption under concepts and does not depend on the meanings the perceiver masters. On the other view, the view Scheler favours, apprehension is no type of belief, involves no subsumption under concepts but does depend on the meanings the perceiver masters. If each of these two views is wrong, then the phenomenological version of the knowledge first view arguably loses much of its plausibility[24]. For if it does not work for what is arguably the simplest type of apprehension, perceptual apprehension, it is unlikely to work for the other cases the phenomenologists apply it to, such as apprehension of mathematical facts or apprehension due to testimony.

Not all convictions and beliefs can be reactions to what is known. What, then, one would like to ask the early phenomenologists, does their account have to say about such cases ? As far as I can see, they offer no worked out answer. But since they think that apprehension is opposed to illusion or deception (*Täuschung*) and to hallucination, and that correct belief, conviction and judgement are opposed to error (*Irrtum*) and that error and illusion are quite distinct phenomena, it would perhaps be in the spirit of their account to say that conviction and belief are reactions to what is known or to what is apparently known[25].

[23] Reinach 1989 120; emphasis mine- KM.

[24] Mertens (1927) discusses and develops the views of Reinach and von Hildebrand.

[25] Cf. Scheler 1955. Why « correct » rather than « true » belief, judgement and conviction ? The early phenomenologists follow Husserl : beliefs and judgements are correct or incorrect,

4. A Greek Folly ?

One early phenomenologist from Göttingen was convinced that Plato's view of knowledge does not make it out to be a species of belief but rather a primitive, indefinable phenomenon - the eminent historian of science, Alexander Koyré. In his 1945 *Introduction à Platon* he argues that the ideal reader of or listener to Plato's Socratic dialogues will come away with definite and positive, true philosophical conclusions. Perhaps unsurprisingly, these are often the conclusions of the early phenomenologists with whom Koyré studied in Göttingen before the Great War. One such conclusion is that knowledge is the "possession" of truth where "possession" does not mean belief or doxa[26]. There is a positive conclusion about the nature of knowledge which can be drawn from the *Theatetus* but which Theatetus himself has not been able to draw[27]. The conclusion that Theatetus the mathematician should have drawn is that

> la *science* qu'elle [la démonstration mathématique] nous donne (et qui peut être le *fondement* d'un jugement ou d'une « opinion ») est tout autre chose qu'une opinion — vraie ou fausse — qui peut être fondée ou infondée, qu'une conviction dont l'âme peut être possédée[28]

He should have seen that

> la circularité nécessaire de toute définition de la science nous révèle le caractère prééminent de cette notion. La définir est tout aussi impossible que « définir » celle de l'Être. Ou du Bien[29]

How, then, can one know what science is ? Koyré answers his question as follows :

---

propositions are true or false ; if the belief that *p* is correct, then it is correct *because* the proposition that *p* is true.

[26] Koyré's notes on lectures by Reinach in 1910 on Plato's philosophy and on Descartes have survived (Koyré 1910). Koyré is particularly concerned to show that the reader of the Socratic dialogues acquires axiological knowledge about « hierarchies » and « scales of value », that this or that value or good is higher or better than some other value or good. The echoes of Max Scheler and Nicolai Hartmann are unmistakeable. Cf. §6 below.

[27] Koyré 1962 74.

[28] Koyré 1962 76.

[29] Koyré 1962 77

> Justement de la même manière dont nous savons ce qu'est l'Être. D'ailleurs, Socrate nous l'a dit *expressis verbis :* la science n'est rien d'autre que la possession de la vérité. Et celle-ci n'est rien d'autre que la révélation de l'Être. Nous avons la science lorsque nous sommes dans la vérité, c'est-à-dire lorsque notre âme, en contact immédiat avec la réalité — avec l'être, — la reflète et la révèle à elle-même. Cet être, cette réalité — faut-il encore le dire? — n'est pas l'amas désordonné d'objets sensibles que le vulgaire (et le sophiste) appellent de ce nom. L'être vulgaire, mobile, instable et passager, n'est pas — ou est à peine — de l'être; il est, et il n'est pas, tout à la fois, et c'est pour cela justement qu'il n'est pas, et ne peut pas, être l'objet de la science, mais tout au plus de l'opinion. Non, l'être que nous avons en vue, c'est l'être stable et immuable de l'essence, que notre âme a contemplée jadis, ou, plus exactement, dont elle possède l'idée, vision dont elle se ressouvient — ou, du moins, dont elle peut se ressouvenir — maintenant, et dont demeurent dans l'âme des traces, des idées « innées »[30]

5. Against conjunctivism

Part of the background to the Göttingen versions of the knowledge first view (and to the account of knowledge given by another realist, Nicolai Hartmann) is the rejection of conjunctivism, in particular of Husserl's thorough-going conjunctivism, by one of the most influential of his early followers, the realist phenomenologist, Max Scheler.

One version of conjunctivism about *perceptual reports* is the view that such reports can be analysed into a conjunction of claims, one of which attributes a perceptual state which differs in no intrinsic way from a state of hallucination and, secondly, a claim to the effect that some suitable object or state of affairs is suitably related to the perceptual state. Conjunctivism about *perception* is the view that a perceptual episode consists of a perceptual state which differs in no intrinsic way from a state of hallucination and of a relation to an object or a state of affairs. Disjunctivism about perception is, then, to begin with, the view that conjunctivism is wrong[31]. If knowledge is a simple, unanalysable relation, then, it may seem that this view entails that conjunctivism about per-

[30] Koyré 1962 78.

[31] An early friend of the view now called disjunctivism is Hinton 1973. The expression „disjunctivism" is apparently due to Howard Robinson.

ceptual knowledge is wrong. Disjunctivism is, of course, more than the mere rejection of conjunctivism. But for present purposes no positive characterisation of disjunctivism is required.

The most important criticisms of conjunctivism in early phenomenology are due to Max Scheler. In 1915 Scheler formulates the view about normal perception he rejects by using an expression due to Pascal Engel's compatriot, Hippolyte Taine: normal perception is "une hallucination vraie", a hallucination which is true. According to this view, normal perception is something which is phenomenally indistinguishable from a hallucination and differs from it only in that the fact that "something real corresponds to it", in the fact that an existential judgement based on it is true. One version of the view, he adds, has it that the state which is phenomenally indistinguishable from an hallucination is caused by the presence of an objective stimulus of the right sort[32]. He objects that this view is incompatible with the "difference of essence between perception and illusion" [33]. Of course, a conjunctivist like Husserl can and does agree that there is a difference of essence between perception and illusion (and between knowledge and erroneous belief). But the conjunctivist view of the essential difference between perception and hallucination is wrong, Scheler thinks, because abnormal cases must be explained in terms of normal cases rather than the other way round. The normal case is not "a special case" of the abnormal case. Austin was to make a related claim: "talk of deception only *makes sense* against a background of general non-deception"[34].

Scheler rejects not only conjunctivism about perception but also conjunctivism about action (a rejection also to be found in the work of von Hildebrand). Scheler's analysis of action shows, he claims, that "it is a phenomenological unity and not composed of an inner act of the will and an external process of movement"[35]; it "cannot be dissolved into any sort of composition or succession of psychological experiences and bodily movements or processes"[36].

Perhaps unsurprisingly, Scheler rejects that version of conjunctivism about knowledge-that and apprehension which presents knowledge as a species of judgment or belief which satisfies certain condition. He does not argue against

[32] Scheler 1955 250.

[33] Scheler 1955 251.

[34] Austin 1962 11. Marion, who quotes this passage, traces the idea back to Prichard (Marion 2000 511, 325 ff..) Criticisms of the use of the normal-abnormal distinction in Oxford ordinary language philosophy also apply to what the phenomenologists do with the distinction.

[35] Scheler 1971 403.

[36] Scheler 1966 475. Scheler's account of action is to be found in Scheler 1966 127-172.

this sort of conjunctivism in anything like the way he argues against conjunctivism about perception and action. But he does state and endorse his own, alternative view: knowledge is irreducibly relational. In 1926 he distinguishes between "the most general concept of knowledge (*Wissen*)" and apprehending. Knowledge in the widest sense is "the end (*Ziel*, aim) of all apprehending". Knowledge as a possession, a state or disposition, is what the episode of apprehending aims at. In other words, although apprehending is ontologically prior to knowledge-that, is what brings it into being, the value of apprehending is determined by the value of knowledge-that. Knowledge must, Scheler also claims, be specified without any reference to "judgement, presentation (*Vorstellung*), inferring". It is an ontological relation (*Seinsverhältnis*, *Seinsbeziehung*), a relation of participation between entities and not any sort of spatial, temporal or causal relation. He seems at one point to call his account of knowledge as a relation of participation a definition[37]. But this is not a very happy use of the term since he does not analyse knowledge into components but rather specifies what sort of relation it is. And indeed a year or two later he writes that "knowledge is an ultimate, *sui generis*, and not further derivable ontological relation between two entities"[38].

6. Knowledge of Values and *Ought*

An account of knowledge which aims at completeness should arguably pay attention not only to the relation between discovery and enduring knowledge but also to the full variety of knowledge. This variety is not limited to the variety of *what is known* – mathematical, social, axiological, scientific etc. facts. Nor to what are sometimes called the different *sources* of knowledge, such as perception, intuition, understanding, proof and testimony. Knowledge may vary in a third way.

This can be seen by considering two different ways of understanding the ideas that knowledge has different sources and different objects. One might think that knowledge may have different sources but is itself always of the same lowest kind. Knowledge which arises out of perception is then, *qua* knowledge, in no way different from the sort of knowledge which is rooted in understanding or in calculation. Similarly, one might think that knowledge of arithmetic and knowledge of value differ only in their objects. But there is also the possibility that the differences between the sources and objects of

[37] Scheler 1960 203.
[38] Scheler 1995 188.

knowledge correlate with differences in types of knowledge, the possibility that knowledge of value and knowledge of arithmetic, say, differ intrinsically.

The accounts of knowledge given by the early, realist phenomenologists and by early Oxford philosophers take very seriously one aspect of the variety of knowledge, the variety of its objects. They aimed to give an account of what might be called theoretical knowledge and of non-theoretical knowledge which, in each case, puts knowledge first. By "non-theoretical" I mean knowledge of axiological and deontic facts, ethical, moral but also, for example, aesthetic. Within early, realist phenomenology this led to the development of the view that non-theoretical knowledge is not of the same lowest kind as theoretical knowledge. This view is to be found elsewhere in the Brentanian tradition but the commitment there to putting knowledge second led to a quite distinct account of the nature of non-theoretical knowledge. This disagreement between heirs of Brentano, as we shall see, parallels a contemporary disagreement about the nature of knowledge of value, a disagreement within the philosophy of mind rather than within mainstream epistemology.

The Göttingen-Oxford project of giving a knowledge first account of both theoretical and non-theoretical knowledge is clearly illustrated by Prichard's famous and influential 1912 paper "Does Moral Philosophy rest on a Mistake ?" and by a paper published one year earlier by the phenomenologist Alfred Brunswig, "Die Frage nach dem Grunde des sittlichen Sollens" . The question posed by Brunswig – does the moral or ethical ought have a ground or justification ? – is also the question addressed by Prichard. Indeed, many of the questions addressed by Prichard are also addressed by Brunswig. The mistake on which moral philosophy rests, according to Prichard, is the view that the demand to "have it *proved* to us that we ought to do" this or that is legitimate. But this demand, he says, is "illegitimate", there is no knowledge to be had which would satisfy the demand[39]. This illegitimate demand, he says, parallels another demand in the Theory of Knowledge, a demand concerning what I have called theoretical knowledge. He contends

> that the existence of the whole subject [Moral Philosophy], as usually understood, rests on a mistake, and on a mistake parallel to that on which rests, as I myself think, the subject usually called the Theory of Knowledge[40].

[39] Prichard 1912 36.
[40] Prichard 1912 21.

> [J]ust as we try to find a proof, based on the general consideration of action and of human life, that we ought to act in the ways usually called moral, so, we, like Descartes, propose by a reflexion on our thinking to find a test of knowledge, i.e. a principle by applying which we can show that a certain condition of mind was really knowledge[41].

He also calls such a test a "criterion" and says that the "search for this criterion and the application of it, when found, is what is called the Theory of Knowledge"[42].

Prichard's alternative to the vain project of trying to find a proof of what we ought to do is the claim that there is "an absolutely underivative or immediate" "apprehension" of moral obligations, of the rightness of actions[43].

Brunswig, too, rejects the demand for a ground of our particular obligations[44] and argues for a direct apprehension of our duties:

> The obtaining of the genuine moral or ethical (*sittlich*) ought is not something which is self-evident in virtue of the concept and value of the moral or ethical nor can it be indirectly deduced, it is rather certain for everyone in certain facts of consciousness. The unconditional obligation to act rightly is as a particular fact directly graspable (*erschaubar*)... [I]t is a state of affairs I apprehend[45].

This *apprehension* that my duty is to do this or that in turn grounds or justifies a *conviction* to this effect[46] but is not any such conviction .The fact apprehended, an ought-to-do, is "in a certain sense... *unprovable* but nevertheless completely certain thanks to the direct intuition every one has"[47]. Brunswig

[41] Prichard 1912 22.

[42] Prichard 1912 34. This characterisation of the theory of knowledge is also that given by Nelson (1908) who also argues that the search for a criterion of knowledge is and must be vain. The title of his book in English is : *On the so called Problem of Knowledge*. In the first part of the book Nelson sets out what he calls a general proof of the impossibility of the theory of knowledge (Nelson 1908 29-105). As Chisholm has pointed, out, referring to a later paper by Nelson: « It is instructive to compare what Nelson says here about theory of knowledge to what H. A. Prichard said about moral philosophy [in Prichard 1912] » (Chisholm 1979 53 n. 2.)

[43] Prichard 1912 27.

[44] Although duties cannot be grounded, Brunswig thinks, like Scheler, that a duty may have a partial ground, which is axiological : if I ought to F, this is in part because x is or would be valuable.

[45] Brunswig 1911 44.46.

[46] Brunswig 1911 44.

[47] Brunswig 1911 49, emphasis mine –KM.

distinguishes sharply between contingent and non-contingent ethical, moral and other axiological and normative facts. The fact that I ought to do this or that, he correctly points out, is a contingent fact, "an empirical fact, not a conceptual necessity"[48]. There are, of course, he thinks, like all the phenomenologists, non-contingent axiological and deontic facts. One such, he argues, is the fact that all acting *ought to be* moral or ethical. (Like other early phenomenologists, Brunswig is a fan of Sidgwick's distinction between *ought-to-do* and *ought-to-be*. Prichard, on the other hand, thinks that all *oughts* are *oughts-to-do*). A similar distinction seems to be implied by various remarks made by Prichard[49].

On one point our Göttingen and Oxford ethical-or-moral-knowledge-first philosophers differ. According to Brunswig my grasp of my duty is a type of practical experience. It is of course also, as we have seen, an apprehending. But it is not the apprehending peculiar to theoretical knowledge. It is "perhaps an act of feeling, of affective apprehending (*fühlenden Erkennens*)[50]". Prichard gives little sign of agreeing with Brunswig's suggestion. The apprehending of mathematical facts and of moral duties differ, it seems, on his view, only with respect to their objects and genesis. Like so many other English intuitionists, Prichard seems to think that intuition is always cold. It is however worth noting that Prichard frequently employs the verb "appreciate" when talking of the apprehension of duties: "the real nature of our apprehension or appreciation of moral obligations"; "appreciation [is] an activity of *moral* thinking"; he even refers to the "sense" that something is owing[51].

The idea that there is an affective apprehending, first published by Brunswig (and Reinach) was enthusiastically endorsed by many heirs of Brentano from around 1907/8. It rapidly came to be thought of as our primary mode of access, not (as Brunswig suggested) to duties or oughts, but rather to value, the value of objects and persons and states of affairs. The view comes in two versions, one for friends of the knowledge first option, the other for enemies thereof.

Husserl and (late) Meinong both argue that in certain optimal circumstances emotions disclose value. Versions of their view have become very popular within one pocket of the philosophy of mind, the philosophy of emotions. Episodic emotions or affects, it is said there, can disclose or reveal

[48] Brunswig 1911 49.
[49] Prichard 1912 28.
[50] Brunswig 1911 47.
[51] Prichard 1912 27, 28, 29.

value[52]. They are able to do this above all because of a property emotions share with beliefs and judgments. Emotions, like beliefs and judgings, are either correct or incorrect. Thus it is argued that an emotion which is correct and which satisfies certain other conditions counts as knowledge of value. I apprehend the injustice of a situation through the emotion of indignation provided my indignation is correct. Thus the view of theoretical knowledge which puts belief (suitably qualified) first has an exact counterpart, the view of non-theoretical knowledge which puts emotions (suitably qualified) first.

Early phenomenological friends of the primacy of knowledge make two claims, as we have seen, which yield objections to this view. First, indignation is triggered by knowledge or apparent knowledge of injustice. This knowledge or apparent knowledge cannot be constituted by an emotion if a regress is to be avoided. Second, indignation, like all emotions, is a reaction and an attitude. But knowledge is neither a reaction nor an attitude. The correct alternative, according to Brunswig, Reinach, Scheler, von Hildebrand and Hartmann, is that the affective apprehension of value involves no emotions but rather *Wertfühlen*, not feelings or emotions but *feeling*: we feel the injustice of a situation. But such feeling of a value, being struck by injustice, shamefulness, dumpiness, elegance or funniness, is not itself either a *pro* or a *contra* stance or attitude; it is what triggers such attitudes, in particular emotions and beliefs. Feeling value, being struck by value, is a type of episode which corresponds to the state or disposition ascribed when we say of someone that he has no sense of or for beauty or injustice, that sensibility to this or that type value is not part of his make-up, that he is blind to this or that range of values.Theoretical apprehension, then, as before, differs radically from non-theoretical apprehension but not because the latter involves emotions. The (apparent) feeling of value is prior to emotions and belief, just as (apparent) theoretical knowledge is prior to belief[53].

Many of Brentano's heirs, then, seem happy to allow that apprehending comes in different, lowest kinds. But, as we have seen, friends and enemies of the knowledge first view give rival accounts of what these kinds are. Must a philosophy of knowledge assert either that knowledge always comes first or that it always comes second ? Scheler is a philosopher who puts knowledge first in his accounts of most kinds of knowledge but allows for one case where knowledge comes second. The case in question concerns knowledge of value-relations. Any account of knowledge of value has to give an account of

[52] Cf. Tappolet 2000, Johnston 2001, Deonna & Teroni 2012.
[53] Cf. Mulligan 2007, 2009, 2010.

knowledge of value-relations, of one thing or state of affairs being worse than another, of relations of height between value (justice is higher in value, more important than prettiness)[54]. According to Scheler, knowledge of relations of height between values is constituted by a type of preferring, preferring which is given as being correct or self-evident[55]. On this matter, then, he finds himself obliged to agree with Brentano, the grandfather of all twentieth century philosophies which put knowledge in second place[56].

7. References

Austin, J. L. 1962 *Sense and Sensibilia*, Oxford University Press.

Bassenge, Fr. 1955 „Über Fragen der Logik", *Deutsche Zeitschrift für Philosophie*, 3:4, 477-496.

Brunswig, A. 1911 „Die Frage nach dem Grunde des sittlichen Sollens", *Münchener Philosophische Abhandlungen*, Lipps Festschrift, Leipzig: Barth, 26-50.

Chisholm, R. 1979 „Socratic Method and the Theory of Knowledge", *Vernunft, Erkenntnis, Sittlichkeit. Internationales philosophisches Symposion Göttingen, vom 27.-29. Oktober 1977 aus Anlass des 50. Todestages von Leonard Nelson*, Hamburg: Felix Meiner Verlag, 37- 54.

Cook Wilson, J. 2002 (1926) *Statement and Inference*, reprint, Bristol: Thoemmes Press.

Deonna, J. & Teroni, F. 2012 *The Emotions : A Philosophical Introduction*, New York: Routledge.

Dutant, J. 2010 *Knowledge, Methods and the Impossibility of Error*, Geneva PhD.

Hildebrand, D. von 1950 *Der Sinn philosophischen Fragens und Erkennens*, Bonn: Peter Hanstein Verlag

[54] Cf. note 25.

[55] Since Scheler also thinks that such preferring is prior to all feeling of value, his entire account of the affective apprehension of value is, contrary to some of his rhetoric, at bottom a knowledge second account.

[56] I am grateful to Julien Deonna for many brief but stimulating discussions of the history of the view that knowledge is justified true belief and for comments on an earlier version of this paper; to Arturs Logins for discussion of the project he is pursuing with Paolo Crivelli on knowledge first views in ancient philosophy; and to Alessandro Salice, Barry Smith and Peter Simons for their suggestions.

— 1960 *What is Philosophy ?* Milwaukee: The Bruce Publishing Company (1973 Chicago); German tr. 1976 *Was ist Philosophie ?*, GW I, Regensburg: Josef Habbel, Stuttgart: Kohlhammer.

Hinton, J. M. 1973 *Experiences*, Oxford University Press.

Johnston, M. 2001 "The Authority of Affect", *Philosophy and Phenomenological Research*, LXIII, I, 181-214.

Koyré, A. 1910 [Koyré's notes on a lecture by Reinach in 1910, „Platons Philosophie"] http://nasepblog.files.wordpress.com/2012/11/reinach-platon-1910.pdf

— 1962 (1945) *Introduction à la lecture de Platon,* suivi de *Entretiens sur Descartes*, Paris : Gallimard.

Mertens, P. 1927 *Zur Phänomenologie des Glaubens*, Fulda.

Marion, M. 2000 „Oxford Realism: Knowledge and Perception", *British Journal for the History of Philosophy*, 8(2): 299–338 & 8(3) 485–519.

— 2002 'Introduction', in Cook Wilson 2002, v–xxvii.

— 2003 « Husserl et le réalisme britannique », D. Fisette & S. Lapointe (eds.) *Aux origines de la phénoménologie. Husserl et le contexte des Recherches Logiques*, 255–286. Québec/Paris: Presses de l'Université de Laval/Vrin.

Mulligan, K. 2006 "Soil, Sediment and Certainty", *The Austrian Contribution to Analytic Philosophy*, ed. Mark Textor, London: Routledge (London Studies in the History of Philosophy), 89-129

— 2007 "Intentionality, Knowledge and Formal Objects", *Disputatio*, Vol. II, No. 23, November 2007, 205-228.

— 2009 „On Being Struck by Value – Exclamations, Motivations and Vocations", *Leben mit Gefühlen. Emotionen, Werte und ihre Kritik*, ed. Barbara Merkel, Paderborn: mentis-Verlag, 141-161.

— 2010 "Emotions and Values", *The Oxford Handbook of Philosophy of Emotion*, ed. P. Goldie, Oxford University Press, 475-500.

— 2013 "Acceptance, Acknowledgment, Affirmation, Agreement, Assertion, Belief, Certainty, Conviction, Denial, Judgment, Refusal & Rejection", ed. Textor, M. *Judgement and Truth in Early Analytic Philosophy and Phenomenology,* (History of Analytic Philosophy Series), London: Palgrave/Macmillan, 97-137.

Mulligan, K. & Smith, B. 1986 "A Relational Theory of the Act", *Topoi,* 5/2, 115-130.

Nelson, L. 1908 *Über das sogenannte Erkenntnisproblem*, Göttingen: Vandenhoeck & Ruprecht.

Prichard, H. A. 1909 *Kant's Theory of Knowledge*, Oxford University Press.

— 1912 "Does Moral Philosophy Rest on a Mistake ?", *Mind*, 21/81, 21-37.

Ramsey, F. 1991 *Notes on Philosophy, Probability and Mathematics*, ed. M-C. Galavotti, Naples: Bibliopolis.

Reinach 1989 (1911) "Zur Theorie des negative Urteils", in eds K. Schuhmann & B. Smith, *Sämtliche Werke*, Textkritische Ausgabe, 2 vols., Vol. 1, *Die Werke*, Munich: Philosophia Verlag, 95-140. An English translation by B. Smith, "On the Theory of Negative Judgement", is to be found in ed. B. Smith, 1982 *Parts and Moments. Studies in Logic and Formal Ontology*, Muinich: Philosophia Verlag, 315-377.

Scheler, M. 1955 (1915) "Die Idole der Selbsterkenntnis" , Vom Umsturz der Werte, Gesammelte Werke, 3, Bern: Francke, 213-292, (English tr. Scheler 1973 "The Idols of Self-Knowledge", in *Selected Philosophical Essays*, Evanston: Northwestern University Press 3-97).

— 1957 (1933) "Lehre von den drei Tatsachen", *Schriften aus dem Nachlass*, Band I, Gesammelte Werke, X, 431- 502, Bern: Francke.

— 1960 (1926) "Erkenntnis und Arbeit", *Die Wissensformen und die Gesellschaft*, Gesammelte Werke, 8, Bern: Francke, 191-382.

— 1966 (1913-16) *Der Formalismus in der Ethik und die material Wertethik*, Gesammelte Werke, 2, Bern: Francke.

— 1971 (1914) "Ethik. Eine kritische Übersicht der Ethik der Gegenwart", *Frühe Schriften*, Gesammelte Werke, I, Bern: Francke, 371-409.

— 1995 (1927/8) "Idealismus-Realismus", *Späte Schriften*, Gesammelte Werke, 9, Bonn: Bouvier, 183-242.

Smith, B. 1984 "Acta cum fundamentis in re", *Dialectica*, 38, 157-178.

Spiegelberg, H 1933 „Sinn und Recht der Begründung in der axiologischen und praktischen Philosophie", *Neue Münchener Philosophische Abhandlungen*, Leipzig: Barth, 100-42.

Tappolet, Ch. 2000, *Emotions et Valeurs*, Paris : Presses Universitaires de France.

Teroni, F. 2005 "Meinong on Memory", *Early Analytic Philosophy: The Austrian Contribution*, Textor, M. (ed.), Routledge, 64-88.

Williamson, T. 2000. *Knowledge and its Limits*, Oxford: Oxford University Press.

PART THREE

Mind and Language

26

How to Account for the Oddness of Missing-Link Conditionals *

IGOR DOUVEN

Abstract Conditionals whose antecedent and consequent are not somehow internally connected tend to strike us as odd. The received doctrine is that this felt oddness is to be explained pragmatically. Exactly how the pragmatic explanation is supposed to go has remained elusive, however. This paper discusses recent philosophical and psychological work that attempts to account semantically for the apparent oddness of conditionals lacking an internal connection between their parts.

Consider this conditional:

(1) If Arsenal wins next year's Champions League final, Great Britain will leave the European Union.

*It is a pleasure and an honor to contribute to this Festschrift celebrating Pascal Engel's sixtieth birthday. When Pascal began working as an analytic philosopher in France, he must have felt like the villagers in the famous French comic strip Astérix, who were surrounded by Roman enemy troops who vastly outnumbered them but nevertheless firmly stood their ground. If this is no longer an accurate description of the situation of any European analytic philosopher, that is to a great extent owing to Pascal, who has done more for the dissemination of analytic philosophy on the European continent than anyone else. I dedicate this essay, in friendship and with affection, to Pascal.

When we read this sentence and try to interpret it, we notice that we try to find some plausible connection between the possible event of Arsenal winning the Champions League final next year and the further possible event of Great Britain leaving the European Union. In this case, it is hard to think of a connection: it seems quite unrealistic to suppose that the occurrence of the former event could give rise to the occurrence of the latter. If we were to overhear someone asserting (1), we might want to point out that whether Great Britain decides to leave the European Union will have nothing to do with which team wins next year's Champions League final. Indeed, laypeople will be inclined to reject (1) as being false. Importantly, this inclination will be independent of what one believes about Arsenal's chances of winning next year's Champions League final as well as of one's beliefs about the possibility of Great Britain leaving the European Union. Even if Arsenal does win next year's Champions League final and Great Britain does leave the European Union, (1) seems defective: it seems to assert the existence of a link between the two events that—we are highly confident—does not exist.

That the absence of a plausible connection between the antecedent and consequent of (1) gives reason to reject that sentence as *false* is, by modern semantic lights, a naive verdict. On none of the currently popular semantics for conditionals is the holding of a connection between a conditional's antecedent and consequent necessary for the truth of that conditional. While it is widely acknowledged that conditionals whose antecedent and consequent are not internally connected—call these "missing-link conditionals"—tend to strike us as odd, the felt oddness is, according to modern semantic theorizing, to be explained along pragmatic lines. Broadly, the idea is that the assertion of a conditional generates the implicature that there is an internal connection between antecedent and consequent. If no such connection exists, that makes the assertion infelicitous, but not—or not necessarily at least—because the asserted sentence is false.

As stated, this is the *broad* idea. Proponents of this idea have been rather unhelpful in providing further details. We have a pretty good understanding of how an assertion of

(2) Some Republicans have a rather feeble grasp of reality.

generates the implicature that not all Republicans have a rather feeble grasp of reality. If a speaker of (2) believed that all Republicans have a rather feeble grasp of reality, he or she should have asserted *that* instead of (2): in doing so the speaker would have made a stronger, more informative claim, and would

thereby have been more cooperative. But how does an assertion of (1) generate the implicature that Arsenal winning next year's Champions League final is somehow relevant to whether Great Britain will leave the European Union?

It might be thought that the word "relevant" in the previous sentence actually holds the key to an answer, in that the answer must have something to do with Grice's [1989] Maxim of Relevance. However, it is unclear how this maxim could help us out. According to the Maxim of Relevance, the cooperative speaker makes his or her contribution to a conversation a relevant one. While it is probably true that missing-link conditionals will, when asserted, bear little relevance to whatever conversation is going on, that this is because there is no relevant connection between their antecedent and consequent is not something that follows from the Maxim of Relevance. Nor, to the best of my knowledge, does this follow from any other principle of contemporary pragmatics.

Some might be willing to accept as a primitive fact that uses of the word "if" generate the *conventional* implicature that there is a connection between antecedent and consequent, just as some have held that uses of the word "but" generate the conventional implicature that there is a contrast between the conjuncts connected by that word. Grant that there are any conventional implicatures at all (which is contentious; see Bach [1999]). Then it is still to be noted that, just as the aforementioned claim about "but" seems false—there does not appear to be any contrast in "He walks slowly, but he walks"—the word "if" does not appear to generate any implicature to the effect that there is some connection in sentences like this:

(3) If the Pope converts to Islam, the US will keep spying on its allies.

In the literature, various types of conditionals have been identified as "non-conditional" or "unconditional" conditionals (Lycan [2001], Merin [2007]). These special types of conditionals include so-called noninterference conditionals (Bennett [2003], Burgess [2004]), of which (3) is an instance. Rather than expressing anything conditional, they state the truth of the consequent—unconditionally!—and use the conditional construction to convey the inevitability of that truth.

Naturally, those trying to devise a semantics for conditionals are in their right to exclude noninterference conditionals or any other special type of conditional as being in the purview of the envisioned semantics: it is a priori rather implausible to hold that there must be one semantics for "conditional" and "unconditional" conditionals alike. Moreover, given how little progress has been made in over two thousand years of theorizing about conditionals, it

is ambitious enough, at least for a while, to try and come up with a semantics for *normal* conditionals (or "conditional conditionals," as one might call them) that is both theoretically satisfactory and empirically adequate. And theorists aiming to explain the oddness of missing-link conditionals along pragmatic lines might likewise want to limit their account to normal conditionals. Those who want to maintain that the oddness is due to a conventional implicature would then have to accept that "if" is polysemous: sometimes it has a meaning that generates the conventional implicature that there is a connection between two propositions, and sometimes it has a meaning—as for instance in (3)—that does *not* generate that implicature. Pretheoretically, this makes as little sense as if someone were to maintain that "but" has different meanings in "Bert will come but Joan won't" (in which it does signal a contrast) and "He walks slowly, but he walks" (in which "but" does not signal a contrast): we simply do not sense any difference in the meaning of "if" as it occurs in (1) and as it occurs in (3).

Might we be able to account *semantically* for the oddness of missing-link conditionals after all? That thereby we would go against all popular semantics for conditionals is not much of an objection, given that none of these semantics is really all that popular. But postulating a link between a conditional's parts as being necessary for the conditional's truth raises the question of what that link might be. Prima facie, candidate-answers abound: the link could be epistemic, causal, explanatory, inferential, to mention just the most obvious candidates. On closer inspection, however, all such suggestions seem to run into trouble.

For instance, it may well be true that in all conditionals we deem reasonable there is an epistemic connection between antecedent and consequent, for instance, in that coming to know or rationally believe the antecedent puts us in a position to know or rationally believe the consequent. But this does not seem to get to the heart of the matter. We can easily imagine a context in which the conditional

(4) If the department is going to hire Harriet, Sue will leave the department as soon as she can.

is a perfectly reasonable assertion. Suppose it is common knowledge that Harriet and Sue hate each others' guts. Then coming to know that the department is going to hire Harriet may well put one in a position to know that Sue will leave the department as soon as she can. However, it seems that if so, then that it is due precisely to there being a more fundamental connection—plausibly

a causal connection in this case—between the event of Harriet being hired by the department and Sue wanting to leave the department. Coming to know the antecedent of (4) puts us in a position to know its consequent for the reason that we perceive this more fundamental connection to obtain.

On the other hand, the conditionals

(5) (a) If $a + 1 = 7$, then $a + 2 = 8$.

(b) If the Seychelles are ruled by a king, then the Seychelles are a monarchy.

seem unassailable in every way. Still, there can hardly be said to exist a causal connection between their antecedent and consequent. Nor does there seem to be an explanatory connection between their parts (if such a connection is different from a causal connection, which some will deny). Rather, in these cases, the connection seems to be that of a deductive inference.

Examples like (5 a) and (5 b) have led some philosophers to hold that the connection between antecedent and consequent must be one of entailment.[1] But that cannot generally be the case. To see this, just consider (4) again. Surely this conditional could be pretheoretically true even if there is a negligible but nonzero chance that Sue will stay in the event that Harriet is hired.

So then how are we to account for the oddness of missing-link conditionals? It is hard to think of a type of connection with both the right specificity (that does not seem to capitalize on a deeper connection) and the right generality (that works uniformly at least for all normal conditionals).

Shortly after I had started thinking about how to account for the felt oddness of missing-link conditionals, and while becoming more and more frustrated with what current semantic and pragmatic theories have to offer on this point,[2] I read a draft of Pascal Engel's [2014]. In this paper, Engel considers the possibility of conceiving of identity in the way that philosophers of mind have proposed to conceive of mental states, to wit, as a second-order functional property. It is now generally accepted that to be in pain, or to be hungry, or happy, or angry, is to have one of a number of properties which naturally belong together in that they all play the same functional role: the pain role, or the hunger role, et cetera. Wright [1996], Lynch [2009], and others have

[1] The Stoic philosopher Chrysippus is famous for having defended the view that a conditional is true iff the falsity of its consequent is incompatible with the truth of its antecedent (Kneale and Kneale [1962, Ch. 2]).

[2] And getting more and more frustrated with more general shortcomings of current semantics for conditionals. I catalogue my most serious misgivings in Douven [2011] and [2013].

proposed that we conceive of truth in a parallel fashion. For instance, Wright holds that the truth predicate is implicitly defined by a list of platitudes concerning truth, such as that truth is correspondence to reality and that truth is stable under the acquisition of further information. In Wright's view, these platitudes are multiply realizable: they can be fulfilled by different first-order properties (like verifiability and warranted assertability) in different domains.

Engel makes the original observation that it is worth considering extending this approach to identity as well. As he points out, if identity can be thought of as a second-order functional property, that would yield an elegant solution to many vexing problems concerning identity, like the well-known problems of constitution (what are we to say about the relation between a statue and the lump of bronze that makes it up?) and change over time (what are we to say about the the relationship between the different stages of Theseus' ship?). He notes that if this approach is taken, we will need an implicit definition of the identity predicate that does justice to the various features our identity judgments are known to have, such as, for instance, being context-sensitive and sometimes failing to be transitive. He argues that the only higher-order characterization that seems to fit the bill is one that defines the identity predicate in terms of similarity in all contextually relevant respects, specifically by stating that two things are identical in a given context iff they are highly similar in all respects relevant in that context. However, while he acknowledges the attraction of this idea, he rejects it in the end because it seems to render identity an epistemic rather than a logical property.

Not all philosophers may regard this as a shortcoming of the contemplated approach to characterizing identity (see, e.g., Douven and Decock [2010] and Decock and Douven [2012]). But that is not the issue now. The present issue is how to account for the oddness of missing-link conditionals. And reading Engel's paper inspired the thought that the connection between antecedent and consequent that a conditional seems to require if it is to be true could also be a second-order functional property, realized by one first-order property in some conditionals, by another first-order property in other conditionals, by a third in again other conditionals, and so on, perhaps. Put differently, the idea is that the truth of a conditional requires the existence of a "connector," linking the conditional's antecedent and consequent, where this connector is to be thought of as a second-order functional property that may be realized by various different first-order properties; for instance, some true conditionals may be true due to the presence of a causal link between their antecedent and consequent—their antecedent and consequent have the first-order property of being causally related—and other true conditionals may be true due to

the presence of an inferential link between their antecedent and consequent—their antecedent and consequent have the first-order property of being inferentially related—and so on.

That is to say, *in principle* there can be many different first-order properties playing the connector role. However, discussions, and joint work resulting from these discussions, with Karolina Krzyżanowska and Sylvia Wenmackers led to a semantics for conditionals that suggests that the variety of realizer properties that must be countenanced to account for the data pertaining to our use of conditionals may still be of a rather limited variety. More specifically, Krzyżanowska, Wenmackers, and Douven [2014] propose a semantics for conditionals that makes the presence of an inferential connection between antecedent and consequent a requirement for the truth of a conditional. Where earlier proposals in this vein erred, according to that paper, was in thinking that the inferential connection must always be of the same variety, namely, deduction. Instead, Krzyżanowska, Wenmackers, and Douven's proposal capitalizes on the fact that there are other varieties of inference, most notably, induction—roughly, inference based on statistical considerations—and abduction—roughly, inference based on explanatory considerations—that may hold between a conditional's antecedent and consequent, and their paper argues that the holding of any of these inferential relations between antecedent and consequent may suffice to render a conditional true. Still more specifically, on the proposed semantics—dubbed "inferentialism" in Douven et al. [2014]—a conditional is true in a given context iff its consequent follows from its antecedent plus (possibly) contextually relevant background knowledge, where "follows" is understood in a broad sense so as to encompass all of the aforementioned types of inference. Indeed, the proposal explicitly allows for the possibility that the consequent follows from antecedent and background knowledge via a number of inferential steps. In that case, the conditional is true, provided all steps are valid either deductively, inductively, or abductively.[3] In this proposal, (5 a) and (5 b) emerge as true—in virtue of the deductive-inferential connection between their antecedent and consequent—but so may (4): even if its consequent does not follow deductively from its antecedent, it may well follow abductively, in that the department's hiring Harriet explains best, or would explain best, Sue's leaving the department.

[3]This is the core of the proposal in Krzyżanowska, Wenmackers, and Douven [2014]. To this core, two clauses are added to take care of certain special cases. The details need not detain us here, except that for later purposes it should be mentioned that for a conditional to be true, its consequent should not *just* follow from background knowledge alone (unless it also follows from the antecedent alone).

Krzyżanowska and coauthors are aware that while we have a formally precise account of the notion of deductive consequence—to wit, classical logic—there are no accounts of inductive or abductive consequence that are even nearly as precise and well worked out. Researchers both in the logic and in the AI community are presently working to fill these lacunae. But, as Krzyżanowska and coauthors note, for now we may as well define inferentialism by saying that, according to it, a conditional is true iff there is a sufficiently strong (i.e., a strong, but not necessarily conclusive) argument from its antecedent plus relevant background knowledge to its consequent. Insofar as it can be vague whether a conditional is true or not, the vagueness inherent in the foregoing characterization of inferentialism (through the vagueness inherent in the expression "sufficiently strong argument") may be considered to be an advantage rather than a weakness.

It is to be emphasized that the proposal explicitly leaves open the possibility that the variety of realizer properties is actually larger than suggested here, and thus that more types of inferential connections, or even connectors of a noninferential type, must be taken on board if the semantics is to provide truth conditions for all normal conditionals. However, research carried out so far indicates that, as it stands, the proposed semantics can account for at least the bulk of normal conditionals.

As Krzyżanowska, Wenmackers, and Douven [2014] point out, inferentialism has some nice theoretical features. For one, it does not suffer from analogues of the paradoxes of material implication: the truth of a conditional does not simply follow from the falsity of its antecedent or from the truth of its consequent, for in neither case need there be an inferential connection between antecedent and consequent. For another, inferentialism gets the Or-to-if principle—which allows us to reason from "Either φ or ψ" to "If not φ then ψ"—right precisely in the kind of cases in which it is intuitively right. If all we know is that either Jim or Hank (or both) committed the crime, then it is eminently reasonable to infer that if Jim did not commit the crime, then Hank did. But now suppose we know that Jim committed the crime. Even though in that case we also know that either Jim or Hank (or both) committed the crime, we are no longer inclined to infer that if Jim did not commit the crime, then Hank did. Krzyżanowska and coauthors show that inferentialism is able to differentiate between the two kinds of cases: in both kinds of cases the disjunction is true, but only in the first kind of case is the conditional true. For a third, inferentialism validates in precisely the right kind of cases the Import–Export principle, according to which "If φ, then if ψ, then χ" and "If φ and ψ, then χ" are logically equivalent; see Krzyżanowska, Wenmackers, and Douven [2014,

Sect. 2] for the details. In fairness, it must be added that, as Krzyżanowska and coauthors also note, modus ponens is not strictly valid on their proposal. This is because induction and abduction are so-called ampliative forms of inference, in the sense that the truth of their premises does not ensure the truth of their conclusion, so that a true conditional with a true antecedent may have a false consequent. However, Krzyżanowska and coauthors argue that while abduction and induction are not *guaranteed* to preserve truth, they may still be taken to be highly reliable guides to the truth: at least, in daily practice we trust them to preserve truth with high probability. Supposing this practice to be justified, modus ponens may be assumed to be still a highly reliable inferential principle, and that may be all there is to our intuition that the principle is valid.

Theoretical support of a somewhat different variety comes from consideration of a problem case presented in Gibbard [1981], which Gibbard and other authors had taken to be strong evidence for the view that conditionals do not express propositions. Krzyżanowska and coauthors show that inferentialism yields an elegant solution to this problem that does not require abandoning that view. In the case that Gibbard presents, two speakers have putatively unassailable and equally strong evidence for conditionals that, according to the principle of Conditional Non-Contradiction (CNC), are jointly inconsistent, that is, for conditionals of the forms "If φ then ψ" and "If φ then not ψ," respectively. Indeed, in Gibbard's view, those speakers' evidence for the conditionals that they assert is so strong and so equal in quality that we cannot but conclude that if either of them has asserted a true conditional, then both have. However, in view of CNC, that cannot be right: at most one of the conditionals can be true. From this, Gibbard concludes that neither conditional is true, and more generally that conditionals lack truth conditions and thus do not express propositions.

Gibbard is not alone in his endorsement of non-propositionalism about conditionals; see, for instance, Adams [1965], [1975], Edgington [1995], and Bennett [2003]. But non-propositionalism faces some well-known problems, not the least of which is the problem of explaining how conditionals connect to the rest of our language: the logical operators that we use to build complex sentences from simpler ones are all *propositional* operators, taking only *propositions* as their arguments; if conditionals do not express propositions, they cannot occur as constituents of complex sentences, which they sometimes do (even if sentences embedding conditionals are not *very* frequent in quotid-

ian language).[4] As Krzyżanowska and coauthors show, given inferentialism, one can explain what is going on in Gibbard's problem case without abandoning either CNC or the view that conditionals express propositions. Their main point is very simple, to wit, that while one of the conditionals Gibbard presents is true relative to the background knowledge of the speaker who asserts that conditional, the other is true relative to the background knowledge of the speaker who asserts *that* conditional. This requires that we read CNC as stating that conditionals of the forms "If φ then ψ" and "If φ then not ψ" can never be jointly true relative to the same background knowledge, but once one considers a semantics for conditionals that makes the truth of conditionals relative to background knowledge, that is the natural reading of CNC anyway.

Krzyżanowska, Wenmackers, and Douven [2014] provide some empirical evidence for the correctness of their analysis of Gibbard's case. In this analysis, the two conditionals at issue embody different inferential connections between antecedent and consequent: in one, the connection is deductive—the consequent follows deductively from the antecedent plus the speaker's background knowledge—while in the other, the connection is abductive, in that the consequent best explains the antecedent in light of the background knowledge of the speaker who asserted that conditional. In earlier empirical work on so-called evidential markers, Krzyżanowska, Wenmackers, and Douven [2013] had (not really surprisingly) found "probably" to be a marker of uncertain inferential connections. In particular, it was shown that inserting "probably" in the consequent of a conditional tends to raise the degree of assertability of that conditional if the conditional embodies either an inductive or an abductive inferential connection, whereas it tends to lower the degree of assertability of the conditional if the conditional embodies a deductive inferential connection. Applying this finding to Gibbard's case and a number of similar cases modeled after Gibbard's, an experiment reported in Krzyżanowska, Wenmackers, and Douven [2014] revealed that the conditional in Gibbard's case that, according to these authors' analysis, embodies a deductive inferential connection relative to the relevant speaker's background knowledge is generally found to be less assertable with "probably" inserted in its consequent than without it, while the opposite was found to hold for the conditional that, in the said analysis, embodies an abductive inferential connection relative to the background knowledge of the person asserting the conditional. (In the experiment, participants were informed about the back-

[4]For a summary of the main attempts to solve this problem, and some reasons for believing these attempts fail, see Douven [2011].

ground knowledge available to a speaker when they were asked to rate the degree of assertability of the conditional asserted by the given speaker.)

As stated, this provides empirical support for Krzyżanowska, Wenmackers, and Douven's [2014] diagnosis of Gibbard's problem case. Empirical support for inferentialism in general was recently obtained in experimental work reported in Douven et al. [2014]. In this study, participants were presented with a soritical color series of fourteen patches, going from clearly green to clearly blue. They were then asked to evaluate conditionals pertaining to these patches. Specifically, they were asked to judge as true, false, or neither true nor false conditionals of the form "If patch number i is X, then so is patch number j," where depending on the test condition X was either "green" or "blue" and, depending on a further test condition, either $j \in \{i-2, i-1, i+1, i+2\}$ or $j \in \{i-3, i-1, i+1, i+3\}$. In this paper, Douven and coauthors derive, for all the main semantics as well as for inferentialism, predictions about which factors will influence the participants' answers. For instance, according to non-propositionalism, nothing matters except the fact that all sentences to be evaluated by the participants are conditionals: that predicts that all answers will be in the "neither true nor false" category. To give another example, the material conditional account—according to which a conditional has the same truth value as its material counterpart (so is false if its antecedent is true and its consequent is false, and is true otherwise)—predicts that only the ranks in the soritical series of the patches referred to in the antecedent and consequent matter (given that these ranks determine the truth values of the conditional's parts). They derive similar predictions for Stalnaker's semantics, according to which a conditional is true iff its consequent is true in the closest possible world in which the antecedent is true, and for various semantics that allow conditionals to take on all three truth values that the participants were given as answer options. As for inferentialism, they show that it implies that participants' answers will be determined by three factors, to wit, (i) the position of the consequent patch relative to the antecedent patch (is the consequent patch located to the left or to the right of the antecedent patch in the soritical series?), (ii) the number of patches lying in-between the antecedent patch and consequent patch, and (iii) the absolute location of the consequent patch in the series of fourteen patches; together these factors determine whether there is a strong argument from antecedent to consequent, for the conditionals in the materials of the experiment. Douven et al.'s [2014] analysis of the data shows that inferentialism provides the best model by far of those data.

To be sure, Douven et al.'s [2014] study used an experimental paradigm of a very specific type and with very specific materials. However, it is not

difficult to conceive of further and rather different experiments that can help us decide between the various semantics for conditionals that are currently on offer. A main distinguishing feature of inferentialism is that it does not validate the inferential principle—commonly called "Centering"—that allows us to infer the conditional "If φ then ψ" from the conjunction "φ and ψ," a principle that *is* valid given *any* of the main semantics for conditionals. So, for instance, given that Obama is the president of the United States and the coming Winter Olympics will be held in Russia, Centering allows us to infer that if Obama is the president of the United States, then the Winter Olympics will be held in Russia, as well as that if the Winter Olympics will be held in Russia, Obama is the president of the United States. My guess is that few people will want to endorse either of these conditionals. Obviously, however, there is no need to go by my guesses, for it should be easy enough to subject the claim to experimental testing (which is in fact on my to-do list).

Another possible experimental approach is to concentrate more directly on the link that, according to inferentialism, exists between arguments and conditionals. If a conditional is true precisely if there is a strong, even if perhaps nonconclusive, argument connecting antecedent and consequent, then people should be expected to endorse "If φ then ψ" to the extent that they judge "φ, therefore ψ" to constitute a strong argument, supposing both the conditional and the argument are evaluated against the same background. Again, it should be quite straightforward to verify (or falsify, as the case may be) this consequence of inferentialism experimentally. (Another item on my to-do list.)

Although the basic idea of inferentialism dates back to antiquity (see note 1), in its present form the position is still in its infancy. For that reason alone, it is not surprising that inferentialism calls not only for further empirical work, but for attention to some outstanding theoretical issues, the two most pressing of which I want to mention here. First, in their experiment, Douven and coauthors found only a relatively small percentage of "neither true nor false" responses (around 10 percent). In other experimental work, however, such responses were seen to be more prevalent (see, e.g., Barrouillet, Gauffroy, and Lecas [2008], Politzer, Over, and Baratgin [2010], and Baratgin, Over, and Politzer [2013]). Specifically, it was seen in earlier experiments that, at least for particular types of materials, people tend to evaluate conditionals with false or indeterminate (neither true nor false) antecedents as indeterminate. (These data are often collectively referred to as "the defective truth table data.") As it currently stands, inferentialism recognizes only the classical semantical values. Naturally, if it is to be both a general semantics for conditionals and

empirically adequate, inferentialism will have to accommodate the defective truth table data, and so will have to be extended to incorporate indeterminacy as well.

There is more than one way to go here, but a prima facie reasonable suggestion is to say that a conditional is true if there is a sufficiently strong argument from antecedent to consequent; false if there is an argument connecting antecedent and consequent but the argument is weak at best (or perhaps there is even a strong argument from the antecedent to the negation of the consequent, or to a contrary of the consequent); and neither true nor false if there is no connection at all. A variant option is to say that falsity coincides only with there being a sufficiently strong argument from antecedent to the negation or to a contrary of the consequent, while either absence of an inferential connection, or presence of only a weak one, coincides with indeterminacy. In fine-tuning inferentialism along these lines, it would be good to let experimental data be our guide. But then we should have the relevant data—which at present we do not have.[5] Another call for more experimental work!

The second issue I want to mention concerns another another body of data, to wit, data on how people evaluate the probabilities of conditionals. Over the past decade, evidence has accumulated for the hypothesis that people tend to evaluate the probabilities of conditionals in line with what is generally known as "the Equation," according to which the probability of a conditional should equal the corresponding conditional probability (i.e., the probability of the consequent given the antecedent).[6] It is unclear what inferentialism predicts about the probabilities of conditionals, and accordingly, it is unclear how the designated data bear on inferentialism.

An attempt to align the position with the aforementioned data starts with the observation that little is known about the provenance of conditional probabilities. A number of authors (e.g., Edgington [1995:266 f]) have provided strong reasons for holding that conditional probabilities are primitive, in the sense that unconditional probabilities derive from conditional probabilities rather than the other way around (as is, for instance, suggested by the Kol-

[5]The defective truth table data are inconclusive in this respect, given that they contain no information about inferential connections perceived by the participants in the various experiments.

[6]See Hadjichristidis et al. [2001], Evans, Handley, and Over [2003], Oberauer and Wilhelm [2003], Over and Evans [2003], Evans and Over [2004], Weidenfeld, Oberauer, and Hornig [2005], Evans et al. [2007], Oaksford and Chater [2007], Oberauer, Weidenfeld, and Fischer [2007], Over et al. [2007], Gauffroy and Barrioullet [2009], Douven and Verbrugge [2010], [2013], Pfeifer and Kleiter [2010], Politzer, Over, and Baratgin [2010], Fugard et al. [2011], and Over, Douven, and Verbrugge [2013].

mogorovian ratio definition of conditional probability, which defines the probability of φ given ψ as the probability of the conjunction of φ and ψ divided by the probability of ψ). If we do not derive conditional probabilities from unconditional probabilities, how *do* we arrive at them? Surely they do not just come out of thin air. This is presently an open question. In particular, it is unknown what goes on, psychologically speaking, when people try to determine a given conditional probability.

Many see the so-called Ramsey Test as providing at least the beginning of an answer. According to this test, we determine the probability of φ given ψ by hypothetically updating our stock of beliefs on ψ and then determining how likely φ is as seen from the resulting hypothetical perspective. This suggestion is found in a famous but short footnote in Ramsey [1929/1990] (whence the label "Ramsey Test"). But this can indeed at best be part of the story about the provenance of conditional probabilities. For how exactly are we to update hypothetically our belief state on ψ? By conditionalization, that is, by setting our (hypothetical) new probabilities to our old (actually current) probabilities conditional on ψ? But that presupposes that we already *have* probabilities conditional on ψ. Furthermore, assume that we have hypothetically updated our belief state on ψ (however we do it). Then how exactly do we determine the probability of φ in that hypothetical belief state in order to arrive at our conditional probability for φ given ψ? The Ramsey Test has nothing specific to say about this.

Here is, very tentatively, a different answer to the question of where conditional probabilities come from, an answer that is at least a bit less sketchy than the Ramsey Test—even if it is still very sketchy—and that points toward a way of reconciling the data about the probabilities of conditionals with inferentialism. To determine the probability of φ conditional on ψ, we assess the *strength* of the argument from ψ plus background knowledge to φ and make our conditional probability of the latter given the former (plus background knowledge) proportionate to that; so for instance, a strong argument from ψ to φ will result in a high conditional probability of the latter given the former, a moderately strong argument in a middling conditional probability, and a weak argument in a low conditional probability. Then to get the probability for the corresponding conditional—in this case, "If ψ then φ"—we simply set that equal to our conditional probability for φ given ψ. For example, I think there is a strong argument from

(6) The emission of greenhouse gases will continue to increase.

to

(7) Sea levels will keep rising.

Accordingly, and in line with the foregoing proposal, my conditional probability for (7) given (6) is high. The same is true of my probability for

(8) If the emission of greenhouse gases continues to increase, sea levels will keep rising.

which is precisely what the proposal predicts.

As it stands, however, the foregoing has a worrying consequence. I do not see any connection, whether inferential or other, between Obama's being the president of the United States and the Winter Olympics' being held in Russia. Still, there *is* an inferential connection between either proposition *together with background knowledge* and the other proposition. After all, from our background knowledge, both propositions follow. Thus, our conditional probability for either proposition given the other in conjunction with background knowledge is 1. That is still consistent with to the above proposal. But now consider that, according to inferentialism,

(9) If Obama is the president of the United States, then the Winter Olympics will be held in Russia.

fails to be true (because the consequent follows from the background knowledge alone; see note 3), and the same holds for

(10) If the Winter Olympics will be held in Russia, then Obama is the president of the United States.

Knowing this, it would be wrong for me to set my probability for either conditional to 1. But then we have a conflict with the Equation, given that my conditional probabilities corresponding to (9) and (10) *are* both 1.

This is worrying, but it is not clear whether this should worry the inferentialist or the advocates of the Equation. For at least so far, none of the empirical studies concerned with the Equation has asked participants to evaluate the probabilities of missing-link conditionals like (9) and (10). I would not be hugely surprised if it turned out that people do *not* assign probability 1 to such conditionals. Still more experimental work ahead! (Better to be overburdened than to be bored.)

In summary, while many have the intuition that the truth of a conditional requires the existence of some kind of connection between the conditional's

parts, the question of the nature of that connection has remained hard to answer. That may have been because theorists so far have sought for a unique and specific nature that this connection must have. But just as truth and identity might be second-order functional properties, and might thus be realized by various different more specific, first-order properties, so the sought-for connection between a conditional's parts might be a second-order functional relation. According to the position we focused on in the foregoing—inferentialism—it is. According to this position, the truth of a conditional requires the existence of a "connector," where this connector may be realized by different inferential relations. In this view, the variety of realizer properties is still rather limited; in any case, they are all of an inferential nature. Whether the view is *too* limited thereby remains to be seen, as other important questions about inferentialism remain to be answered. But it is encouraging to see that an old philosophical idea combined with a rather new philosophical idea—that some properties and relations which, on one level of analysis, appear as one may on a lower level be realized by a variety of different properties or relations—has led to a formulation of a position that, at least so far, is backed by both theoretical and empirical considerations.[7]

1. References

Adams, E. W. [1965] "The Logic of Conditionals," *Inquiry* 8:166–197.

Adams, E. W. [1975] *The Logic of Conditionals*, Dordrecht: Reidel.

Bach, K. [1999] "The Myth of Conventional Implicature," *Linguistics and Philosophy* 22:327–366.

Baratgin, J., Over, D. E., and Politzer, G. [2013] "Uncertainty and the de Finetti Tables," *Thinking and Reasoning*, in press.

Barrouillet, P., Gauffroy, C., and Lecas, J.-F. [2008] "Mental Models and the Suppositional Account of Conditionals," *Psychological Review* 115:760–771.

Bennett, J. [2003] *A Philosophical Guide to Conditionals*, Oxford: Oxford University Press.

Burgess, J. P. [2004] Review of Bennett [2003], *Bulletin of Symbolic Logic* 10:565–570.

Decock, L. and Douven, I. [2012] "Putnam's Internal Realism: A Radical Restatement," *Topoi* 31:111–120.

[7] I am grateful to Janneke Huitink for helpful comments on a draft version of this paper.

Douven, I. [2011] "Indicative Conditionals," in L. Horsten and R. Pettigrew (eds.) *A Companion to Philosophical Logic*, London: Continuum Press, pp. 383–405.

Douven, I. [2013] "The Epistemology of Conditionals," *Oxford Studies in Epistemology* 4:3–33.

Douven, I. and Decock, L. [2010] "Identity and Similarity," *Philosophical Studies* 151:59–78.

Douven, I., Elqayam, S., Singmann, H., Over, D. E., and Huitink, J. [2014] "Conditionals and Inferential Connections," manuscript.

Douven, I. and Verbrugge, S. [2010] "The Adams Family," *Cognition* 117:302–318.

Douven, I. and Verbrugge, S. [2013] "The Probabilities of Conditionals Revisited," *Cognitive Science* 37:711–730.

Edgington, D. [1995] "On Conditionals," *Mind* 104:235–329.

Engel, P. [2014] "Is Identity a Functional Property?" in *Themes from Ruth Marcus*, Ontos Verlag, in press.

Evans, J. St. B. T., Handley, S. J., Hadjchristidis, C., Thompson, V., Over, D. E., and Bennett, S. [2007] "On the Basis of Belief in Causal and Diagnostic Conditionals," *Quarterly Journal of Experimental Psychology* 60:635–643.

Evans, J. St. B. T., Handley, S. J., Neilens, H., and Over, D. E. [2006] "The Influence of Cognitive Ability and Instructional Set on Causal Conditional Inference," *Quarterly Journal of Experimental Psychology* 63:892–909.

Evans, J. St. B. T., Handley, S. J., Neilens, H., and Over, D. E. [2007] "Thinking About Conditionals: A Study of Individual Differences," *Memory and Cognition* 35:1759–1771.

Evans, J. St. B. T., Handley, S. J., and Over, D. E. [2003] "Conditionals and Conditional Probability," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29:321–355.

Evans, J. St. B. T. and Over, D. E. [2004] *If*, Oxford: Oxford University Press.

Fugard, A., Pfeifer, N., Mayerhofer, B., and Kleiter, G. [2011] "How People Interpret Conditionals," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37:635–648.

Gauffroy, C. and Barrouillet, P. [2009] "Heuristic and Analytic Processes in Mental Models for Conditionals: An Integrative Developmental Theory," *Developmental Review* 29:249–282.

Gibbard, A. [1981] "Two Recent Theories of Conditionals," in W. L. Harper, R. Stalnaker, and G. Pearce (eds.) *Ifs*, Dordrecht: Reidel, pp. 211–247.

Grice, H. P. [1989] "Logic and Conversation," in his *Studies in the Way of Words*, Cambridge MA: Harvard University Press, pp. 22–40.

Hadjichristidis, C., Stevenson, R. J., Over, D. E., Sloman, S. A., Evans, J. St. B. T., and Feeney, A. [2001] "On the Evaluation of 'if p then q' Conditionals," in J. D. Moore and K. Stenning (eds.) *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, Edinburgh, pp. 381–386.

Kneale, W. and Kneale, M. [1962] *The Development of Logic*, Oxford: Oxford University Press.

Krzyżanowska, K. H., Wenmackers, S., and Douven, I. [2013] "Inferential Conditionals and Evidentiality," *Journal of Logic, Language and Information* 22:315–334.

Krzyżanowska, K. H., Wenmackers, S., and Douven, I. [2014] "Rethinking Gibbard's Riverboat Argument," *Studia Logica*, in press.

Lycan, W. G. [2001] *Real Conditionals*, Oxford: Oxford University Press.

Lynch, M. P. [2009] *Truth as One and Many*, Oxford: Oxford University Press.

Merin, A. [2007] "Unconditionals," available at http://semanticsarchive.net/Archive/WUwZTk5M/unconditionals.pdf.

Oaksford, M. and Chater, N. [2007] *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*, Oxford: Oxford University Press.

Oberauer, K., Weidenfeld, A., and Fischer, K. [2007] "What Makes Us Believe a Conditional? The Roles of Covariation and Causality," *Thinking and Reasoning* 13:340–369.

Oberauer, K. and Wilhelm, O. [2003] "The Meaning(s) of Conditionals: Conditional Probabilities, Mental Models and Personal Utilities," *Journal of Experimental Psychology: Learning, Memory and Cognition* 29:688–693.

Over, D. E., Douven, I., and Verbrugge, S. [2013] "Scope Ambiguities and Conditionals," *Thinking and Reasoning*, in press.

Over, D. E. and Evans, J. St. B. T. [2003] "The Probability of Conditionals: The Psychological Evidence," *Mind and Language* 18:340–358.

Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., and Sloman, S. A. [2007] "The Probability of Causal Conditionals," *Cognitive Psychology* 54:62–97.

Pfeifer, N. and Kleiter, G. D. [2010] "The Conditional in Mental Probability Logic," in M. Oaksford and N. Chater (eds.) *Cognition and Conditionals*, Oxford: Oxford University Press, pp. 153–173.

Politzer, G., Over, D. E., and Baratgin, J. [2010] "Betting on Conditionals," *Thinking and Reasoning* 16:172–197.

Ramsey, F. P. [1929/1990] "General Propositions and Causality," in his *Philosophical Papers*, edited by D. H. Mellor, Cambridge: Cambridge University Press, 1990, pp. 145–163.

Weidenfeld, A., Oberauer, K., and Horning, R. [2005] "Causal and Non-Causal Conditionals: An Integrated Model of Interpretation and Reasoning," *Quarterly Journal of Experimental Psychology* 58:1479–1513.

Wright, C. [1994] *Truth and Objectivity*, Cambridge MA: Harvard University Press.

27

Reference, Truth, and Biological Kinds

MARCEL WEBER

Abstract This paper examines causal theories of reference with respect to how plausible an account they give of non-physical natural kind terms such as 'gene' as well as of the truth of the associated theoretical claims. I first show that reference fixism for 'gene' fails. By this, I mean the claim that the reference of 'gene' was stable over longer historical periods, for example, since the classical period of transmission genetics. Second, I show that the theory of partial reference does not do justice to some widely held realist intuitions about classical genetics. This result is at loggerheads with the explicit goals usually associated with partial theories of reference, which is to defend a realist semantics for scientific terms. Thirdly, I show that, contrary to received wisdom and perhaps contrary to physics and chemistry, neither reference fixism nor partial reference are necessary in order to hold on to scientific realism about biology. I pinpoint the reasons for this in the nature of biological kinds, which do not even remotely resemble natural kinds (i.e., Lockean real essences) as traditionally conceived.

1. Introduction: Reference and conceptual change

There are occasions in the history of science that are of particular interest with respect to the metaphysical question of how concepts relate to the world. I am thinking of such episodes where some newly discovered thing generates controversy as to how exactly it should be classified. A recent example has been widely publicised: the question of whether trans-Neptunian object 2003 UB313 is or is not a planet. In the history of biology, there are many cases like this. Here are two examples. First, at the beginning of the 19th century, naturalists argued as to whether a newly discovered creature from Australia was a mammal or not. A very strange creature indeed, the duck-billed platypus *Ornithorhynchus anatinus* (first named *Platypus paradoxus*!) appeared to have features from mammals and from reptiles and birds. In fact, some British naturalists, on being shipped the first specimens from Australia, thought it was a colonial prank.[1] Here is a second example: At the dawn of molecular biology in the 1940s, scientists discussed whether bacteria and viruses have genes. Both questions have been settled by the scientific community in the meantime: the platypus' status as a mammal is secure, and bacterial and viral genes are all over the scientific journals. By contrast, UB313 didn't make it and took poor Pluto down as well.

Cases like these may be seen as supporting a certain philosophy of language. According to a position known as "meaning finitism," the extension of a term is not determined. This indeterminacy is such that, whenever a new case arises, there is no fact of the matter as to whether it belongs to the concept's extension or not. The inclusion or exclusion of any referent of a concept is always subject to negotiation by the scientific community, meaning finitists argue (Barnes 1982, Bloor 1997, Kusch 2002). Clearly, meaning finitists will see cases like the platypus and the microbial genes as confirming instances for their philosophy of language: They will argue that, prior to the closure of these processes of negotiation, there was no fact of the matter as to whether the platypus belonged to the class *Mammalia*. By the same token, it was not determined whether bacteria contain any entities that are of the same kind as the genes of higher organisms. However, it must be stressed that a mere lack of consensus among a group of speakers alone does not prove that there are no reference-constituting facts, that is, facts that make it so that some thing falls

[1] For a history of platypus biology, see Moyal, A. (2001) *Platypus. The Extraordinary Story of How A Curious Creature Baffled the World*. Washington D.C.: Smithsonian Institution Press. For a detailed analysis of this case from the perspective of the theory of reference, see LaPorte, J. (2003). *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

under a concept. If a group of speakers disagree whether some thing instantiates a certain concept or not, this could mean two things: It could mean that there *is* no fact that makes a certain thing belong to a concept before a relevant group has made a collective decision. But it could also mean that it is merely not *known*, or not known with certainty, if some reference-constituting facts actually obtain or not. According to some philosophers, there are reference-constituting facts associated with a term that may not be accessible or transparent to the relevant linguistic community.

Taking that second line is a challenge. Anyone who wants to argue that there are reference-constituting facts must be able to give a philosophically adequate answer of what determines the reference of scientific terms. And note that my concern is not whether the reference of terms can be determinate under some ideal conditions. It is rather whether, in these historical situations at hand, there were reference-constitutive facts that eluded the scientific community or where it was not known with certainty whether some such facts obtained or not.

I would like to examine whether a certain kind of theory of reference is able to establish the existence of such elusive reference-constituting facts about scientific terms, namely causal theories of reference. The first causal theories of reference have been developed by Saul Kripke (1980) and Hilary Putnam (1973). Such theories claim that the reference of terms may be fixed by the ostension of samples of some natural kind. After an initial "baptism", the term remains rigidly attached to whatever shares a common essence with the original samples. For example, the term "water" is rigidly attached to a certain molecular structure, H_2O, which provides the underlying essence. This essence may be unknown, but the interesting cases are those where this essence is later discovered. According to the original version of the causal theory, such later discoveries of essences leave reference unchanged. I shall refer to this view as "reference fixism".

As is widely recognised today, the original version of the causal theory faces severe difficulties. Probably the most serious one is known as the “qua” problem. This problem arises because a sample may instantiate different kinds. A sample of water may also be viewed as instantiating the kind of liquids or hydrogen compounds, for example. Nothing in the original apparatus of the causal theory can distinguish between them. For this reason, many authors have modified the theory to allow certain content-bearing mental states to be involved in reference fixing (e.g., Nola 1980; Sankey 1994; Psillos 1999; Stanford and Kitcher 2000). Such theories are known as ”causal-descriptive theories of reference”. Because this quite a mouthful, I will refer to them simply as

"causal theories of reference".

One of the goals of this paper is to show that under the assumptions of such a theory of reference, reference fixism about biological kind terms fails. The main example I shall use is the case of genes. After giving some historical background (Section 2), I will show that a refined causal theory of reference fails to establish reference fixism about the term "gene" (Section 3). Further, I would like to show that the reasons for this failure are philosophically interesting; they tell us something about the nature of kinds in biology and perhaps also in other special sciences. I will locate the reasons for the failure of reference fixism in the salient sameness of kind relations that underlie the classification of biological entities (Section 4). In Section 5, discuss the notion of partial reference and the attempt to use it as a basis for a realist semantics.[2] I show that, in the context of biology, partial reference theory has consequences that are opposed to its realist goals. Finally, I will show that the failure of reference fixism and of partial reference is not a problem for realism about biological theories (Section 6).

2. The case of the gene

I would like to use the gene concept as an example, but I believe that some of the results may be of more general relevance. The history of the gene concept is extremely complex. Here are just some stages in its historical development (Carlson 1966; Portin 1993; Waters 1994, 2004; Weber 2005).

[2] Such accounts are usually developed with the aim of countering forms of anti-realism that are based on Laudan's pessimistic meta-induction and/or Kuhnian considerations that involve incommensurability (Laudan, L., 1984: A Confutation of Convergent Realism. In J. Leplin (Ed.), *Scientific Realism* pp. 218-249, Berkeley: University of California Press. Kuhn, T. S., 1970: *The Structure of Scientific Revolutions* 2nd ed., Chicago: The University of Chicago Press). This kind of challenge begins by observing that there are historical predecessors of our contemporary scientific theories that were empirically successful, yet their theoretical vocabulary contains either terms such as "phlogiston" or "ether" that are thought to have no reference. In response to this challenge, realists have tried to show that at least some of the terms of these theories successfully referred (e.g., the terms "dephlogisticated air" or "transversal electromagnetic wave") and that this referential success supported important truths.

Time	Concept	Mendelian behavior	Gene-trait relation	Functional role	Material basis	Structure	Individuation
late 19thc.	pangene	no	?	pangenesis	particulate	open	?
1900–1919	unit-character	yes	one-one	trait-determination	open	open	via trait
1915–1950s	classical	yes	many-many	phenotypic difference maker	chromosomes	subgenes (?)	complementation
1950s	neo-classical	optional	many-many	cistron	DNA	linear	complementation
1960s	molecular	optional	many-many	protein coding	DNA	colinear w/ protein	via gene product
1970s–present	contemporary	optional	many-many	protein/RNA coding	DNA / RNA	intron/exon (optional)	via gene product(s)

I shall try to simplify this story by trying to answer the simple question that I raised at the beginning: Did the term "gene" around 1940 refer to bacterial and viral genes, even though the latter had not yet been discovered?

Let us assume that the reference of the term "gene" was originally fixed with the help of a few experimental systems, in particular the fruit fly *Drosophila*. This was the main model organism used by Thomas Hunt Morgan and his associates to develop the classical theory of the gene in the years 1910-1915 (Morgan, et al. 1915). In their writings, these geneticists introduce the term "gene" by describing certain patterns of inheritance of certain trait differences in the fruit fly. These patterns include the segregation of certain traits according to the Mendelian ratios, and the independent assortment of pairs of traits.

They also show that these patterns of genetic transmission can be explained by assuming the existence of independent factors or genes that are located on the fly chromosomes. Genes that are located on the same chromosomes tend to be transmitted together, a phenomenon that was termed "linkage". But with a certain frequency, this linkage was broken. Morgan and his associates argue that the observed frequencies can be explained by assuming that the genes are arranged linearly on the chromosomes. They are also very careful in pointing out that the relationship of factors and traits was many-many: most genes affect many traits, and most traits are affected by many genes.

It is tempting to suggest that the experimental practices of these early geneticists rigidly attached the term "gene" to the things that were causally responsible for the trait differences, behaved in accordance with these Mendelian patterns and everything else that shares some kind of essence with these things. This is what a causal theory of reference suggests for this case. I will work out this suggestion in more detail in the following part, using a refined version of the causal theory of reference due to Stanford and Kitcher (2000). Then I will show that such an attempt to defend reference fixism about the term "gene" fails.

3. Reference Fixism: Stanford and Kitcher

Let us assume that the reference of the term “gene” was fixed by the following means (this is a slightly modified version of a causal-descriptive theory of reference that has been developed by Stanford and Kitcher 2000):

> a) A range of experimental systems consisting of different strains of fruit flies and a few other organisms showing both instances and counter-instances of Mendelian inheritance

b) A complex conjunctive predicate $\Phi(x)$ composed of predicates $\varphi_1 x \& \varphi_2 x \& \ldots \& \varphi_n x$ such that each instance satisfies $\Phi(x)$ and each counter-instance fails to satisfy $\Phi(x)$

We are not assuming that the constitutive predicates $\varphi_n x$ are purely observational. In other words, inferences are permitted when applying these predicates.

And these might be the relevant φ-properties in our present example:

φ_1 : is arranged linearly on chromosomes

φ_2 : segregates and assorts in accordance with Mendel's laws (three kinds of Mendelian inheritance according to T.H. Morgan 1917: autosomal, sex-linked and due to unusual distribution of chromosomes)

φ_3 : exhibits linkage to other factors located on the same chromosome

φ_4 : crosses over with a frequency roughly proportional to the distance between two factors

φ_5 : complements alleles residing at different loci

φ_6 : mutates spontaneously or under the influence of ionising radiation or certain chemicals

φ_7 : causes heritable phenotypic differences when mutated (difference makers, not total causes!)

The question now is of this apparatus is sufficient to attach the term "gene" to a class of things sharing an essence. What might this essence look like?

Of course, today we have the molecular gene concept according to which genes are DNA sequences that determine the linear structure of a protein or RNA molecule. Could we not view this *coding property* as something like an essence that is shared by all genes, including bacterial and viral genes? In asking this question, it is important to note that causal theorists of reference will not be worried about the fact that Morgan and his associates did not *know* the molecular essence of genes. Causal theories of reference were developed for precisely such cases.

The crucial question is whether the classical gene concept picked out a molecular, relational essence. The question is far from being trivial. Reasons can be given both for affirming or for denying such a thesis for referential continuity. Many of the genes isolated in Morgan's lab were later described at the

molecular level. I have shown that the classical gene concept and the associated operational criteria were actually *used* for isolating molecular genes in *Drosophila* (Weber 2005, Ch. 6). The molecular concept, on the other hand, was worked out mainly by using bacteria and bacteriophage as model organisms.

However, none of this really proves that the reference of the term "gene" as, it was introduced by classical geneticists, extended to bacteria in 1940. For bacteria did not exhibit the patterns of inheritance known from fruit flies and other higher organisms. They have no chromosomes in the classical, cytological sense of the term. They don't exhibit Mendel's laws. Something like phenomena of linkage and crossing-over can be observed, but only under highly contrived experimental conditions. These include, for example, double infections of bacterial cells with two different strains of virus. What is interesting to note is that Seymour Benzer, who was the first to apply the technique of complementation analysis to bacteriophages by using this technique, had strong reservations about the term "gene" (Benzer 1955).

The only property that bacteria showed from the beginning was random mutation. This was shown in a classic study by Max Delbrück and Salvador Luria that was published in 1943 (Luria and Delbrück 1943). In the conclusion section of their paper, Delbrück and Luria wrote: "Naming such hereditary changes 'mutations' of course does not imply a detailed similarity with any of the classes of mutations that have been analyzed in terms of genes for higher organisms. The similarity may be merely a formal one." Clearly, they were reluctant to draw any close parallel between the processes they studied in bacteria and those studied by *Drosophila* geneticists. Of course, this will not worry causal theorists of reference because, in their view, reference-constituting facts may obtain irrespectively of what scientists actually believe.

However, what causal theorists of reference must show is that the scientists' mental states together with the experimental systems originally used when some term was introduced uniquely pick out some essence, for example to DNA sequences that have the coding property. What made it so that the classical term "gene" referred exactly to the set of DNA sequences that share the coding property in their cellular context?

One suggestion might be that the coding property is the function that explains all of the properties traditionally associated with genes, in other words, the φ-properties according to our present account. I mean "function" in a minimal causal role sense, that is, not in the sense of proper function. We could further modify Stanford's and Kitcher's theory of reference. They suggest that natural kind terms refer to "the set of those things having the inner constitution that is a common constituent in the total causes of the presence of each

of the φ-properties in each of the samples." This is not applicable to our case, because genes are not structural kinds. We need to substitute function for structure.

So is there some function that is a common constituent in the total cause of each of the φ-properties? This does not seem right. Some of the phenomena studied by classical geneticists, in particular the Mendelian regularities, are explained simply by the way in which the chromosome align and separate in the formation of germ cells, not by the coding property.

Note also how important the Mendelian behavior was for the initial referential success of classical geneticists. To drop the Mendelian behavior from the list of properties involved in reference fixing means also to drop the chromosomal location of genes. But this allows the *qua*-problem to run amok. Because then it is not at all clear what functional properties the geneticsts were ostending when they introduced the term "gene" into discourse. The reference of "gene" then might include all sorts of things that are involved in heredity, including cytoplasmic factors. If classical geneticists succeeded in referring to anything, it was something that is located on a chromosome and, therefore, exhibits the Mendelian patterns.

Stanford and Kitcher suggest that "a principal motivation for causal theories lies in the possibility of discovering that some members of a natural kind lack properties originally used in picking out that kind". They suggest that this was the case in the example they have studied, which is the chemical term "acid". So Stanford and Kitcher, it seems, would allow that some referents of a kind term are later shown to lack some of the properties that once were crucial for referential success. This would allow bacteria to have genes, even though they the lack Mendelian inheritance. However, I will show now that their account cannot be modified in a way that would allow us to say that the classical gene concept picked out the molecular essence of genes.

4. Sameness of Kind: Why Reference Fixism Fails

In section 3, I raised some scepticism concerning the idea that the classical term “gene” may have referred to some molecular essence prior to the advent of molecular biology. Now, it is time to provide some metaphysical grounds for this scepticism.

Genes are no kind like those that have been discussed in physics and chemistry. Here are some differences, most of which have been consistently ignored in discussions of reference and biological kinds:

(1) Genes are a *relational* kind. To be a gene is not an intrinsic property of some chemical substance. Some DNA sequences are only genes because there exist cellular contexts that contain specific biochemical machinery of gene expression. While some of the machinery can recognize DNA sequences from other species, most genes are only properly expressed by cells derived from the same species (unless the sequences are tampered with by genetic engineers). Thus, while to be a H_2O molecule is an intrinsic property that a thing can possess independently of anything else, to be a gene is not an intrinsic property.[3]

(2) Genes are a *functional* kind, in the sense that they are individuated by their causal role in a system. This is partly responsible the relational character of genes mentioned above. In the molecular sense, genes are also *structural* kinds, because only things made of nucleic acid are called "genes" today. Thus, genes are a mixed-functional kind (Waters 2000).

(3) Genes are a *variable* kind. All water molecules are the same. By contrast, genes vary enormously both within and between species.

(4) Genes are a *generic* kind. Genes come in billions of subkinds such as "the human PAX6 gene" or "the *Drosophila melanogaster white* gene", etc. The generic kind of gene and these subkinds are related in the same way as the kind "species" is related to the kind "*Homo sapiens*". Every species *taxon* (e.g., *H. sapiens*) is an instance of the species *category* (to use Ernst Mayr's terms). By the same token, every specific gene (e.g., the human PAX6 gene) is an instance of the generic kind "gene". I shall use the terms *specific gene kinds* and *generic gene kind* to distinguish these.

(5) Genes are *sortal* kinds. You can count genes, and a statement of the form "there are less than 50'000 human genes" (meaning specific gene kinds) or "this plasmid contains three genes" (meaning tokens of arbitrary specific gene kinds) are complete with requiring extra sortal terms. By contrast, statements

[3] Neumann-Held, E. M. (1999). The Gene is Dead - Long Live the Gene! Conceptualizing Genes the Constructionist Way. In P. Koslowski (Ed.), *Sociobiology and Bioeconomics. The Theory of Evolution in Biological and Economic Thinking* pp. 105-137). Berlin: Springer. has argued that genes ought to be conceptualized as *containing* all the biochemical machinery necessary to express them. I don't see the need for such a radical departure from molecular biological uses of the term "gene". That genes are *relational* with respect to this machinery does not mean that they *contain* it (in a mereological sense).

of the form "there are about 1'000'000'000 H_2O in this sample" is incomplete without addition of general sortal term such as "molecule". This sortal term, by the way, is unfit for gene talk, as one molecule of DNA or RNA may contain an arbitrary number of genes.

Another way of expressing this characteristic is by pointing out that the term "gene" is a *count noun* (see Rosenberg 2006, 114). Most of the natural kind terms that have been discussed in the philosophy of science are not count nouns, but *mass nouns*. Examples include "oxygen" or "water". Even "acid" is a mass noun. You can count acid-*types* (sulphuric acid, acetic acid), but not acid-*tokens*, unless you introduce another sortal expression such as "molecule". Examples of count nouns in the physical sciences are "atom" or "electron" but – curiously – these are not the terms that have been discussed the most in debates over reference and concepts in science.

I would like to claim that some of these characteristics are responsible for the difficulties of applying a causal theory of reference to the kind "gene". For such theories to work, it is instrumental that there is a Lockean real essence (i.e., an inner constitution or common structure) that furnishes the salient sameness of kind relation. This real essence had better be nomologically linked to the properties used to identify instances of the kind (Locke's "nominal essence"). The relational nature of genes is not compatible with there being such nomological connections. Gold atoms and the laws of physics (should those be in some sense independent of the intrinsic properties of gold, which some metaphysicians doubt, see Ellis 2001) make it so that lumps of gold exhibit the same properties in many different contexts in which they can exist. This is not so in the case of the gene. DNA or RNA as a chemical compound may satisfy this requirement, but not any piece of DNA or RNA contains genes. The gene-making relations are *context-dependent*. A piece of human DNA will not be biologically active in most cellular contexts, even if in its original context (a human cell) it contains a fully functional gene. Therefore, with respect to their biological (as opposed to purely chemical) properties, genes lack the kind of context-independent nomological relations to other properties.

Does this matter at all? This will depend on whether there is some sort of *unique causal role* that all and only genes share and that, perhaps, could constitute their relational essence. According to molecular biology, there appears to be such a role: The causal determination of the linear sequence of

either RNA or protein molecules (in the appropriate cellular environment).[4] The question is if this causal role is specific enough to delimit all and only genes. There are reasons for doubt. First of all, the notion of "causally determining the linear sequence of a biomolecule" is in need of explication. Probably the best explication for this causal notion is this: The salient sense of causal determination here is to be explicated in terms of *causally specific actual difference-making causes*. Ken Waters (forthcoming) has recently used James Woodward's manipulationist theory of causation in order to explicate this concept. This account starts by differentiating between *potential* and *actual* difference-making causes in a population of entities (e.g., the population of proteins in a cell). Actual difference-making causes are those that actually vary in the population and that account for the variation of the dependent variable. Potential difference-making causes are capable of this, but they don't *actually* vary in the population. Where the actual difference-making cause fully accounts for the variation in the dependent variable, Waters speaks of *the* actual-difference-making cause. If the independent variable accounts for the variation in the dependent variable only partially, Waters refers to the former as *a* actual difference-making cause (whether a given variable is "independent" or "dependent" is to be analyzed in accordance with Woodward's theory of causation. Basically, independent variables (causes) are those that can be manipulated such as to change the value of another variable (effects) in a way that does not alter the value of any other variables that could do the same.

According to Waters, this apparatus can be used to specify a unique role for certain nucleic acids in determining the linear structure of other nucleic acids or proteins, for example, prokaryotic genes (where there is no post-transcriptional modification). An additional causal concept is needed to single out a unique role for eukaryotic genes: the concept of causal specificity. Waters borrows this notion from Lewis (2000). Briefly, specific causes are causes where a multiplicity of different states of an independent variable are causally linked to a comparative multiplicity of states of the dependent variable. Using

[4] As Ken Waters has argued, in molecular biology the use of the term "gene" is context-sensitive: depending on the stage of gene expression that is being talked about, a gene may include or exclude certain DNA sequences. For example, in a context where biologists talk about primary transcript, they will mean the term "gene" in a sense that includes the introns (non-coding intervening sequences that are spliced out after transcription). By contrast, in a context where they speak about mRNA or finished proteins, the gene will exclude the introns (Waters, C. K. 1994. Genes Made Molecular. *Philosophy of Science, 61,* 163-185.). This may be an extra complication for a causal theory of reference, but it seems to me that it fades in comparison to those that I discuss in the text.

this notion, Waters argues that eukaryotic genes are the only causally specific actual difference-making causes in RNA- and protein synthesis.

I have argued elsewhere that the notion of causal specificity admits of *degrees* (Weber 2006). Causal specificity may be viewed as special kind of *invariance* in the sense of Woodward (2003), namely a relationship such that a change in the independent variable (e.g., a DNA sequence) would bring about a change in a dependent variable (e.g. protein sequence) in a way as it is specified in the relationship. Causally specific relationships are such that they relate *discrete* variables. Now, depending on how many different values these variables can take, the relationship is more or less causally specific.

Let us now analyze the causal influence of eukaryotic genes on proteins. DNA is an actual-difference making cause, but so are certain agents that are responsible for alternative splicing (the production of different polypeptides from a single RNA molecule by cutting and joining the exons or coding sequences is different ways). So far, these factors are causally on a par (Oyama (2000), Sterelny and Griffiths (1999)). However, DNA is more causally specific a variable than the splice agents, because it could take a much larger number of different values (= nucleotide sequences). Thus, we may define the causal role of genes as that of being the most highly specific actual difference-making causes in the synthesis of RNA and protein in a cell.

Now add to this the properties from the nominal essence of genes according to the classical theory, i.e., chromosomal location, complementation, Mendelian inheritance, mutation, recombination. A causal theorist of reference might suggest that the term "genes", as it was used in the classical period" denoted exactly those parts of the *Drosophila* chromosomes that had these φ-properties *in the fruit flies* and any other thing that shares the causal role of being the most highly specific actual difference-making cause of the linear structure of RNA and protein in the cell (but that need not have any of the φ-properties).

I think the problem with this suggestion is obvious: This account of the reference of "gene" attributes to Morgan and his associates mental states that they did not have. They may have had mental states bearing contents such as "difference-making cause", perhaps even "highly specific actual difference-making cause" (perhaps implicitly so). But they did not have thoughts containing the idea that genes are the most highly specific difference-making cause *of the linear structure* of protein and RNA. It was not yet known at that time that genes play this biochemical role. But it is necessary to *spell out* this role in order to secure reference to the kind of things recognized as genes by contemporary biology. If that is left out, all we have is a bunch of fly genes

and everything else that has the same causal role. But these genes play many causal roles, so the reference of "gene" would include way too many things.

It seems to me that the general problem is this: There might be no other way to pick out a function short of actually *specifying* the function. I can point to a particular space-time region, say, one containing liquid water and say "I am talking about that stuff, and everything else that has the same structure" and succeed in referring. To be precise, we can succeed *provided* that we can solve the *qua* problem by specifying some appropriate φ-properties such as boiling temperature to exclude that I am talking about the natural kind of liquids, for example.

But if I want to fix the reference of the term "heart", I can't just point to my chest, saying "I am talking about that thumping thing in there, and everything else that has the same function." This would pick out far too many things. For example, this might pick out all things that make thumping noises or that produce heat and carbon dioxide (note that I mean "function" in a minimal causal role sense). In order to succeed in referring, I need to specify *what* function I am talking about, for instance, the blood-pumping function. Therefore, it is not possible to refer to such an essence without already *knowing* it. But this is exactly what causal theories of reference would require.

As LaPorte (2003) points out, we should not judge a theory of reference on the basis of whether or not it makes reference determinate. A theory of reference shouldn't see referential determinacy where there is none. However, the whole point of bringing causal theories of reference to the philosophy of science so far has been to establish referential continuity in the face of theoretical and conceptual change in science. The upshot of my analysis, so far, is that the case of the gene lacks such continuity, and there are *in principle* reasons for this, reasons that have to do with the nature biological kinds.

I now turn to examining whether the case of the gene exhibits *partial reference*. This is a form of referential indeterminacy, but presumably one that does not beget radical conceptual change of the kind that spells doom for scientific realism.

5. Partial reference and truth

The idea of partial reference was introduced by Field (1973). Using the transition from Newtonian to relativistic mechanics as his main example, Field argued that there is no fact of the matter as to what the term "mass" referred to prior to Einstein. It did not refer to proper mass, nor did it refer to relativistic

mass (which are taken as the real properties). Reference was indeterminate, and the Newtonian concept of mass was lacking in discriminatory capacity with respect to this distinction.[5] Nonetheless, Field suggested that there is a relation of "partial denotation" between the term "mass" as it was used before Einstein and the real properties relativistic mass and proper mass. This means that the term did not refer to either property; rather it *partially* referred to *both*. According to Field, a similar relation obtains between the classical term "gene" and the units of recombination, of function (Benzer's cistron), and of mutations. Before the advent of molecular biology, the term "gene" lacked discriminatory power to distinguish these different units.[6] Thus, the term "gene" partially denoted all of them.

Stanford and Kitcher[7] also end up endorsing partial reference towards the end of the paper. In their main example, which is the chemical term "acid", chemists abandoned some properties as being essential for acids that were once thought to be essential. This would not be a problem if there were only one salient natural kind in the relevant domain. In the latter case, it would be possible that the properties used to identify acids (i.e., its nominal essence) could change, while the term "acid" would still refer to the same real essence. But there are several natural kinds that once were candidates for the reference of the term "acid". Hence, reference of the term was partial.

[5] T.S. Kuhn, famously, argued that these concepts are incommensurable, meaning that there is no way of expressing one concept solely in the vocabulary of the other theory. According to Carrier, M. (2001). Changing Laws and Shifting Concepts: On the Nature and Impact of Incommensurability. In P. Hoyningen-Huene, & H. Sankey (Eds.), *Incommensurability and Related Matters* pp. 65-90). Dordrecht: Kluwer. Kuhnian incommensurability means that it is impossible to preserve both the conditions of application and the standing inferential relations in an attempt to translate statements containing concepts into the language of a theory that contains concepts that are incommensurable tot he first. Field's argument can be seen as an attempt to salvage a weak form of realism in the face of the Kuhnian challenge.

[6] It is often said that classical geneticists such as the school of T.H. Morgan thought that these units coincide. This is historically incorrect, see Weber, M. (1998). Representing Genes: Classical Mapping Techniques and the Growth of Genetical Knowledge. *Studies in History and Philosophy of Biological and Biomedical Sciences, 29*, 295-315.

[7] A few years earlier, Kitcher had developed a different account of reference, the theory of reference potential (Kitcher, P. 1978. Theories, Theorists and Theoretical Change. *Philosophical Review, 87*, 519-547.; Kitcher, P. 1982. Genes. *British Journal For The Philosophy Of Science, 33*, 337-359.; Kitcher, P. 1993. *The Advancement of Science. Science Without Legend, Objectivity Without Illusions*. Oxford: Oxford University Press.). On this account, different tokens of a term may refer differently, depending on the intentions of speaker who produced the token. In the 2000 paper, Kitcher suggests that this earlier account is similar to partial reference. However, Christina McLeish (McLeish, C. 2005 Scientific Realism Bit by Bit: Part I. Kitcher on Reference. *Studies in History and Philosophy of Science, 36*, 668-686) argues that the theory of reference potential ultimately fails whereas a modified version of partial reference is defensible.

What is the point of introducing the notion of partial reference? This becomes evident when we ask what the notion of reference was once introduced for: truth. A statement of the form *Fa* is true exactly if *a* belongs to *F*'s extension. Reference is the relation between a predicate's extension and its term, and true statements are such that they predicate a predicate of a member in its extension.

Partial reference is only an interesting relation to the extent in which it can support truths. The whole point of saying that Newton's term "mass" partially referred (as opposed to complete failure of reference as suggested by Kuhn and Feyerabend) is to enable Newton and other pre-Einsteinian physicists to have said at least some true things about the world, even though their theory on the whole was false. Field suggested the following way of allowing for truth with partial reference. Assume that a scientific term such as "mass" is associated with different *structures* that map this term to different referents. One such structure may map the term "mass" to relativistic mass, while another may map it to rest mass. Any statement may now be true or false with respect to a given structure. For example, with respect to a structure that maps "mass" to relativistic mass, the statement "momentum equals velocity times mass" is true (by the lights of relativity theory), while the same statement is false with respect to a structure that maps "mass" to proper mass. So long as this is the case, i.e., when different structures give rise to different truth-values to statements containing partially referring terms, we can't say that the statement is true. Its truth value is indeterminate. However, there is the logical possibility that *all* the structures of such a statement return the value "true". An example would be "in a given frame of reference, the mass of the Earth is less than the mass of the sun". No matter how "mass" is interpreted in this sentence, it comes out true (again, by the lights of relativistic mechanics). In such cases, Field allows a sentence to be true even if contains partially referring terms.

McLeish (2006) has argued that Field's account is too restrictive on truth. It will recognize precious little truths to have been spoken in the history of science. Furthermore, Field's account of truth under partial reference is in conflict with some strong intuitions. McLeish therefore suggests the following amendment of Field's account. First, any partially referring term is not only associated with a set of structures that map the term to some set of referents. It also contains a structure that maps the term to the *empty* set. Thus, a statement like "dephlogisticated air does not exist" is true under at least one structure if that term refers partially. This is in line with our intuition that, in a sense, there is no such thing that fits the description that Priestley et al. gave

of dephlogisticated air. But at the same time there is a way of interpreting some statements made by phlogiston chemists according to which "dehphlogisticated air" referred to oxygen, such that "dephlogisticated air supports respiration" is true. Of course, there is no single interpretation that makes the absurd sentence "dephlogisticated air does not exist and supports respiration" true. This is how it should be.

A second modification introduced by McLeish is to say that sentences containing partially referring terms are true if there is at least *one* structure that makes the statement true. Thus, a statement such as "dephlogisticated air supports the respiration of mice" may be true, namely if there is a structure that maps "dephlogisticated air" to oxygen and "oxygen supports the respiration of mice" is true. In contrast to Field's original account, which is conjunctive, McLeish's account is *disjunctive*. This makes it much more permissive with respect to truth.

McLeish's account has several attractive features. First, it does not make reference of a term used in a statement made in the past a matter of whether that statement is true (which would put the cart before the horse. Successful reference begets truth, not vice versa). Second, it does not privilege any of the descriptions of theoretical entities or magnitudes that scientists used in the past.[8] Thirdly, the account does not need to appeal to *our* intuitions as to whether some past tokening of a term referred. Thus, it avoids a certain kind of whiggism.

I want to leave open question as to whether McLeish's theory gives a correct account of the reference of terms from the physical sciences, such as "mass" or "dephlogisticated air". Of course, there cannot be much hope that this account has no difficulties of its own. McLeish's account is in danger of making reference a vacuous relation. To avoid vacuity, it must be able to show how *reference failure* is possible. We can't have *any* term from the history of science partially refer, e.g., things like Darwin's "gemmules," just because it may be associated with a structure from some class that contains one good structure. I will not delve on this issue here. What I would like to do instead is to show that the aim of allowing truths to be spoken in the past can be reached without partial reference, at least in biology.

[8] By contrast, Kitcher's (Kitcher, P. 1993. *The Advancement of Science. Science Without Legend, Objectivity Without Illusions*. Oxford: Oxford University Press.) account privileges certain descriptions contained in what he calls the reference potential as ensuring referential success.

6. Classification and general truths in biology

When Priestley spoke about "dephlogisticated air", this term may have partially referred to oxygen. Oxygen is a traditional natural kind in that all samples of oxygen share an essential property that all and only the kind members instantiate (Locke's "real essence", given by atomic number according to contemporary chemistry). If McLeish is right, then some of Priestley's statements may have been true, given that there is a partial structure associated with this term that maps "dephlogisticated air" to oxygen. If some general statement endorsed by Priestley was true, e.g., "dephlogistocated air sustains respiration of mice", then it was true of all the members of the natural kind that was partially denoted by "dephlogisticated air."

Now contrast this example with some claim made by a classical geneticist, for example, "genes cross over with a frequency that is roughly proportional to the distance of their separation on the chromosome". Is there a partial structure in Field's and McLeish's sense that maps Morgan's use of the term "gene" to a natural kind? Perhaps there is, provided that this partial structure excludes everything that fails to exhibit this classical genetic regularity. But note that we can just as well say that there is a *subkind* of what is today recognized as genes that is *fully* (as opposed to partially) denoted by Morgan's term gene, namely, all eukaryotic genes that reside on the same chromosome of a diploid, sexually reproducing organism. This subkind is *variable*; it contains different genes from the same species and genes from different species. There is nothing wrong with some use of the term "gene" refer to a subkind of what is today recognized as the class of genes. If we compare this to the oxygen case, we notice that this is not a life option there. You can't refer to a subkind of the natural kind "oxygen", because there aren't any.[9]

We are now ready to consider the problem mentioned in the introduction, to wit, if "gene" referred to bacterial genes before the advent of bacterial genetics. If it did so refer, then it can only have referred partially. For to say that it fully referred to bacterial genes requires that we privilege some description of genes as the dominant one. This can only be done by the lights of molecular biology, which begs the question, see McLeish. However, partial reference is in danger of being a vacuous relation. Is there any way out of this dilemma?

I think there is. We can simply say that Morgan et al. only referred to some *subkinds* of the molecular kind of genes, namely *Drosophila* genes and

[9] There may be different isotopes of oxygen, but these do not differ chemically. Genes, by contrast, come in different subtypes that differ biologically.

perhaps the genes of some sufficiently similar organisms. Thus, we should read general sentences from classical genetics as ranging only over subkinds that do *not* include things such as bacterial genes. If we attribute to Morgan's term "gene" the full (not partial) reference of all molecular genes, this makes most of his general beliefs plainly false. (Bacterial genes show very few of the characteristics that Morgan et al. discovered in *Drosophila*.) This violates the intuition that his group of researchers discovered important truths about inheritance in sexually reproducing organisms. The flight to partial reference is cumbersome, for the reasons indicated. But we don't need partial reference: We can say that Morgan's sentences were not referring to bacterial genes at all. Instead, these sentences were only about the model organisms used back then plus, perhaps, a few others.

By the way, many of the specific genes that classical geneticists talked about were later re-described at the molecular level (Weber 2005, Ch. 7). Thus, there is no difficulty in saying that when Morgan talked about the *Drosophila white* gene, he referred to the same class of DNA sequences as a modern biologist (or the *Drosophila* genomic database known as "Flybase"). Many of the *subkinds* of the generic kind "gene" are quite stable throughout the history of genetics. However, the *generic* term "gene" has not been stable, as many authors have suggested (Kitcher (1982); Burian (1985); Waters (1994); Burian, et al. (1996)). The reference of this term has been "floating" incessantly as new mutants were discovered, as new model organisms and new experimental systems were developed (Weber 2005, Ch. 7).

Biology does not aspire to the kind of generality known from physics or chemistry. General claims in the latter disciplines range over the whole universe. Oxygen atoms, electrons or mass have the same properties and enter into the same nomological relations no matter where they are found. Theories that describe the interactions of fields and particles are universal. Biological theories are much more local. No-one expects there to be a universal genetics. The genetic code, which is found in most organisms on Earth, is about as universal as it gets in biology. This is a far cry from the generality of physical and chemical theories.

As a result, truth comes much easier in biology, unless biologists overgeneralize. Of course, they have been known of over-generalizing. But there is no indication to think that the theory of the gene, as it was proposed by Morgan et al., was supposed to cover all life on Earth, including bacteria and archae. There is certainly no indication in the works of these authors that would suggest that they thought their theories would have this kind of scope. Therefore, to ascribe to their term "gene" such a wide reference as to include

bacteria is both uncharitable and unnecessary. What is more, this is uncharitable and unnecessary *before* we even begin to consider the further difficulties that this will incur, especially those of partial reference.

In comparison to physics and chemistry, biological theories are only of restricted scope. There may be generalizations that are true of all genes (in the molecular sense), but there are also generalizations that are true of some subclasses of genes. (By contrast, there are no physical theories that are only true of some samples of oxygen, or some instances of mass). The theories of classical genetics should not be interpreted as making claims about all kingdoms of life; this takes the theory further than its own fathers would have been willing to defend. For this reason, it is best interpreted as having established *full* reference (as opposed to partial), but not to the *full set of things* that are recognized as genes today. Reference was only to some subclasses of the kind.

At this point, it may be asked if partial reference does not make a similar claim: Does it not also say that a term may partially refer to different kinds which often stand in some hierarchy of kinds? For example, according to partial reference theorists, "dephlogisticated air" also partially referred to gases, which contains oxygen as a subkind. Why should we not say that the classical term "gene" partially referred to some subclass of genes (e.g., those studied by Morgan & Co.), but it also had the full set of molecular genes as a partial referent? As long as partial reference is construed along the lines of McLeish's disjunctive account, this still allows some sentences produced by classical geneticists to be true.

The difference becomes clear if we ask to what extent the different accounts assign the same truth-value to different sentences. Take a sentence such as "all genes are located in the cell nucleus". If "gene" is read in the molecular sense, this sentence is false (bacteria don't have a nucleus, and in eukaryotes there are also mitochondrial and chloroplast genes). On my analysis, the sentence is true if said or thought before the advent of molecular biology. Because I maintain that the reference of "gene" did not reach very far beyond the organisms that were experimentally accessible back then. But on McLeish's account, this sentence may also be regarded as true, provided that there is a structure that maps "gene" to just the nuclear genes of higher organisms (even if there is also a structure that maps "gene" to the set of molecular genes that makes the sentence come out false). Thus, in this case, the two accounts assign the same truth-value. So far so good.

But now comes the rub: There are also sentences such as "all genes segregate in accordance with Mendel's laws of segregation and independent assortment". This sentence was known to be false as early as 1916. Many genes

don't obey Mendel's laws, in fact, the whole history of early 20th century genetics may be described as the discovery of a series of anomalies to these laws (Darden 1991). One of the first anomalies was sex-linked inheritance, another was linkage. I would say that while, originally, the term "gene" only referred to things that obey Mendel's two classical laws, the reference of the term was expanded to accommodate new cases as genetics was developed (Weber 2005, Ch. 7).

Here, McLeish's account exhibits its difference, and also its difficulties. I see no reason why the partial reference theorist should not say that things that obey Mendel's laws of segregation and independent assortment belong to a Field/McLeish-style structure. It's as good as the other structures that we have examined so far. But this has the undesirable consequence that the sentence "all genes obey Mendel's laws of segregation and independent assortment" comes out true, therefore attributing to Morgan & Co. *false beliefs that they did not entertain*.

Could McLeish's account not be saved from this difficulty by saying that "things that obey Mendel's laws of segregation and independent assortment" was not among the descriptions that Morgan et al. used to refer to genes? In fact, statements can be found in their texts that explicitly exclude this description as reference-relevant.

However, this move is not open to the partial reference theorist. For the partial reference theory *forbids* us to privilege some descriptions in determining reference. If descriptions that widen reference are parts of a Field/McLeish structure, then so are descriptions that narrow reference. But as soon as this is accepted, the damage is done: This makes statements true that are clearly false by *any* lights, be it our best contemporary theories or some historical predecessor.

Why does this problem not arise in the more traditional cases such as oxygen? It seems to me that, in the latter cases, there is a *smallest causally homogeneous kind* the (partial) denotation of which by some scientific vocabulary is responsible for the truth of certain sentences. We are there in the tidy world of physics and chemistry, which is neatly divided into causally homogeneous kinds of truly cosmic extensions. This is not so in the messy world of biology. Here, causal homogeneity is a matter of degrees, and a matter of relations. Some class of entities may be causally homogeneous in relation to some specific mechanism (i.e., the gene expression machinery of a bacterial species) but causally heterogeneous in relation to another mechanism. Causal homogeneity is a matter of context in biology. Hence, there is no smallest causally homogeneous kind that the theory of partial reference needs in order to avoid to

make far too many statements true. Even if it works for physical and chemical kinds (which I doubt), it cannot do justice to the nature of biological kinds.

It is time to take the special character of biological kinds into account when speaking about reference and truth in biology. I suggest that taxonomies of kinds in biology should be viewed as *open classification systems*, much like biological systematics itself. In contrast to classifications systems such as the period table or the standard model in particle physics, there is no limit to the number of kinds that such a system could accommodate. It is always possible to introduce new taxa, to lump or split existing taxa, or to enlarge or contract existing taxa. Such classificatory choices will be informed by the theoretical goals that the classificatory system is supposed to serve (and perhaps practical goals and interests as well). The species category, for example, can accommodate an unlimited number of species. It had better be able to so, for new species arise by evolution all the time, while existing species go extinct. When a new species arises, this does not correspond to the filling of a pre-existing slot (unlike when an atom of some chemical element forms for the first time). By the same token, an extinct species does not leave an empty slot behind.

It was a mistake to model the reference of biological terms on the model of oxygen or mass, as Field (1973) or Kitcher (1982) have done. Biological systematics is a much better model. The term "gene" is more similar to the term "species" than it is to "electron" or "acid". It is generic term that comes in many subtypes. "The human PAX6 gene" is related to "gene" like "Homo sapiens" is related to "species". As in the case of species, new genes arise all the time by evolution. When that happens, there is no filling of a pre-existing slot. Even though not infinite in the mathematical sense, the number of possible genes is not limited in any relevant way.

Biology differs enormously from physics with respect to the generality of its theoretical claims—this is hardly news (Beatty 1995; Waters 1998; Weber 1999; Mitchell 2000). But what has not been sufficiently appreciated are the implications of this insight for the theory of reference. Today, in the age of genomics, generalizations such as those of classical genetics (Waters 2004) generalize over *subkinds* of all the things that are classified as genes. By contrast, in the era of classical genetics, these generalizations ranged over the whole extension of the term "gene". This makes for a substantial reference shift. At the same time, this does justice to the intuition that Morgan and co-workers discovered important truths. What is more, none of the other accounts of reference that have been proffered in the history and philosophy of science do proper justice to this intuition. The view of reference fixism attributes to classical geneticists many false beliefs, because it has them make general claims

about genes that differ radically from the genes they had experimental access to (e.g., bacterial genes). There is no historical evidence that these scientists actually held such beliefs. The theory of partial reference, as I have shown, makes statements true that classical geneticists (correctly) thought to be false. Thus, a view of "floating reference" (Weber 2005, Ch. 7) does the best job in attributing true beliefs, and not too many false ones, to classical genetics.

Of course, we should not judge a theory of reference solely on the basis of what kinds of statements it makes true. On the other hand, intuitions about the truth of historical theories has been a major motivation to develop such theories in the first place. Cleary, the alternative theories of reference that I have discussed have problems other than what kinds of truths they support. We can now add to these problems the fact that, with regard to biological kinds, these theories are not necessary in order to hold on to the view that the historical predecessors were tracking important truths.

7. Conclusions

I have examined various theories of reference and conceptual change with respect to what they say about biological kinds, in particular the case of the gene. I have shown that genes are unlike any of the kinds that have been discussed as paradigm cases of natural kinds, such as "oxygen" or "acid". The kind "gene" is relational, functional (or mixed-functional), variable, generic, and sortal. These properties, as I have shown, are toxic for reference fixism. There may be causal[10] as well as descriptive elements involved when experimental biologists attached the term "gene" to some class of unknown factors, as Stanford and Kitcher and others have suggested, however, this causal-descriptive apparatus was never sufficient to pick out anything remotely resembling Lockean real essence, i.e., a molecular constitution or something of this sort. Reference of the term "gene" was *floating*; it changed with every new major model organism and investigate technique deployed. So reference fixism fails.

The theory of partial reference runs into the difficulty that it makes historical statements come out true that were known to be false by the relevant

[10] Ultimately, a causal element in reference-fixing will have to be involved to fence off meaning finitism. Note that my rejection of reference fixism does not commit me to meaning finitism, at least not in its full-blown form. There are many new instances of scientific concepts that are clear-cut and do not require a community choice (like in Kuhnian normal science, perhaps). But there are also new instances that require revision of the existing conceptual taxonomy (like in scientific revolutions, but not necessarily as radical).

historical actors (on McLeish's disjunctive account). I have located this difficulty in the fact that, in biology, there is usually no smallest or most basic causally homogenous kind that could be responsible for a theory's success in speaking truths. Thus, nothing stops such a semantic theory from assigning positive truth values to a motley of statements that may be true about some subkinds of a general kinds. So partial reference is counter-intuitive in biology on top of the other philosophical difficulties in faces.

I hope to have shown that we don't need reference fixism or partial reference to account for the intuition that an area such as classical genetics discovered important truths. General claims made back then generalized only over parts of the domain of molecular genetics, that is, over subkinds of the contemporary gene concept. Some of these general claims ranged over the full extension of the term "gene" as it was used then. This kind of conceptual change, which I have termed floating reference, is different from partial reference in that there was no ambiguity in reference, and it is different from reference fixism in that there were substantial reference shifts associated with new developments in experimental techniques and with new model organisms. Floating reference provides a more adequate truth-conditional, realist semantics for biological science, while something like partial reference may be required for defending a realist semantics in physical science.

Finally, I have suggested that biological kinds are typically part of open classification systems that resemble biological taxonomy itself. Such systems admit lumping and splitting, as new specimens are discovered and new investigative techniques are developed. Nature's biological joints are always in motion, and so is the language of those who try to carve them.

8. References

Barnes, B. (1982). On the Extensions of Concepts and the Growth of Knowledge. *Sociological Review, 30,* 23-44.

Beatty, J. (1995). The Evolutionary Contingency Thesis. In G. Wolters, J. G. Lennox, & P. McLaughlin (Eds.), *Concepts, Theories, and Rationality in the Biological Sciences. The Second Pittsburgh-Konstanz Colloquium in the Philosophy of Science* pp. 45-81). Konstanz/Pittsburgh: Universitätsverlag Konstanz/University of Pittsburgh Press.

Benzer, S. (1955). Fine Structure of a Genetic Region in Bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America, 41,* 344-354.

Bloor, D. (1997). *Wittgenstein, Rules and Institutions*. London: Routledge.

Burian, R. M. (1985). On Conceptual Change in Biology: The Case of the Gene. In D. Depew, & B. Weber (Eds.), *Evolution at a Crossroads: The New Biology and the New Philosophy of Science* pp. 21-42). Cambridge Mass.: MIT Press.

Burian, R. M., Richardson, R. C., & Van der Steen, W. J. (1996). Against Generality: Meaning in Genetics and Philosophy. *Studies in History and Philosophy of Science, 27,* 1-30.

Carlson, E. A. (1966). *The Gene. A Critical History*. Philadelphia: Saunders.

Carrier, M. (2001). Changing Laws and Shifting Concepts: On the Nature and Impact of Incommensurability. In P. Hoyningen-Huene, & H. Sankey (Eds.), *Incommensurability and Related Matters* pp. 65-90). Dordrecht: Kluwer.

Darden, L. (1991). *Theory Change in Science: Strategies from Mendelian Genetics*. Oxford: Oxford University Press.

Ellis, B. (2001). *Scientific Essentialism*. Cambridge: Cambridge University Press.

Field, H. (1973). Theory Change and the Indeterminacy of Reference. *The Journal of Philosophy, 70,* 462-481.

Kitcher, P. (1978). Theories, Theorists and Theoretical Change. *Philosophical Review, 87,* 519-547.

Kitcher, P. (1982). Genes. *British Journal For The Philosophy Of Science, 33,* 337-359.

Kitcher, P. (1993). *The Advancement of Science. Science Without Legend, Objectivity Without Illusions*. Oxford: Oxford University Press.

Kripke, S. A. (1980). *Naming and Necessity*. Cambridge: Harvard University Press.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (2nd ed. ed.). Chicago: University of Chicago Press.

Kusch, M. (2002). *Knowledge by Agreement. The Programme of Communitarian Epistemology*. Oxford: Oxford University Press.

LaPorte, J. (2003). *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Laudan, L. (1984). A Confutation of Convergent Realism. In J. Leplin (Ed.), *Scientific Realism* pp. 218-249). Berkeley: University of California Press.

Lewis, D. (2000). Causation as Influence. *The Journal of Philosophy, XCVII,* 182-197.

Luria, S., & Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics, 28,* 491-511.

McLeish, C. (2005). Scientific Realism Bit by Bit: Part I. Kitcher on Reference. *Studies in History and Philosophy of Science, 36,* 668-686.

McLeish, C. (2006). Realism Bit by Bit: Part II. Disjunctive Partial Reference. *Studies in History and Philosophy of Science, 37,* 171-190.

Mitchell, S. (2000). Dimensions of Scientific law. *Philosophy of Science, 67,* 242-265.

Morgan, T. H., Muller, H. J., Sturtevant, A. H., & Bridges, C. B. (1915). *The Mechanism of Mendelian Heredity*. New York: Henry Holt & Co.

Moyal, A. (2001). *Platypus. The Extraordinary Story of How A Curious Creature Baffled the World*. Washington D.C.: Smithsonian Institution Press.

Neumann-Held, E. M. (1999). The Gene is Dead - Long Live the Gene! Conceptualizing Genes the Constructionist Way. In P. Koslowski (Ed.), *Sociobiology and Bioeconomics. The Theory of Evolution in Biological and Economic Thinking* pp. 105-137). Berlin: Springer.

Nola, R. (1980). Fixing the Reference of Theoretical Terms. *Philosophy of Science, 47,* 505-531.

Oyama, S. (2000). Causal Democracy and Causal Contributions in Developmental Systems Theory. *Philosophy of Science (Proceedings), 67,* S332-S347.

Portin, P. (1993). The Concept of the Gene: Short History and Present Status. *Quarterly Review of Biology, 68,* 173-223.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Putnam, H. (1973). Meaning and Reference. *The Journal of Philosophy, 70,* 699-711.

Rosenberg, A. (2006). *Darwinian Reductionism. Or, How to Stop Worrying and Love Molecular Biology.* Chicago: The University of Chicago Press.

Sankey, H. (1994). *The Incommensurability Thesis*. Aldershot: Ashgate.

Stanford, P. K., & Kitcher, P. (2000). Refining the Causal Theory of Reference for Natural Kind Terms. *Philosophical Studies, 97,* 99-129.

Sterelny, K., & Griffiths, P. E. (1999). *Sex and Death: An Introduction to Philosophy of Biology*. Chicago: University of Chicago Press.

Waters, C. K. (1994). Genes Made Molecular. *Philosophy of Science, 61,* 163-185.

Waters, C. K. (1998). Causal Regularities in the Biological World of Contingent Distributions. *Biology and Philosophy, 13,* 5-36.

Waters, C. K. (2000). Molecules Made Biological. *Revue Internationale De Philosophie, 214,* 9-34.

Waters, C. K. (2004). What Was Classical Genetics? *Studies in History and Philosophy of Science, 35,* 783-809.

Weber, M. (1998). Representing Genes: Classical Mapping Techniques and the Growth of Genetical Knowledge. *Studies in History and Philosophy of Biological and Biomedical Sciences, 29,* 295-315.

Weber, M. (1999). The Aim and Structure of Ecological Theory. *Philosophy of Science, 66,* 71-93.

Weber, M. (2005). *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

Weber, M. (2006). The Central Dogma as a Thesis of Causal Specificity. *History and Philosophy of the Life Sciences, 28,* 565-580.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

28

Wittgenstein's Essentialism

ROGER POUIVET

Abstract Wittgenstein is often described as strongly anti-essentialist. The famous passage about "games" in his *Philosophical Investigations* is generally read as a declaration of war against essentialism. But when Wittgenstein says that "essence is expressed in grammar", how can we attribute to him the view that the notion of essence makes no sense? How could essence be simply a metaphysical toy, not to be taken seriously, if grammar - which is certainly taken very seriously in Wittgenstein's philosophy! - expresses it? I must recognize that a lot of good commentators understood this formula as a critic of the very notion of "essence". The formula is sometimes supposed to mean that essence is not at all what metaphysicians call "essence". But what do they are supposed to call "essence"? Against the received interpretation, I want first protest that the notions of definition and essence do not appear at all in the Philosophical Investigations' passage about games. One could answer that they are implicit. But not at all, I think. This passage is clearly about *explanation.* Wittgenstein does not ask "How to define a game?", or "Can we give the essence of games?". He is examining the question to know how can we explain to someone what is a game. He does not say that we cannot give a definition, but that a definition would not be of efficient use. We can give a lot of definitions of games and of numbers. Wittgenstein does not say that giving definitions is a stupid activity or that all definitions would inevitably be bad or crazy. Simply, most of definitions are not useful at all to explain to someone what a word means. So, in this famous passage that some Neo-Wittgensteinians interpreted as strongly anti-essentialist, Wittgenstein fights against the notion of "proper original signification", not especially against the notion of essence. And I think that it would be disputable to identify the two

or to authorize oneself from this passage to say that Wittgenstein rejects the true notion of essence. "Essence is expressed by grammar" perhaps means that meaning is not an ideal or mental object that speakers must have in mind when they speak about something, or that philosophers would be expert to discover. Meaning is related to use means that such an ideal or mental object is a myth. But I wonder why we had to consider that it says that the notion of "essence" is a myth. To be expressed by grammar is not exactly to be a myth that therapeutic philosophy would happily eradicate.

1. The falsehoods of idealism and the stupidities of empiricist realism

At the very beginning of his paper "The Problem with W*ttg*nst**n"[1], Pascal Engel says:

> No one can deny that there is a problem between Wittgenstein and analytic philosophers. To put it mildly, there are tensions between Wittgenstein's and Wittgensteinian styled reflections and the views and practice of a lot of contemporary analytic philosophers, such that they often seem to be strange bedfellows, when they are bedfellows at all.

I will comment on the tensions between the Wittgensteinian styldes reflections and and practice of analytic philosophers. But I will try to show why on the dominant interpretation about him, according to which Wittgenstein would be anti-essentialist and he would give reason to be so as much as him, is disputable. Pascal Engel speaks about Wittgensteinians who do not hesitate, when they think it fit, to defend theses, and he distinguishes them from Wittgensteinian Quietists. Surely, what follows will situate me in the first group! And if I am right in what follows, if Wittgenstein could be an essentialist, he has only a very weak family resemblance with W*ttg*nst**n.

Wittgenstein's essentialism? Is it a joke? He says: "essence is expressed by grammar". This formula is often quoted as if its sense is clear. It is in fact enigmatic. What does it exactly means? I think that its sense indicates a middle way between the falsehoods of idealism and the stupidities of empiricist realism. "It is enormously difficult to steer in the narrow channel here:

[1] P. Engel, "The Problem with W*ittg*nst**n", *Rivista di Estica*, *Homagio a Diego Marconi*, 2007.

to avoid the falsehood of idealism and the stupidities of empiricist realism"[2], says Elizabeth Anscombe. Let us try to find our way in between, if there is one. But Anscombe is right: on such topic, it is enormously difficult not to be wrong and not to be silly, and we are even in danger to be both. Wittgenstein tried hard to find the middle way, and this is why he interests me, and is not so closed to some received interpretation of his philosophy.

2. The received interpretation and its critic

I must recognize that a lot of good commentators understood the formula "essence is expressed by grammar" as a critic of the very notion of "essence". The formula is sometimes supposed to mean that essence is not at all what metaphysicians call "essence". But what do they are supposed to call "essence"? Locke said that essence in the "proper original signification" of the word, it is "the very being of any thing, whereby it is, what it is"[3]. In short, the essence of something, X, is what X is, or what it is to be X. In another locution, X's essence is the very *identity* of *X*.

So, by grasping the essence of something, you have at your disposal, necessary and sufficient conditions for *X* to be what it is. And Wittgenstein is supposed to have shown that in fact you have not such conditions. This is the received interpretation of the famous §66 in the *Philosophical Investigations.* And even if you never read this book, you know that empirical concepts has been showed by Wittgenstein to be predicate of family resemblance. Metaphysics is dead and Wittgenstein definitively buried the Aristotelian notions of concept and essence. For, we are perfectly able to use the word "game" even if we are unable to give necessary and sufficient conditions for something to be a game. So, meaning of the word "game" is not related to essence, the very identity of game, but to uses of this word in different language games. We have no essence by similarities between many uses of the word "game".

> I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family: build, features, colour of eyes, gait, temper-

[2] G.E.M. Anscombe, "The Question of Linguistic Idealism", From Parmenides to Wittgenstein, Oxford: Blackwell, 1981, p. 115.

[3] See J. Locke, *An Essay Concerning Human Understanding*, ed. P. H. Nidditch, Oxford: Clarendon Press, 1975, III, III, 15.

> ament, etc. etc. overlap and criss-cross in the same way. — And I shall say: "games" form a family."[4]

Well, if "games" form a family, the metaphysical notion of essence is supposed not practicable.

Against this received interpretation, I want first protest that the notions of definition and essence do not appear at all in this passage. One could answer that they are implicit. But not at all, I think. This passage is clearly about *explanation.* Wittgenstein does not ask "How to define a game?", or "Can we give the essence of games?". He is examining the question to know how can we explain to someone what is a game. He does not say that we cannot give a definition, but that a definition would not be of efficient use. We can give a lot of definitions of games and of numbers.[5] Wittgenstein does not say that giving definitions is a stupid activity or that all definitions would inevitably be bad or crazy. Simply, most of definitions are not useful at all to explain to someone what a word means.

Wittgenstein rejects the idea that to be useful a concept must be determined by a system of rules, and that the understanding of a word (or a sentence) is based on or supposes to use defined rules. For, even if we have such rules, they could be interpreted in many ways. This is the Platonist illusion: believing that there is something behind our use of words, something that would justify philosophically our linguistic uses and practices. Meaning would be the rule of the right use. Not at all, says this famous passage on games in Wittgenstein's *Philosophical Investigations*. But it does not only challenge Platonist stance. Philosophers gave other status to this notion of meaning as a rule: the notion of idea in the seventeenth and eighteenth century philosophy played in part this role, and the notion of mental state in contemporary philosophy can play the same. It is the general philosophical tendency to identify meaning to an ideal or mental object, which could be grasp (especially by philosophers) and would serve to determine, as a sort of rule, what we are speaking about when we use a term. Geach called it the "Socratic fallacy"[6].

So, in this famous passage that some Neo-Wittgensteinians interpreted as strongly anti-essentialist, Wittgenstein fights against the notion of "proper original signification", not especially against the notion of essence. And I

---

[4] L. Wittgenstein, *Philosophical Investigations,* tr. G.E.M. Anscombe, Oxford: Blackwell, 3rd ., 1967, I, 67.

[5] For numbers, see *Philosophical Investigations*, I, 68.

[6] P. Geach, "Plato's *Euthyphro*", Logic Matters, Oxford: Blackwell, 1972.

think that it would be disputable to identify the two or to authorize oneself from this passage to say that Wittgenstein rejects the true notion of essence. "Essence is expressed by grammar" perhaps means that meaning is not an ideal or mental object that speakers must have in mind when they speak about something, or that philosophers would be expert to discover. Meaning is related to use means that such an ideal or mental object is a myth. But I wonder why we had to consider that it says that the notion of "essence" is a myth. To be expressed by grammar is not exactly to be a myth that therapeutic philosophy would happily eradicate.

3. Some absurd questions

As Elizabeth Anscombe suggests, that the formula "essence is expressed by grammar" is related to the absurdity of certain sentences as "Where does this pencil uncle live?", "What is the shape of dust?", "What is a rainbow made of?", "How many legs has a tree?", "What does a chair feel?", "Do bacteria think?".[7]

In fact, she's in partly wrong. It would be possible to answer these questions. For example, "What is the shape of dust?" could be the beginning of a poem. Imagine this one:

> "What is the shape of dust?
>
> This is the way I feel my love
>
> This is the way I need your glove
>
> During the night, during the day,
>
> Trembling, crying, I must."

I do not suggest that my first and only poem in English is aesthetically good, but that it would be possible to ask, poetically, this question "What is the shape of dust?", and even to give, poetically, an answer (what follows this question in my poem). There is a language game where a question like "What is the shape of dust?" is in use. "How many legs has a tree?" is also the kind of question you could ask to a child, and he could answer: "Well, poor tree,

[7] Anscombe, "Human Essence", Human Life, Action, and Ethics, ed. by M. Geach & L. Gormally, Exeter: Imprint Academic, 2005.

it has no legs and so it cannot run", and even be sorrow for a moment about this tree.

But Elizabeth Anscombe is right for the essential. To say that "Essence is expressed by grammar" means: "Look at the way some questions would be considered as absurd, except in special circumstances, in poems, in conversations with kids. To grasp the essence of the thing one refers in this question is to realize that such questions make no sense. It is because you possess the essence of the thing one refers to, that the question appears crazy."

When you ask, for example, "Do bacteria think?", this is a sin against what Wittgenstein calls "grammar", but not if you ask "Do human beings think?". This last question makes sense. And to say that the first question is a sin against grammar and that the last question makes sense, that is not to make a remark inside the domain of biology, or in the domain of psychology. This observation is fully a grammatical one – it is about the grammar of the verb "to think". But *at the same time,* and without entering into the domains of biology and psychology, this remark says something about what bacteria *are* and what human beings *are*. Our human ability to grasp the essence of bacteria and the essence of human beings is nothing else and more than to have the disposition to consider that the question "Do bacteria think?" makes no sense, and that the question "Do human beings think?" makes sense. For example, you could say: "Well, of course, human beings think, for man is a rational animal". This answer would be a grammatical one, because finally it says that "Do *X* thinks?" makes no sense if X does not symbolize "human beings", "angels", "God", and also, at the limit and in certain circumstances when compared to lower animals, "dog", "rabbit". But definitively it makes no sense concerning "bacteria" or "woodlouse". By answering this way, you apply some rules belonging to what Wittgenstein calls "grammar", but you also indicate what are the essence of human beings and the essence of woodlice.

To grasp essences of things around us is not a specific philosophical activity. Even a child grasps the essence of human beings and woodlice by remarking that "Do human beings think?" makes sense and "Do woodlice think?" makes not. And a child does it even if he cannot use correctly the word "essence" in its metaphysical sense.

Sometimes, and perhaps often, men are highly irrational; some men do not develop the capacity to use spoken language; human babies are unable to reason; passionate lovers seem crazy. But it does not change anything to the fact that "Do woodlice think?" makes not sense and "Do human beings think?" makes one. And nobody would say that her baby is for the moment a woodlouse or close to it, because she does not think. Nobody would say that

these lovers are closed to woodlice because they seem not to think at all about the present and the future. Even if a baby does not speak and is clearly unable to have a rational behaviour, even if your good friend Jack, terribly in love, makes incredibly stupid things, the baby is already a human being and Jack still one. Why? For they have an essence or a nature. And everybody knows this, not only metaphysicians. Metaphysicians are only those who remark that we are able to grasp essences, or contest that we do it or can do it.

4. Linguistic idealism, dolphins and pain

Here, someone can object that it is a little bit strange to embark Wittgenstein in the defense of an apparently naive or commonsensical notion of reality and essence. What I mean by "naive or commonsensical realism" is the thesis that there is some differentiation in the world before we came to experience it. A lot of modern philosophers, following Hume and Kant, and developing sometimes strongly anti-realist accounts about reality and language, characterizes themselves by the thesis that there is no differentiation independently of the way we experience it. I would like to suggest that if they insist so strongly on the notion of "experience", the first and final word for a lot of modern philosophers, it is because they think that everything is inside this experience. And some of them think that this experience is itself mainly language-laden. For example, Hilary Putnams says:

> We can and should insist that some facts are there to be discovered and not legislated by us. But this is something to be said when one has adopted a way of speaking, a language, a "conceptual scheme". To talk of "facts" without specifying the language to be used is to talk of nothing; the word "fact" no more has its use fixed by the world itself than does the word "exist" or the word "object".[8]

Very often today, philosophers who think this way authorize themselves from Wittgenstein, to say that everything is inside language. So "Essence is expressed by grammar" would mean the linguistic nature of essence. Reality

[8] H. Putnam, *Representation and Reality*, Cambridge: The MIT Press, 1988, p. 114. Note that the proliferation of "..." is a good indication that you are in the realm of antirealism. To speak about truth is metaphysical naivety; to speak about "truth" is anti-realistic or even deconstructionist enlightenment.

would not exist as a "thing in itself", independently of our linguistic activity. Reality and essence is only mirroring the way our language represents it.

But Wittgenstein's formula is not necessarily linguistic idealism, even at its best, and even if it resembles a lot to it. At least, I think it is possible to give a sense to Wittgenstein's formula without embracing linguistic idealism or a conception dangerously close to it.

But could we reintroduce the notion of "essence" and of "nature of things" without losing the therapeutic effect of Wittgenstein's philosophy? If grammar is a way to metaphysical knowledge, what difference there is between, from one side, Aristotle and Thomas Aquinas, and from another supposed very different side, Wittgenstein? Well, perhaps not a lot, and this is a conviction I have since some years that Wittgenstein's philosophy is inside a Thomistic tradition in philosophy.[9] To say "essence is expressed by grammar" is not at all for me a way out of this tradition.

I go back to the idea that "Do human beings think?" makes sense, and "Do woodlice think?" makes no sense. Let us examine the difference as a grammatical one. We will discover that it has also, at the same time, a metaphysical one. For this, let us introduce dolphins and ask "Do dolphins think?" Not to answer these questions as a zoologist would do, but simply to ask what it metaphysically means that the first question makes sense and not the second.

Let us first read this passage by Alasdair McIntyre:

> Although our differences from all other species are certainly of crucial importance, it is also important that both initially in our earliest childhood activities and to some significant extent thereafter we comport ourselves towards the world in much the same way as other intelligent animals. In transcending some of their limitations we never separate ourselves entirely from what we share with them. Indeed our ability to transcend those limitations depends in part upon certain of those animal characteristics, among them the nature of our identity.[10]

The way we speak about some animals, saying that they think, makes sense. But it does not mean that we have the same nature. Our way of speaking and especially asking about things register fine-grained ontological differences: it makes sense to say that human beings think, and to be shocked if someone

[9] See my *After Wittgenstein, saint Thomas,* South Bend: St Augustine's Press, 2007.

[10] A. McIntyre*Dependent Rational Animals,* Chicago: Open Court, 199, p. 8.

says that dolphins do not think for they are only animals, and also to consider that "Do woodlice think?" makes no sense. In the great chain of beings, dolphins and human beings, even if they do not have the same nature, are closer than woodlice and human beings. Conversely, what makes so disgusting – I do not find another word to characterize it – Descartes' insistence that non-human animals not only lacks thoughts and intelligence, but also genuine perceptions and feelings? Simply that, if someone beat a dog in front of you, it would be difficult to maintain that the question "Do animals suffer?" makes no sense, or that true philosophy (as Malebranche suggested) would permit us to correct this false impression that such a question makes sense.

"Essence is expressed by grammar" does not mean that essence is *created* by grammar. Wittgenstein's formula, even if it is enigmatic, is quite far from the cosmic-porridge view. According to it, there would be an indeterminate stuff, the cosmic porridge itself, and our concepts and our words would cut into it *at libitum*. Putnam – or a Putnam-temporal-slice – for example endorsed this kind of view: "We cut up the world into objects when we introduce one or another scheme of description"[11]. Goodman would say that the notion of this cosmic porridge makes no sense, for we cut and re-cut not the same stuff, but inside already artificial schemes, re-schematizing them. Both Putnam and Goodman are agree that realism, for which there is some essential differentiation in the world before we came to experience it, is metaphysical bullshit. Well, but if essence is expressed by grammar, it is likely there before we speak about it. And when we speak we are not simply cutting into the cosmic-porridge or into previous schematization. By refusing sense to "Do woodlice think?", we indicate a metaphysical sensibility, by making sense both to "Do dolphins think?" and "Do human beings think?" too, even if it does not mean of course that dolphins and human beings share some nature. But there are closer than human beings and woodlice, and the question of the dolphins' thought makes sense, when the question of the thought of woodlice makes not.

"If we assent to 'Essence is expressed by grammar', we may very likely say 'The words for what I am talking about *have to have* this grammar'"[12], says Elizabeth Anscombe. One Wittgensteinian example of this account would be that the language for talking about sensation *must have* first-third person asymmetry. This is a feature of our language. But it also means something about the *nature* of those beings that have sensations, especially human be-

[11] H. Putnam, *Reason, Truth, and History*, Cambridge: Cambridge University Press, 1981, p. 52.
[12] "The Question of Linguistic Idealism, p. 112, my italics.

ings. This is the same reasoning than for question like "Do human beings?" and "Do dolphins thin?", "Do woodlice think?". Inside grammar, we speak about the nature of things, not only about words and the way we use them. Attention for uses and for practices does not mean that reality is evanescent and that things have no natures or essences.

Wittgenstein rejects "Platonism" if it consists in saying that the grammar of our language must correspond to an independent reality. No doubt, he rejects the mirror account of language. But we, and Wittgenstein like us, want to be assured that what we say to be actually exists and is not mere projection of the way we are speaking. If I say that human beings are rational animals, I do say that if a human being does not speak or is irrational, however he is a rational animal. For rationality is his essence. This is expressed in language and it concerns the way we are using some words. But even if to have the concept of "human being" consists for example in applying rightly "human beings" even to irrational persons, without be impressed by the irrational behaviour, it does not mean that the case for the nature of human beings is a product of that grammar.

"You learned the *concept* pain when you learned language"[13] says Wittgenstein. Does it mean that pain is a product of a word? No, it means that it is not experiencing pain that gives you the meaning of the word "pain". For the word applies to that experience, but also to another one, yours and another one experience. And so, the term cannot find its meaning in a private experience.

Essence of pain is expressed by grammar. But it does not mean that the notion of essence of pain makes not sense. It means that it is not something that we could grasp without mastering language and the right use of "pain".

5. On riding a horse

Elziabeth Anscombe says also:

> If there never had been human beings around talking about horses, that is not the slightest reason to say there wouldn't have been horses. These essences, then, which are expressed by grammar, are not created by grammar. It must be misunderstanding of 'essence' to think otherwise: to think, for example, that though there doubt-

[13] *Philosophical Investigations*, I, 384.

> less would have been horses, the essence expressed by "horse" would not have existed but for human and thought.[14]

Could we be in the situation where according some "general facts of nature", human beings did not have the concept "horse", but they were horses around them.[15] Wittgenstein seems to say that these human beings devoid of the concept "horse" would not *miss* something that we, with this concept, realize. We are tempted to interpret this affirmation in an anti-realistic sense. "Horse" would be a concept of our own, without correspondence with anything that exists independently of the projection of this concept on reality. But Wittgenstein could mean something very different. These human beings without the concept "horse" could not miss anything if they had *another way,* in their own language, to speak about horses. For example, as Anscombe suggests, they could have a verb meaning horse-presence, but without the concept "horse". So, not to miss something when you have not the concept "horse", does not mean at all that horses have no essence. It is possible not to miss something when you have not the concept "horse" and that there is some differentiation in the world before we came to experience it. To pass from Wittgenstein's motto, "essence is expressed by grammar", to anti-essentialism, seems clearly now, I hope so, to be *anti-realist wishful thinking*. Such an interpretation embarks Wittgenstein in an anti-metaphysical fight he seems not to have been at all a strong partisan. He is simply disguised in a Post-modern philosopher he was not.

Let us suppose that someone says: "Do you see this horse?" Wittgenstein suggests that the intelligibility of such a question is not related to the existence of an ideal entity that would give a meaning to the word or to the concept "horse", and it is no related to an image or a representation in the mind. He rejects the kind of theory of meaning that flourishes during the Seventeenth Century, when the notion of "idea" begun to be central (Descartes, Malebranche, Locke, Berkeley, Hume). I think that Wittgenstein is philosophically impressive for the way he has been able to criticize this very strong paradigm in philosophy which still influences a lot of current philosophy (and peculiarly cognitive sciences).

I think that we can understand Wittgenstein's account another way. When one explains Wittgenstein, it is very difficult not to simply quote or paraphrase Wittgenstein, two operations which explain generally nothing... So I prefer to reconstruct what I take to be a Wittgensteinian account of the meaning of

[14] "The Question of Linguistic Idealism", p. 114.

[15] See *Philosophical Investigations,* II, xii.

a word, and to show why the notion of "Wittgenstein's essentialism" is not completely absurd, even if it is not orthodox.

- By saying "horse", I do not refer to anything else than this thing in front of me I ask someone if he sees it. I am not asking him about an idea, a mental reality, an intentional object, a noem, or I don't know what (the kind of "objects" philosophers like, but which you are unable to find out of the philosophical class), but about *this.*
- By saying "horse", I do not say "Black Jack". I do not indicate the *name* of this horse, but *this*. And by doing it, I indicate what it is, its *essence,* shared by all horses, and among them this one.
- It is by the grammar expressing the essence that the word I am using expressed a kind of animal, and that I mean exactly what I mean by asking the question.
- By asking "Do you see this horse?", I do not speak about the essence, and I do not mean it.
- The essence is *through what* I understand, think, mean, something, here a horse am I asking someone if he sees it.
- I master the use of the word "horse". It shows that *I know what is a horse*. (This could be a mule, not a horse. What shows that I know what a horse is, is that I could say to the person who asks me "Do you see this horse?" – "This is not a horse, it is a mule!" It would be silly to say that I have simply show that I am mastering two words "horse" and "mule", and not that I know what is a horse and what is a mule, even that I know t better than my interlocutor.)
- If my interlocutor answers me "Well, yes. And have you seen this other one?", the essence (of a kind of things) is present in our conversation. (It would also be the case if I answer: "This one is not a horse, my dear, but a mule". To know an essence, it is to be able to do such distinctions.)
- If I say "The farmer made this horse during the night", my interlocutor could look at me by wondering what I mean. Surely, the question would be to know if I am truly mastering the word "horse" use. But there is also a problem concerning if I know what is a "horse", if I possess the essence of "cability".

- Finally, it can be useful to quote Elizabeth Anscombe:

> If, then, seeing a donkey, he supposes it too is a horse, he might say "But isn't it the same as you pointed to before?" showing that the identity in question is identity of kind. 'Pointing twice the same' is an expression that does not yet determine what counts as that: the question has only a determinate answer when we know what identity, what method of counting, is relevant. A horse has been counted and another horse comes along; if the procedure is to say 'we've counted that one', but to assign a *new* number to, say, a giraffe (a giraffe not having been counted before) – then it appears that one is counting kinds. But what one is counting is in any case out there before one, and not in either case a 'creature of mind".[16]

To know the essence of something is closely related to the ability to count things, for example kinds of animals in a zoo. (It is clear for me that the visit to the zoo with children is a metaphysical moment: "What exists in the Creation?" – "It exists this, and this, and this, and that…")

To say that essence is expressed by grammar means: to count kinds is different from counting individuals. For example, there could be here many horses, but it means that by saying "This is an horse, and this is an other one, and still an other one", I indicate that these things are the same. My mastering of the word "horse" in this case, means something about the kind of generic identity these things have, and so the kind of things they are. To do this is grasping the essence of horse. By grasping this essence, I am not accessing to Platonic Forms that make all the horse-copies of Horse (or cabality), or a general idea of cabality. I am simply able to characterize something as an horse, it means the same than other things which are horses.

That the meaning of expressions depends upon linguistic practice (that includes a lot of non linguistic elements, for example certain moves, like to look in some direction when you are asked "Do you see this horse?"), this does not mean that human ability to recognize things for what they are makes no sense. And so essence can be expressed by grammar without loosing is metaphysical value to be through what we grasp things as they are. Essence is not an ethereal thing, ideal, mental, out of language. It is related to our linguistic use, but this one is not purely verbal. There is a way in between the stupidities of realism and the falsehoods of idealism. I am not sure that it is so narrow as it seems.

[16] "The Question of Linguistic Idealism", p. 116.

6. From Wittgenstein to Aristotle

If the formula "essence is expressed by grammar" can be understood the way I propose, it seems to me that the meaning of this phrase is not so far from what Aristotle says in the well-known chapter 2 of the *Categories*. This is the passage:

> Of things themselves some are predicable of a subject, and are never present in a subject. Thus "man" is predicable of the individual man, and is never present in a subject. By being "present in a subject", I do not mean present as parts are present in a whole, but being incapable of existence apart from the said subject. Some things, again, are present in a subject, but are never predicable of a subject. For instance, a certain point of grammatical knowledge is present in the mind, but is not predicable of any subject; or again, a certain whiteness may be present in the body (for color requires a material basis), yet it is never predicable of anything. Other things, again, are both predicable of a subject and present in a subject. Thus while knowledge is present in the human mind, it is predicable of grammar. There is, lastly, a class of things which are neither present in a subject nor predicable of a subject, such as the individual man or the individual horse. But, to speak more generally, that which is individual and has the character of a unit is never predicable of a subject. Yet in some cases there is nothing to prevent such being present in a subject. Thus a certain point of grammatical knowledge is present in a subject.

This is surely among the most commented passage in all the history of philosophy. (I let aside the question to know if this passage represents a first account in Aristotle's philosophical development, superseded by his doctrine in *Metaphysics*.) It celebrates the wedding of metaphysics with the doctrine of predication. The very notion of "theory of predication" seems at odd with the idea of language games. Is it not a theoretical study of language cut from our uses and practices, exactly what Wittgenstein presented as a fundamental philosophical error? But in other sense, what is a theory of predication except what Wittgenstein calls "grammar"? If we examine the chapter 2 of *Categories,* is it so far from a reflection on language games? Some of you will answer: "Oh, yes, very far. It does not represent at all what are linguistic practices! It is metaphysics, exactly what Wittgenstein rejects!" But others may be tempted to recognize that Aristotle is simply trying to examine *what we say*

about things; for him, this is a way to determine what really exists. His method is to follow what we say to be or not to be predicable of a subject and to be present or not in a subject. I am not a scholar in ancient philosophy. I don't know if Aristotle's text is very corrupted or not. But I would bet that a phrase have been lost in this passage. The beginning surely was what Aristotle said at his students: "Look and see what we say about things! For if you look at them you will see that we speak about substances and say that they have ways of beings, and other properties, like to be colored, to be in certain places, to endure, and so on! Don't turn yourself to Forms, which are supposed to be supreme realities imitated by empirical things and by meanings of the words we use to represent them in language. Don't speak about participation of empirical and relative things to transcendent absolute one. Don't use a dubious metaphor. Don't try to determine this way what particular things have in common. Examine simply how we speak about things."

Aristotle proposes a kind of immediate and commonsensical ontology that appears when we examine our way to speak about things. I agree, this is not Wittgenstein's way to speak. He is ontologically abstinent. When you wrote *the Tractatus logico-philosophicus,* metaphysically you are like a repented alcoholic. You promise not to drink even a drop of metaphysical alcohol. But if you have always be able to drink, without becoming addicted, a glass of good wine – I mean of good metaphysics –, you have no reason to go to the Alcoholics Anonymous or to be metaphysical abstinent. If you have never to rely too heavily on metaphysics, pretending that you can discover supreme realities or ideal meanings, if you simply say that by saying there is a class of things which are neither present in a subject nor predicable of (say of) a subject, such as the individual man or the individual horse, you have really no reason to be abstinent. So I propose a non-abstinent interpretation of Wittgenstein's formula "essence is expressed by grammar". Grammar tells us what there is, and essence of things.

Interpretations of this passage in *Categories* go from one extreme to another. At one, we find the great German scholar Friedrich Tredelenburg (and the French linguist Émile Benvéniste). Thought cannot be separated from language. Categories, and among them the category of substance who indicate what something is, gives the thing nature, are simply grammatical characterizations. Call it the linguistic interpretation. It says that predication, and so grammar, constitutes categories, and so essences. Another great German scholar, Hermann Bonitz, situated himself at the other extreme (and he was followed by Brentano at the beginning of his career). Aristotelian categories are the genres of being, the many meanings of being. Categories, and so

essences, are foundations of predications. A non-abstinent Wittgensteinian interpretation of Aristotle would say that truth about this passage is in between the falsehood of idealism (Tredelenburg?) and the stupidities of empiricist realism (Bonitz?). You cannot go directly to essences without examining grammar (predication), but grammar (predication) is not the last word. World is the last word, it means what exists and what are things that exist, their essences, their identities.

*

It could be objected that Wittgenstein is not so interesting if he simply says the same than Aristotle. But you remember what Peter Strawson says in the preface of *Individuals*:

> If there are no new truths to be discovered, there are old truths to be rediscovered. For though the central subject-matter of descriptive metaphysics does not change, the critical and analytical idiom of philosophy changes constantly. Permanent relationships are described in an impermanent idiom, which reflects both the age's climate of thought and the individual philosopher's personal style of thinking.[17]

Wittgenstein reflected the age's climate of thought, and he has unquestionably a personal style of thinking. He rediscovered, sometimes chaotically and in his own style, old truths. The equilibrium he tries to find between linguistic idealism and naïve realism is, I think, the same than the one proposed by Aristotle in *Categories,* chapter 2. If everything does not depend to our linguistic scheme, it does not mean that we are committed to the contrary proposition that nothing depends on it. So it is both possible to insist on language games and practices, and to say that things have essences. It means that one can say: "essence is expressed by grammar".

[17] P. Strawson, London: Methuen, 1959, p. 10-11.

29

The speech acts account of derogatory epithets: some critical notes

CLAUDIA BIANCHI

Pascal's research interests are extremely comprehensive, venturing into many fields and touching many interdisciplinary themes, often with an uncommon attention to the civil and public relevance of philosophical issues. I hope, then, that he won't dislike the – in many ways so unlikeable – topic of this paper.

1. Introduction

Derogatory epithets are terms such as 'nigger', 'bitch' and 'faggot' targeting individuals and groups of individuals on the basis of race, nationality, religion, gender or sexual orientation. In recent years they have become an inspiring object of analysis in research fields as diverse as philosophy of language, linguistics, ethics, political philosophy, philosophy of law, feminist philosophy and critical race theory.[1] There is no consensus on the best treatment of derogatory epithets: each theory accounts for certain intuitions, but none seems completely satisfactory. The aim of my paper is to evaluate a proposal recently put forward by Rae Langton, the speech acts account (SAA). Assessing SAA is far from an easy task, since the proposal is little more than an outline, deeply intertwined with Langton's general view on hate speech and pornography. My goal is first of all to disentangle a coherent account from Langton's observations, mostly in Langton 2012 and Langton, Haslanger & Anderson 2012; second, I will raise and partially address some key objections against it. I will argue that, although SAA gives us significant insights into a number of phenomena, it is in need of a clearer formulation and further investigation.

2. Strategies of treatment of derogatory epithets

Derogatory epithets (from now on I will use the term 'epithets' for short) target individuals and groups of individuals on the basis of race, nationality, religion, gender or sexual orientation. They generally have a neutral counterpart, i.e. a non-derogatory term possessing at least the same extension of the derogatory one: 'nigger' and 'African-American' or 'black', 'bitch' and 'woman', 'faggot' and 'male homosexual'.

There are several alternative taxonomies of treatments of epithets. I will adopt here a classification in three perspectives: semantic, pragmatic and deflationary.

a) From a semantic perspective the derogatory content of an epithet is part of its conventional meaning (i.e. part of the truth-conditions of the sentence containing the term); therefore it is expressed in every (nonfigurative or ironic) context of utterance. In a simplified version, the meaning of 'nigger' may be expressed as 'African-American and despicable because of it' (Hom 2008:

[1] Dummett 1973: 454; Kaplan 1999, Hornsby 2001, Hom 2008, Potts 2008, Richard 2008, Williamson 2009, Predelli 2010, Anderson and Lepore 2011.

416). This strategy accounts for the largely shared intuition that epithets *say* offensive and derogatory things. In other words, the sentence

(1) Tom is a nigger

(having as a neutral counterpart

(2) Tom is an African-American)

says something we may paraphrase with

(3) Tom is an African-American and despicable because of it.

b) According to the pragmatic perspective,[2] the derogatory content of an epithet doesn't contribute to the truth-conditions of the sentence containing it, but is merely conveyed in context. The pragmatic perspective is usually spelled out in terms of presuppositions, tone or conventional implicatures. According to the strategy in terms of presuppositions, the offensive content of (1) isn't expressed or said but merely presupposed. According to the strategy in terms of Fregean tone, 'nigger' and 'African-American' are synonymous, and differ only in coloring or connotation. Finally, according to the strategy in terms of implicatures, (1) and (2) have the same truth-conditions, and the offensive content of an epithet may be assimilated to a conventional implicature.

c) The deflationary perspective opposes both strategies working in terms of content (a) and b)). There is no difference in content (expressed or implicitly conveyed) between 'nigger' and 'African-American': (1) and (2) have the same meaning. In a deflationary perspective, derogatory epithets are prohibited words not in virtue of any content they express or communicate, but rather because of edicts surrounding their prohibition – issued by relevant entities (targeted members, groups, or institutions).[3] Deflationists like Anderson and Lepore take a *silentist* stance: they suggest removing epithets from language until their offensive potential fades away, and avoiding any use or mention of them in any context, including so-called pedagogical contexts, where the speaker makes explicit the derogatory import of epithets or objects to discriminatory discourse.[4]

[2] I adopt here Hom's label 'pragmatic' to indicate strategies claiming that the derogatory content does not contribute to the truth-conditions of the sentence containing them (Hom 2008: 416). More particularly, I dub the strategies in terms of conventional implicatures and presuppositions 'pragmatic', although their (semantic or pragmatic) status is far from settled.

[3] Anderson and Lepore 2011: 16: '*once relevant individuals declare a word a slur, it becomes one*'.

[4] Anderson and Lepore 2011: 16: 'A use, mention, or interaction with a slur, *ceteris paribus* (...), constitutes an infraction'. On pedagogical contexts see *infra*, §3.

3. Adequacy conditions

Over the last fifteen years scholars working on epithets have identified a number of features characterizing how derogatory terms work in our ordinary conversations: these very features constitute a set of adequacy conditions that any satisfactory strategy of treatment of epithets must meet. For present purposes I will adopt Christopher Hom's list of features.[5]

1. *Derogatory force: Epithets forcefully convey hatred and contempt of their targets*. Epithets are generally perceived as more offensive than pejoratives (terms like 'stupid', targeting individuals and not groups of people).

2. *Derogatory variation: The force of derogatory content varies across different epithets*. Some epithets are perceived as more offensive than others: 'nigger' is considered by many the most insulting racial epithet in the USA.[6]

3. *Derogatory autonomy: The derogatory force for any epithet is independent of the attitudes of any of its particular speakers*. A speaker uttering a derogatory epithet expresses or conveys hatred or contempt towards an individual and a group of individuals independently of her beliefs or intentions.[7]

4. *Taboo: Uses of epithets are subject to strict social constraints, if not outright forbidden*. According to semantic and pragmatic perspectives, uses of epithets are acceptable only within quotations, fictional contexts, appropriation (see *infra,* 7). According to the deflationary perspective, there are no acceptable uses of epithets, and the taboo extends even to expressions that are phonologically similar but semantically unrelated.[8]

5. *Meaningfulness*. According to Hom 'Sentences with epithets normally express complete, felicitous, propositions' (Hom 2008: 427). In what follows, I will return to the alleged felicity of the speech acts performed with sentences containing epithets.

6. *Evolution: The meaning and force of epithets evolve over time to reflect the values and social dynamics of its speakers*. Expressions like 'gay' or 'Tory' were insulting in the past but are no longer perceived as offensive.

[5] Hom 2008: 426-430. Quotations from Hom are in italics.

[6] For a more cautious opinion, see Jeshion 2011.

[7] Cf. Alston 2000: 103-13 and Hornsby 2001: 138.

[8] Cf. the term "niggardly" (Kennedy 2003: 94-97). *The New Oxford American Dictionary* warns that the terms "niggard" and "niggardly" may cause unintended offense.

7. *Appropriation: Targeted members or groups may appropriate their own slurs for non-derogatory purposes*, in order to demarcate the group, and show a sense of intimacy and solidarity – as in the appropriation of 'nigger' by the African-American community, or the appropriation of 'gay' and 'queer' by the homosexual community.

8. *NDNA uses: Epithets can occur in nonderogatory, nonappropriated (NDNA) contexts*. According to Hom there are acceptable uses of epithets in so-called pedagogical contexts, where the speaker is objecting to discriminatory discourse, like:

   (4) Institutions that treat Chinese people as chinks are racist,
   (5) There are no chinks; racists are wrong,
   (6) Chinese people are not chinks,
   (7) Yao Ming is Chinese, but he's not a chink.[9]

9. *Generality: The account of derogatory force for epithets needs to generalize to similar language; for example, sexist, gender-biasing, religious epithets and approbative terms*.

4. The speech acts account (SAA)

I previously stated that there is no consensus on the best account of derogatory epithets. There are indeed well-known problems with all three perspectives: for an overview of the main difficulties of the semantic, pragmatic and deflationary perspectives see respectively Anderson & Lepore 2011, Hom 2008 and Bianchi 2014. Each perspective accounts for certain intuitions, but none seems completely satisfactory; hence, it may be worthwhile to examin an alternative account belonging to the pragmatic perspective, recently put forward by Rae Langton. Drawing on Austin's speech acts theory, Langton focuses not on what derogatory epithets *say*, but on what they *do*. The derogatory content of an epithet isn't part of its conventional meaning: epithets are expressions used to do things, to perform certain speech acts.

As is well known, Austin emphasizes the performative dimension present in any use of language: with a famous slogan, 'to say something is to do something'. Within the same total speech act – the uttering of a sentence like

[9] Hom, 2008: 429. Predelli 2010 rightly classifies (7) as offensive and suggests adopting the sentence resulting from omission of the contrastive conjunction: (7') *Yao Ming is Chinese, not a chink*.

(8) Stay here!

Austin distinguishes three different acts: locutionary, illocutionary and perlocutionary. The *locutionary* act is the act of saying something, the act of uttering certain expressions, well-formed from a syntactic point of view and meaningful. The *illocutionary* act corresponds to the act performed in performing a locutionary act, to the particular force that an utterance like (8) has in a particular context: order, request, entreaty, challenge, and so on. According to Austin, by uttering a sentence we can bring about new facts, "as distinguished from producing consequences in the sense of bringing about states of affairs in the 'normal' way, i.e. changes in the natural course of events" (Austin 1975: 117): by uttering a sentence we may undertake obligations and legitimate attitudes and behaviors, institute new conventions and sometimes even modify the social reality. The *perlocutionary* act corresponds to the effects brought about by performing an illocutionary act, to its consequences (intentional or non-intentional) on the feelings, thoughts or actions of the participants.

Following Catharine MacKinnon 1987, Langton identifies a particular kind of illocutionary act: acts of *subordination*. An utterance of

(9) Blacks are not permitted to vote

in South Africa in order to enact legislation that reinforces apartheid may be conceived as an illocutionary act of subordination: it makes it the case that blacks are not permitted to vote. The same holds for a sign reading

(10) Whites only (MacKinnon 1987: 202).

According to Langton, the sign counts as an illocutionary speech act, ranking blacks as inferior, depriving them of certain important powers, demeaning and denigrating them, and legitimating discriminatory behavior: "it orders blacks away, welcomes whites, permits whites to act in a discriminatory way towards blacks. It subordinates blacks" (Langton 1993/2009: 35).

Moreover, Austin's framework is exploited by Langton in order to offer a defence of MacKinnon's controversial claim that pornography *subordinates* women by violating their civil right to equal civil status, and *silences* them by violating their civil right to freedom of speech.[10] According to Langton works of pornography can be understood as *speech acts* of subordinating women and silencing women.[11] More precisely, works of pornography may be conceived as speech acts in two distinct senses:

[10] See MacKinnon 1987.

[11] Cf. Langton 1993, Hornsby and Langton 1998, West 2003.

- as *perlocutionary* acts that cause subordination, and produce changes in attitudes and behaviours, including discrimination, oppression and violence;
- as *illocutionary* acts that can in themselves subordinate women, legitimate attitudes and behaviours of discrimination, advocate oppression and violence.

Further extending her view on pornography to racial and hate speech, Langton argues that epithets are expressions used to do things, to perform certain speech acts: "Austin's distinction between illocutionary and perlocutionary acts offers a way to distinguish speech that *constitutes* racial oppression, and speech that *causes* racial oppression".[12] As in the case of pornography, speech acts performed with the help of epithets may then be conceived as speech acts in two distinct senses:

- as *perlocutionary* acts that *cause* discrimination, and produce changes in attitudes and behaviours, including oppression and violence;
- as *illocutionary* acts that *constitute* racial or gender discrimination, legitimate beliefs, attitudes and behaviours of discrimination, advocate oppression and violence.

Let us consider the illocutionary thesis in more detail.

5. Three classes of illocutionary acts

According to SAA, derogatory epithets are apt for performing certain illocutionary speech acts. More precisely, Langton outlines a distinction between three classes of illocutions that S can perform by using a derogatory expression.

> **a.** Assault-like speech acts such as *persecuting* and *degrading*. By using an epithet S may directly attack, persecute or degrade her targets. Epithets are here "weapons of verbal abuse" (Richard 2008): the focus is on the targeted group and individuals. By uttering (1),

[12] Langton, Haslanger & Anderson 2012: 758. They underline that a similar approach is already present in Richard 2008 (a supporter of the expressivist view), p. 1: 'what makes a word a slur is that it is used to do certain things, that it has…a certain illocutionary potential'.

S isn't merely *asserting* something, but performing an illocutionary speech act of persecuting, degrading or threatening – an act directed towards Tom and all blacks.

b. Propaganda-like speech acts as *inciting* and *promoting* racial discrimination, hate and violence. Shifting the focus from targets to addressees ("prospective haters"[13]) S's utterance of (1) may be regarded as an act of propaganda, an act that incites and promotes racial oppression.[14]

c. Authoritative subordinating speech acts as *enacting* a system of racial oppression: derogatory expressions are used to classify people as inferior, to legitimate racial oppression, religious or gender discrimination, to deprive minorities of powers and rights.

6. Objections to SAA

As mentioned in the Introduction, evaluating SAA is far from an easy task. The proposal is little more than an outline, deeply intertwined with Langton's general view on hate speech and pornography. In this paragraph I will raise and partially address some key objections: my ultimate goal is to disentangle a coherent account from Langton's observations.

1. According to SAA, by uttering sentences containing derogatory epithets, S may perform a variety of acts of subordination: persecuting her targets (a), promoting racial oppression (b) or legitimizing behaviors of discrimination (c). Is Langton saying that the *mere presence* of an epithet makes (1), say, an act of persecution? Is the epithet a sort of illocutionary force indicating device (IFID)? Langton doesn't explicitly make this suggestion, but such a claim would actually fit well within Austin's conventionalist framework. Austin characterizes the illocutionary act as the *conventional* aspect of language (to be contrasted with the perlocutionay act). For any speech act "there must exist an accepted conventional procedure having a certain conventional effect, that procedure to include the uttering of certain words by certain persons in certain circumstances" (condition A.1, Austin 1975, p. 14): if the conventional procedure is executed according to further conditions, the act is successfully

[13] Langton, Haslanger & Anderson 2012: 758.

[14] "Promoting" may be understood in a perlocutionary, causal sense, and in an illocutionary, constitutive sense: cf. Langton 2012, p. 130: "'promote' is a verb that straddles both sides of Austin's distinction".

performed. Illocutionary acts are – *inter alia* – performed via conventional devices (like linguistic conventions): in this framework epithets may be regarded as conventional devices apt for performing acts of persecution.[15]

2. Langton spells out what kind of speech acts fall under c., but is far less explicit about a. and b.

Using Austin's taxonomy, Langton classifies authoritative subordinating speech acts (c.) as verdictives or exercitives. In the class of verdictives Austin includes acts (formal or informal, and concerning facts or values) of giving a verdict, estimate or appraisal (such as acquitting, reckoning, assessing, diagnosing). In the class of exercitives Austin includes acts of exerting powers, rights or influence (such as appointing, voting, ordering, warning). In Langton's view, derogatory expressions are used

- to classify people as inferior (verdictives: "a judgment that it is so", Austin 1975: 155);[16]
- to legitimate racial oppression, religious or gender discrimination, to deprive minorities of powers and rights (exercitives "a decision that something is to be so", Austin 1975: 155).

I suggest classifying a. and b. along the same lines:

Assault-like speech acts (a.) may be seen as verdictives, "a judgment that it is so". In other words, to perform an assault-like speech act amounts to assigning an institutional status (inferior) to a natural fact (being black).

Propaganda-like speech acts (b.) may be seen as exercitives, "a decision that something is to be so". To perform a propaganda-like speech act amounts to creating (or reinforcing) a new institutional fact (the subordination of blacks).[17]

3. It is unclear whether a. and b. are two distinct speech acts at all, or the same speech act as perceived by – or directed to – different audiences: its

[15] There is a potential objection: it is widely held that any expression serving as an indicator of illocutionary force must be without semantic content (Stenius 1967, 258–259). Nonetheless, Green 2000, in accounting for the behavior of a range of parenthetical expressions, argues for the idea that a part of speech can simultaneously have semantic content and indicate force.

[16] While Austin distinguishes between *Expositives* (acts that clarify reasons, arguments, or communications) such as describe, class, identify, call and *Verdictives* such as diagnose and describe (where describe appears in two different categories), Searle admits only one class of "assertive illocutionary verbs": Searle 1979, p. 25. Cf. Berdini and Bianchi 2013, Sbisà 2001 and 2013.

[17] Cf. Bach and Harnish 1979.

targets (a.) and so-called prospective haters (b.).[18] Langton is aware of this possibility, but focuses only on propaganda acts used as assault acts: "The distinction here [between assault and propaganda] is a context-sensitive one. Propaganda aimed at turning its hearers into racists could also be used as an attack on an individual" (Langton 2012, p. 131). In my opinion, the reverse case is equally interesting in this context: assault acts may be regarded as propaganda acts. By uttering (1), S is not simply attacking Tom and all blacks, but also promoting racial hatred and discrimination: (1) *constitutes* an incitement to discrimination, directed to addressees and bystanders.

4. Langton claims that hate speech is typically *a more ordinary illocution as well*: "it *asserts* that there is a Jewish conspiracy... *orders* blacks to keep away".[19] She seems to extend her claim to utterances containing epithets and argue that S is performing an act of subordination *by asserting* (1). This is common in other cases: we use assertions like 'I will come to your party' in order to perform acts of promising. Someone may object that in this way, subordinating speech acts must be conceived as *indirect* speech acts. However, as in the case of promises, SAA isn't committed to such a conclusion; as Kissine points out, "the fact that an utterance corresponds to the performance of two speech acts does not necessarily imply that one of them is indirect. Arguably, a speech act is indirect only if its content is distinct from that of the corresponding direct speech act".[20]

5. Another powerful objection concerns "authoritative speech acts"; for the sake of simplicity, let's focus on exercitives.[21] According to Langton, speech acts performed via epithets are exercitives – illocutions conferring or taking away rights or privileges, i.e. fixing what is permissible in a certain domain. Langton further claims that speech acts performed via epithets enact permissibility conditions that *subordinate* blacks because they i) unfairly rank blacks as having inferior worth; ii) legitimate discriminatory behavior towards blacks; iii) unjustly deprive blacks of certain important powers.

According to Austin, exercitives (and verdictives) are "authoritative speech acts": they presuppose that the speaker has a certain kind of authority or influence. In other words, authority is a crucial *felicity condition* for subordinating

[18] We should, perhaps, investigate whether other speech acts (like *teasing*) exhibit the same pattern.

[19] Langton, Haslanger & Anderson 2012: 758.

[20] Kissine 2013: 177.

[21] McGowan 2003 argues convincingly that verdictives may be reduced to exercitives.

speech acts. Yet, in most cases, speakers using epithets lack formal authority: (1) may be uttered in an ordinary conversation by an ordinary speaker. In order to account for this objection, Langton addresses the question along the same lines as the analysis of pornography – relying on McGowan's model of conversational exercitives.[22] According to McGowan 2003, any conversational contribution invokes rules of accommodation in Lewis' sense, and therefore changes the bounds of what is permitted in that conversation (in this sense, it is an exercitive). Hence, an utterance of (1) changes what is permissible in that conversation. The question of authority is less critical as far as conversational exercitives are concerned: the authority required of S is limited to the relevant domain, and any conversational participant must have authority over the actual conversation in which she is contributing: "It is clear that a competent contributor to a conversation is an authority over the conversation that he or she is creating".[23]

This solution has some unwelcome consequences. First, McGowan holds that all speech is, in some way, exercitive. She denies that this has the result of trivializing exercitive force; nevertheless, her claim seems bound to undermine Langton's thesis about speech that has the power to subordinate. Second, conversational exercitives seem to enact permissibility facts that are "easily reversible". Third, each conversational participant "seems just as able to change the permissibility facts of the conversation as any other participant".[24] These are three features we don't want to ascribe to subordinating speech acts.

6. One last point. Apparently, no peculiar authority is required in order to successfully perform an act of persecuting (a), promoting racial oppression (b) and legitimizing behaviors of discrimination (c). Langton doesn't specify when (if ever) acts of persecution, propaganda or subordination are *infelicitous*. What are the felicity conditions of acts of subordination?

7. SAA and conditions of adequacy

In §3 I presented a number of features characterizing the behavior of derogatory terms: these features constitute a set of adequacy conditions that any

[22] Cf. MacKinnon: "authoritatively *saying* someone is inferior is largely how structures of status and differential treatment are demarcated and actualized" (MacKinnon 1993 *Only words*, p. 31). Actually, Austinian exercitives have many features that do not fit well with Langton's claim: see McGowan 2003, pp. 164-169.

[23] McGowan 2003, p. 180.

[24] McGowan 2003, p. 187.

satisfactory account of epithets must meet. In this paragraph I will briefly examine whether SAA meets Hom's adequacy conditions.

1. *Derogatory force: Epithets forcefully convey hatred and contempt of their targets*.

According to SSA, using an epithet is far more insulting than using a pejorative like "stupid". As a matter of fact, by uttering sentences containing derogatory epithets, S may perform illocutionary acts of subordination: persecuting her targets (a), promoting racial oppression (b) or legitimizing behaviors of discrimination (c).

2. *Derogatory variation: The force of derogatory content varies across different epithets*.

Some epithets are perceived as more offensive than others: the derogatory force varies with the strength of the discriminatory system that acts of subordination contribute to enact and reinforce. It is crucial to SAA that uses of derogatory epithets are but an ingredient of a more comprehensive subordinating system.

3. *Derogatory autonomy: The derogatory force for any epithet is independent of the attitudes of any of its particular speakers*. According to SAA, by uttering a derogatory epithet, S performs an act of subordination towards an individual and a group of individuals independently of her beliefs or intentions. We have said that in an Austinian framework illocutionary acts are performed – *inter alia* – via conventional devices or linguistic conventions. More particularly, epithets may be regarded as conventional devices apt for performing acts of persecution, autonomous from the beliefs, attitudes and intentions of individual speakers.

4. *Taboo: Uses of epithets are subject to strict social constraints, if not outright forbidden*. Because epithets are conventional devices apt for performing acts of persecution, there are rigid social limitations ruling their use. Their use is appropriate only in quotations, fictional contexts and appropriation.

5. *Meaningfulness: Sentences with epithets normally express complete, felicitous, propositions*. The felicity of the *speech acts* (and not of the *propositions*, as Hom erroneously holds) performed with sentences containing epithets cannot be presupposed but must be argued for. Langton herself doesn't specify

the felicity conditions of acts of subordination, and seems to hold that acts performed with sentences containing epithets are always felicitous.

6. *Evolution: The meaning and force of epithets evolve over time to reflect the values and social dynamics of its speakers*.

We have said that epithets are devices used to enact and reinforce more comprehensive systems of oppression: those very systems may evolve over time, leading to changes in the derogatory force of the acts of subordination associated with them.

7. *Appropriation*. Targeted members or groups may appropriate their own slurs for non-derogatory purposes, in order to demarcate the group, and show a sense of intimacy and solidarity. I have argued elsewhere that appropriated uses may be conceived as echoic uses, in Relevance Theory terms: in-groups echo derogatory uses in ways and contexts that make manifest the dissociation from the offensive contents.[25] A second approach that SAA could adopt treats appropriation as a type of *pretense*. On this approach, a targeted member uttering (1) in an appropriated context is not performing an act of subordination but merely pretending to perform an act of subordination, while expecting her audience to see through the pretense and recognize the critical or derisive attitude behind it.[26]

8. *NDNA uses: Epithets can occur in nonderogatory, nonappropriated (NDNA) contexts*. SAA focuses not on what derogatory epithets *say*, but on what they *do*. In NDNA contexts the speaker isn't performing acts of subordination, but completely different speech acts: *objecting* to discriminatory discourse, *pointing out* the racist contents carried by epithets, *denouncing* the racist, misogynist, homophobic presuppositions that come with ordinary uses of epithets. Of course Langton owes us a detailed explanation of how a conventional device for subordination may be put to a new, non-derogatory, use.

9. *Generality: The account of derogatory force for epithets needs to generalize to similar language; for example, sexist, gender-biasing, religious epithets and approbative terms.*

[25] See Bianchi 2014. My echoic account suggests a solution compatible with the semantic and the pragmatic perspectives, that is with strategies of treatment of epithets in terms of content (expressed or conveyed).

[26] See Walton 1990.

More than alternative views, SAA provides us with a general framework for hate speech: derogatory expressions are used to classify people as inferior, to legitimate racial oppression, religious or gender discrimination, to deprive minorities of powers and rights. Furthermore, SAA offers a straightforward explanation for *approbative terms* as "angel", "blessed", "stud", "goddess" (Hom, 2008: 439): approbative terms are terms apt for performing acts of approval, praise and commendation.

8. Conclusion

The aim of my paper was to evaluate the speech acts account, recently put forward by Rae Langton. Assessing SAA is a challenging task for at least two reasons. First of all, the account is little more than a draft, not fully developed in its consequences and assumptions. Second, the model is deeply intertwined with Langton's arguments against pornography: it inevitably inherits some of the weaknesses of her general view on hate speech. I have argued that SAA needs a clearer formulation and further investigation. Nonetheless I hope to have shown that the proposal has interesting advantages over alternative views, gives us significant insights into a number of phenomena and certainly deserves careful consideration and further development.

9. References

Anderson, Luvell and Lepore, Ernest 2011. Slurring words. Nous, 1-27.

Austin, J. L. 1962. How to Do Things with Words, J. O. Urmson and M. Sbisà (Eds.), Oxford: Oxford University Press, 2nd edition 1975.

Berdini F. and Bianchi C. 2013 "John L. Austin", IEP – Internet Encyclopedia of Philosophy (`http://www.iep.utm.edu/austin`).

Bianchi, Claudia 2013. Slurs: un'introduzione. E/C, anno VII, n. 17, pp. 41-46,

Bianchi, Claudia. 2014. Slurs and Appropriation: An Echoic Account, forthcoming.

Croom, Adam 2011. Slurs. Language Sciences, 33, 343-58.

Dummett, Michael 1973. Frege's Philosophy of Language. Oxford: Clarendon Press.

Green, Mitchell 2000. Illocutionary Force and Semantic Content, Linguistics and Philosophy 23: 435–473.

Grice, Herbert Paul 1978. Further notes on logic and conversation. In: P. Cole (Ed.), Syntax and Semantics 9: Pragmatics. New York: Academic Press, 113-27. Now in Grice 1989, 41-57.

Grice, Herbert Paul 1989. Studies in the Way of Words. Cambridge, MA: Harvard University Press.

Hom, Christopher 2008. The semantics of racial epithets. Journal of Philosophy, 105, 416–40.

Hornsby, Jennifer 2001. Meaning and uselessness: how to think about derogatory words". In: P. French and H. Wettstein (Eds.), Midwest Studies in Philosophy, XXV, 128–41.

Hornsby, Jennifer and Langton, Rae 1998. Free Speech and Illocution, Journal of Legal Theory 4, 21-37.

Jeshion, Robin 2011 "Dehumanizing Slurs" Presented at the 2011 Society for Exact Philosophy meeting, Winnipeg, Manitoba (ms.)

Kaplan, David 1999. The Meaning of ouch and oops: explorations in the theory of meaning as use, ms, UCLA.

Kennedy, Randall 2003. Nigger: The Strange Career of a Troublesome Word. New York: Vintage.

Kissine, Mikhail 2013. Speech act classifications. In Sbisà and Turner 2013, 173-201.

Langton, Rae 1993. Speech Acts and Unspeakable Acts, Philosophy and Public Affairs, 22, 293-330. Now in Langton 2009, 25-63.

Langton, Rae 2009. Sexual Solipsism: Philosophical Essays on Pornography and Objectification, Oxford: Oxford University Press.

Langton, Rae 2012. Beyond Belief: Pragmatics in Hate Speech and Pornography. In McGowan and Maitra (Eds.) What Speech Does, Oxford: Oxford University Press.

Langton, Rae, Haslanger Sally and Anderson Luvell 2012. Language and Race. In Gillian Russell and Delia Graff Fara (Eds.) Routledge Companion to the Philosophy of Language, Routledge, 753-67.

Langton, Rae and West C. 1999. Scorekeeping in a Pornographic Language Game, Australasian Journal of Philosophy, 77, 3, 303-319. Now in Langton 2009, 173-195.

MacKinnon, Catharine 1987. Feminism Unmodified: Discourses on Life and Law, Cambridge (Mass.), Harvard University Press.

Potts, Christopher 2007. The centrality of expressive indexes. Reply to commentaries. Theoretical Linguistics, 33, 2, 255-68.

Potts, Christopher 2008. The pragmatics of conventional implicature and expressive content. In: C. Maienborn and P. Portner (Eds.), Semantics: An International Handbook of Natural Language Meaning. Berlin: Mouton de Gruyter.

Predelli, Stefano 2010. From the expressive to the derogatory: on the semantic role for non-truth-conditional meaning. In: S. A. Sawyer (Ed.), New Waves in Philosophy of Language. New York: Palgrave-MacMillan, 164-85.

Predelli, Stefano 2013. Meaning without Truth. Oxford: Oxford University Press.

Richard, Mark 2008. When Truth Gives Out. Oxford: Oxford University Press.

Sbisà, Marina 2001. Illocutionary Force and Degrees of Strength in Language Use. Journal of Pragmatics 33, 1791-1814.

Sbisà, Marina 2013. Locution, Illocution, Perlocution. In Sbisà and Turner 2013, 25-75.

Sbisà, Marina and Ken Turner (Eds.) 2013. Pragmatics of Speech Actions, Handbooks of Pragmatics, Vol. 2, Berlin, Mouton de Gruyter.

Searle, John 1979. Expression and Meaning. Studies in the theory of speech acts. Cambridge, Cambridge University Press.

Stenius, Erik 1967. Mood and Language-Game, Synthese 17, 254–274.

Walton, Kendall 1990. Mimesis as Make-believe: On the Foundations of the Representational Arts. Harvard University Press, Cambridge, MA.

Williamson, Timothy 2009. Reference, inference, and the semantics of pejoratives. In J. Almog and P. Leonardi (Eds.), The Philosophy of David Kaplan. New York: Oxford University Press, 137–58.

30

Knowledge Attribution, Warranted Assertability Manoeuvre and the Maxim of Relation

JACQUES-HENRI VOLLET

Abstract According to the traditional account of knowledge, what turns a true belief into knowledge is only a matter of truth-relevant factors (evidence, reliability, etc.), and these epistemic standards do not vary across contexts. Still, our knowledge attributions seem to vary depending on practical factors. Some philosophers account for this variation by a shift in the warranted assertability conditions. The main way of fleshing out this proposal is based on the idea of conversational implicature generated by the maxim of relation (Rysiew 2001, 2005, Brown 2006). In this paper, I argue that this is not promising. The proposition that is supposed to be implicated in such cases concerns whether the subject is in a position to eliminate a salient alternative that is not knowledge-destroying. I will claim that in such contexts, this consideration is not more relevant than the question whether the subject knows, even if the stakes are high.

1. Introduction

The bank cases

Suppose that two subjects truly believe that p, and are in the same epistemic position with respect to p. According to the traditional account of knowledge, both of these subjects know or do not know. The traditional account of knowledge is intellectualist. What turns a true belief into knowledge is only a matter of truth-relevant factors (evidence, reliability, etc.). It is also invariantist. The epistemic standards that a subject must meet in order to count as knowing do not vary with the context.

Against this traditional view, some philosophers have argued that the truth-value of knowledge attributions (utterances of sentences of the form "S knows that p") is sensitive to practical factors. There are two main ways of understanding this idea. For epistemic contextualists, "know" is a context-sensitive term. What is semantically expressed by an utterance of this word is determined by the context of utterance, and this context includes practical factors.[1] For subject-sensitive invariantists, there is a pragmatic condition on knowledge that must be satisfied in the subject's practical environment.[2] For instance, some philosophers have put forward the following condition: If a subject S knows that p, then it should not be a problem for S to act as if p.[3]

The idea that the truth-value of knowledge attributions is sensitive to practical factors rests in part upon pairs of cases that show that our ordinary knowledge attributions vary according to the stakes and the possibilities of error that are mentioned. One of these pairs of cases is the following DeRose (2009):

> Bank Case A (Low Stakes). My wife and I are driving home on a Friday afternoon. We plan to stop at the bank on the way home to deposit our paychecks. But as we drive past the bank, we notice that the lines inside are very long, as they often are on Friday afternoons. Although we generally like to deposit our paychecks as soon as possible, it is not especially important in this case that they be deposited right away, so I suggest that we drive straight home and deposit our paychecks on Saturday morning. My wife says, "Maybe the bank won't be open tomorrow. Lots of banks are

[1] Cohen (1988), Lewis (1996), DeRose (2009).
[2] Hawthorne (2004), Stanley (2005), Hawthorne & Stanley (2008), Fantl & McGrath (2009).
[3] Fantl & McGrath (2002, 2007).

> closed on Saturdays." I reply, "No, I know it'll be open. I was just there two weeks ago on Saturday. It's open until noon."
>
> Bank Case B (High Stakes). My wife and I drive past the bank on a Friday afternoon, as in Case A, and notice the long lines. I again suggest that we deposit our paychecks on Saturday morning, explaining that I was at the bank on Saturday morning only two weeks ago and discovered that it was open until noon. But in this case, we have just written a very large and important check. If our paychecks are not deposited into our checking account before Monday morning, the important check we wrote will bounce, leaving us in a very bad situation. And, of course, the bank is not open on Sunday. My wife reminds me of these facts. She then says, "Banks do change their hours. Do you know that the bank will be open tomorrow?" Remaining as confident as I was before that the bank will be open then, still, I reply, "Well, no. I'd better go in and make sure."

In case A, Keith says that he knows, which seems true. In case B, he says that he does not know, which also seems true. The only difference between the two cases is that the stakes are high in case B (a large cheque has been written) and a different possibility of error has been raised (banks do change their hours).

At first glance, this is a problem for the intellectualist and invariantist account. Indeed, according to the intellectualist and invariantist view, since Keith is in the same epistemic position in both cases and the epistemic standards do not shift, he knows in case A if and only if he knows in case B. Most intellectualists are also non sceptics, and hence agree that Keith knows in case A.[4]

Thus, according to the intellectualist and invariantist view, it is true that Keith knows in case A and in case B. However, what Keith says in case B, namely that he does not know, seems true. If so, the intellectualist and invariantist account is in contradiction with our intuitions.

Warranted assertability manoeuvres

The main intellectualist and invariantist reply consists in questioning the claim that what Keith says in case B is true. According to this line of thought, in case

[4]Hereafter, I will put aside the sceptical approach to the problem.

B it is right for Keith to say that he does not know, and wrong for him to say that he knows. However, it can be right to say something false.

This line of argument is based on a distinction between the semantic and pragmatic content of an utterance. Whether or not the utterance of a sentence is right depends on two things: the truth-value of the proposition that is the semantic content of the sentence (the literal meaning of the sentence), and the proposition that is communicated by the utterance of the sentence.

Thanks to this distinction, the invariantist intellectualist can make a so-called warranted assertability manoeuvre in order to explain what happens in case B. It is true that Keith knows in case B. However, it is wrong for him to utter the sentence that has this proposition as semantic content in this case. Indeed, given the conversational context of case B, this utterance would communicate a proposition that is false. Furthermore, it is right in this context to say something false since this communicates something true.

To defend this warranted assertability manoeuvre, one needs to show that there is a general rule of conversation which is operative in case B, which explains that saying a truth ("I know") in this context prompts an implicature that is wrong in terms of conversational rationality, and that saying something false ("I do not know") prompts an implicature that is right in terms of conversational rationality. Indeed, the manoeuvre is *ad hoc* if there is no such a general rule at play.[5]

The most influential account along these lines is Rysiew's (2001, 2005). It has been modified by Brown (2006). That is why I will mainly focus on these views. The common basic idea is that what Keith says is right due to a conversational implicature based on the general maxim of Relation (be relevant!). According to this view, on the assumption that Keith makes a relevant speech act, he addresses the worry raised by his wife. Therefore, it is natural to interpret what he says in light of this worry in the following way. In uttering "I do not know that p", Keith implicates that he is not in a position to eliminate his wife's salient alternative, which is true. By contrast, in saying the literal truth, that he knows, he would implicate that he is in a position to eliminate this alternative, which is false.

Plan

I propose a new argument against this strategy, based on the idea that the rule of Relation cannot prompt the pragmatic implication required to explain the

[5] DeRose (2009, 88–9).

intuition that Keith says something right.

My plan will be as follows. I will first expose Rysiew's argument in further details. This will allow me to question one of his presuppositions: that the knowledge attribution or denial must *merely* take the salient alternative into account. By contrast, I will show that in order to be fully rational in conversational terms, Keith must take the general goal of the conversation into account: to decide whether or not he must come back the following day to deposit his pay cheque. Taking this general goal into consideration leads to oppose to the idea that uttering "I do not know that p" might be a relevant speech act in this context due to an implicature regarding the salient alternative. In this context, I will argue, the consideration of a salient possibilities of error is not conversationally *more* relevant than the question whether the subject knows.

It might be replied to my argument that high stakes can explain why it is relevant to consider this salient alternative. This point has been stressed by Brown (2006). I will argue that this reply fails. On the assumption that it is not common knowledge that Keith knows, the question whether he knows he still more relevant.

2. Rysiew's manoeuvre

Rysiew's warranted assertability manoeuvre

I will first present the background assumptions of Rysiew's manoeuvre. Then I will show how they are used to explain our intuitions by claiming that in Bank Case B, there is a true proposition that is pragmatically communicated by Keith's utterance. Then, I will put forward objections.

Background

Rysiew rests upon the idea that in general speakers do not make the distinction between (or are not interested in) the semantic (or literal) meaning and the pragmatic meaning. The semantic meaning of an utterance is the proposition that is literally expressed by this utterance. The pragmatic meaning is the proposition that is imparted, or communicated, by the act of uttering the sentence. For instance, let us suppose that John has three children. If I say "John has two children", what I say is literally true, but misleading. Indeed, due to the rule of Quantity (assert the stronger!), what is pragmatically communicated is that John only has two children, and therefore not three.

According to Rysiew, if there is a confusion between the semantic and pragmatic meaning, one can explain our intuitions about the alleged variation in truth-value in the bank cases. In case A, what Keith says is literally true, and what is pragmatically communicated is identical to what is literally said. In case B, what Keith says is literally false, but what is pragmatically communicated is different, and true. Therefore, the intuition that it is true that Keith does not know in case B relies on this confusion between the semantic and pragmatic meaning of Keith's utterance.[6]

Now, what does a knowledge attribution literally express? Rysiew adopts a relevant alternative theory about knowledge, even though his general view is compatible with other accounts. According to such a view, an utterance of "S knows that p" literally expresses the proposition that S has a true belief that p, and that S's epistemic position is good enough to eliminate all the alternatives to p that are relevant. An alternative is relevant in this epistemic sense if and only if it is knowledge-destroying.[7]

A possibility of error is knowledge-destroying if its likelihood is above a certain threshold. This threshold does not vary depending on conversational and practical factors (such as what is salient in the conversation, what is at stake, etc.) but is set independently. That is why his account is intellectualist and invariantist. The degree of likelihood for a possibility of error is set by what a normal human would think of it in the circumstances in question. Therefore, the set of knowledge-destroying alternatives for a proposition does not vary across contexts.

Rysiew distinguishes these knowledge-destroying possibilities from possibilities of error that are merely salient. A possibility of error is salient if it has been raised in the conversational context. A possibility of error is salient and not knowledge-destroying if and only if it has been raised in the conversational context but does not prevent a subject from knowing, even if he is not in a position to eliminate it. The set of salient alternatives is a variable set.

As regards pragmatic meaning, it is standardly assumed that what an utterance pragmatically imparts depends on the conversational context. Let us suppose that there is a feature, X, that can roughly specify the conversational context, and that is determined by the goal or direction of the conversation, what is salient, etc.[8] Whether or not a speech act is right in conversational

[6]The idea that speakers are not always aware of, or interested in, the distinction between semantic content and pragmatic implicatures is disputable, though I will put aside this question.

[7]Rysiew uses the expression "relevant alternatives", but to avoid ambiguities, I will use "knowledge-destroying alternatives".

[8]One can think of X as the "conversational score". See Lewis (1979).

terms depends on X. If conversational participants are governed by a cooperative principle, they infer from what is literally said what best rationalizes others' speech acts given X. Therefore, what an utterance pragmatically imparts depends on X.

Given the variety of contexts, one can use an expression to communicate many different things. What is communicated in a given context depends on the semantic content of an utterance, along with the general rules of conversation. Then, according to Rysiew, it is natural to think that an utterance of "S knows" in some conversational contexts can pragmatically impart the proposition that S is in a good enough epistemic position to satify the epistemic standards for X. Indeed, in the semantic content of "S knows", there is the idea that S is in a good enough epistemic position. And if in some contexts this is what best rationalizes an utterance of "S knows", then this is part of what will be pragmatically communicated in this context.

The underlying idea is that X, the goal of the conversation, the interests of the speakers, etc., can determine some relevant epistemic standards in a context C. According to this view, an utterance of "S knows" in C can pragmatically impart that S meets these epistemic standards.

Bearing all these ideas in mind, it is possible to reconstruct Rysiew's warranted assertability manoeuvre.

Rysiew's account

Let me present Rysiew's account in detail. I will first explain what a knowledge attribution is supposed to communicate according to Rysiew, in a context where there are salient alternatives, if this attribution satisfies Grice's maxim of Relation (be relevant!). I will show that there are contentious assumptions in this picture. Then, I will focus on my main point.

An utterance is conversationally relevant in a context only if it takes the ingredients of the context into account. Thus, one can agree with Rysiew that a knowledge attribution in a context where some alternatives have been raised is conversationally relevant only if it takes the alternatives that have been raised in this context into account. Moreover, since a knowledge attribution communicates something about the subject's epistemic position, it is also very plausible that it can communicate something concerning the assessment of the subject's epistemic position with respect to the epistemic standards of the context. Rysiew uses both of these ideas to claim that a knowledge attribution in a context where there are salient alternatives can be conversationally

relevant if it communicates something concerning the assessment of the subject's epistemic position with respect to the salient alternatives.

In other words, if you apply the general idea that a conversationally relevant speech act must take the conversational context into account, to a case where the speech act is a knowledge attribution and the conversational context includes salient possibilities of error, you can predict that the knowledge attribution communicates something relative to these salient possibilities of error. And given the semantic content of a knowledge attribution sentence, it is natural to think that what is communicated is something related to the subject's epistemic position relative to these salient possibilities of error.

However, Rysiew goes further. He claims that it is natural to think that a knowledge attribution in this kind of context can be conversationally relevant in communicating that the subject's epistemic position with respect to the operative standards in the context is "good enough".[9] How does Rysiew support this claim? Here is the explanation:

> For it is only if speakers are understood to be intending to communicate information about how the subject fares vis-à-vis the contextually operative standard(s) — hence, i.a., about whether the subject can or cannot rule out any contextually salient alternatives — that they can be seen as striving to be maximally relevantly informative (hence, as conforming to CP) (Rysiew 2005, 48).[10].

It is not clear how the argument works. There is a difference between communicating information about the value of a subject's epistemic position with respect to an operative epistemic standard in a context (in Keith's case, a standard which is a function of salient doubts), and communicating that the subject's epistemic position is "good enough" with respect to the operative epistemic standards.

It may seem that Rysiew goes from one idea to the other in assuming that it is a semantic fact that if S knows that p, then S's epistemic position with respect to p is good enough. The idea would be that when you utter "S knows that p", you can communicate an assessment of S's epistemic position with

[9] See for instance Rysiew (2005, 48): "...it's because speakers strive to conform, and are known to so strive, to the maxim of relation ("Be relevant") that, in uttering a sentence of the form, 'S knows that p', the speaker is naturally taken to intend/mean that S's epistemic position with respect to p is 'good enough' given the epistemic standards that are operative in the context in question."

[10] See also (Rysiew 2001, 491–492).

respect to X, as well as the idea that the subject's epistemic position is good enough with respect to X.

Let us note that this is not obvious, in particular if one adopts Rysiew's fallibilist framework. If non sceptical fallibilist invariantism is true, the semantic fact is that if S knows that p, then S's epistemic position with respect to p is good enough *to a certain extent*. Therefore, there is no clear reason to think that an utterance of "S knows that p" would impart the idea that the subject's epistemic position is *good enough* to satisfy the epistemic standards of the context. We will come back to this point below.

Let us also note that the use of the rule of Relation is based on the question of the specific relevance of an utterance. The question is not to recognize the falsity of what is literally said (in which case the conversational rule at work would be the rule of Quality). Therefore, the maxim of Relation can be put at work only if one assumes that the hearer is led to seek for a different (and pragmatic) meaning because the semantic meaning of the uttered sentence lacks specific relevance.

Then, in bank case B, the rule of Relation is operative only if the knowledge attribution or denial in this context lacks specific relevance. If one can show that the denial in case B does not lack specific relevance in this context, one can doubt that a warranted assertability manoeuvre using the rule of relevance is promising. This worry has been raised in particular by DeRose (2009). We will come back to it below.

For the time being, let us apply Rysiew's general idea to Keith's case. When Keith utters "I do not know", the semantic content of the sentence uttered in this context lacks specific relevance. (This would also have been the case if Keith had uttered "I know that the bank is open"). Indeed, this literally means: "I cannot eliminate the knowledge-destroying alternatives", and in the context, the question is about the salient alternative raised by Keith's wife. If the conversational context does not concern knowledge-destroying alternatives, but mere salient alternatives, one needs to find a meaning that is different from the literal meaning, so as to make the utterance specifically relevant (so as to make the speech act in conformity with the maxim of relevance).[11] This different meaning is the pragmatic meaning, namely that Keith cannot eliminate the salient alternative.

[11] See for instance Rysiew (2001, 491–492): "Whether or not what he [Keith] says is strictly speaking true, if the speaker *didn't* think that *S*'s epistemic position were good enough *in the relevant sense of "good enough"*, whatever that is, he would not say "*S* knows that *p*"; and regardless of the truth-value of the sentence itself, he wouldn't say "*S* doesn't know that *p*" unless he thought there were salient no-*p* possibilities that the speaker could not rule out".

Given this analysis, it is possible to rephrase Rysiew's account using the following claims:

(1) In general, an utterance of "S knows that p" or "S does not know that p" is specifically relevant in a conversational context only if it means something concerning the assessment of the strength of S's epistemic position with respect to the standards required by X, where X is set by the conversational context C.

(2) In general, an utterance of "S knows that p" in C, if it is specifically relevant, means that S's epistemic position is *good enough* with respect to the operative standards in C.

(3) In general, an utterance of "S does not know that p" in C, if it is specifically relevant, means that S's epistemic position is *not good enough* with respect to the operative standards in C.

(4) When C is the context of case B, X is the salient alternative.

Suppose also that:

(5) Keith is not in a position to eliminate the salient alternative in case B.

Then, by (1) - (5):

(6) In case B, Keith's utterance of "I know that p", if it means something that is specifically relevant with respect to the assessment of Keith's epistemic position concerning the salient alternative, would mean something that is false (namely that Keith can eliminate the salient alternative)

(7) In case B, Keith's utterance of "I do not know that p", if it means something that is specifically relevant with respect to the assessment of Keith's epistemic position concerning the salient alternative, means something that is true (namely that Keith cannot eliminate the salient alternative)

Therefore, supposing that it is conversationally wrong to communicate something false and conversationally right to communicate something true:

(8) It is conversationally wrong for Keith to say "I know that p" in case B.

(9) It is conversationally right for Keith to say "I do not know that p" in case B.

Note that this account is supposed not only to explain why it would be improper for Keith to say that he knows, namely that this would prompt a false implicature, but also why it is correct to say that he does not know: this utterance takes the context into account, and communicates something true.

First possible objections

I will now consider the main objection that has been proposed so far, and that has been put forward by DeRose. It states that the semantic content of the knowledge denial is relevant in case B, and hence the hearer cannot be expected to infer a pragmatic meaning from a lack of relevance. In other words (1) can be true even though there is no pragmatic implicature at play in case B. I will argue that Rysiew's account can escape this objection. Then I will consider a different objection that seems more compelling to me. However, my main argument will be more radical.

According to DeRose (2009, 118–124), the semantic content of an utterance of "S does not know that p" does not lack specific relevance in the context of case B. Indeed, if this utterance literally means that S is not in a position to eliminate the knowledge-destroying alternatives, this directly implies that S is not in a position to eliminate the salient alternatives. It may seem that one needs to be in a better epistemic position to eliminate all the salient alternatives than to eliminate the knowledge-destroying ones only. Therefore, not only does this utterance seem specifically relevant, but it also communicates further information: that S is not in a position to eliminate the knowledge-destroying alternatives. Now, suppose that it is actually false that S does not know. Then, this utterance is actually misleading. Since the hearer cannot appeal to a lack of relevance for working out the implicature (since this utterance does not lack any specific relevance), and since the proposition expressed is false, the utterance violates the maxim of Quality. Therefore, the invariantist intellectualist cannot in this way explain why it seems correct.

A possible reply would be to say that it is common knowledge that S does not know. However, this is not a manoeuvre that uses the maxim of Relation, but the maxim of Quality. In addition, it is possible to figure out cases where this common knowledge is absent (DeRose 2009, 122–123).

Still, there is another possible rejoinder available to Rysiew. Indeed, the idea that the semantic content of an utterance is relevant is ambiguous. This can mean that the semantic content helps to reach the proposition that is meant to be communicated, or that it is sufficient to reach the proposition that is meant to be communicated. It does not seem that the knowledge denial in

case B is relevant in the latter sense. Let us assume that the important question is whether Keith can eliminate a salient alternative, and suppose that the semantic content of this knowledge denial does not lack relevance in the sense that it can "settle the question" concerning this alternative (DeRose 2009, 122). It might be argued that the way this semantic content can settle the question supposes a kind of inference that is led to the hearer. This inference may be of the following form: "If Keith does not know that the bank is open, then he is not in a good enough epistemic position to eliminate this salient possibility of error". Clearly, if the hearer is not able to make this inference, he can be puzzled by the answer (given that the most relevant issue concerns a salient possibility of error). Therefore, the knowledge denial in case B is not sufficient to communicate what is meant to be communicated, and in this sense it is not fully relevant. That is why even if it can be used to settle the question, it can lack relevance.

To illustrate, let us take DeRose's own analogy, and suppose that we are in a context where "tall" means "at least 6 feet tall". Suppose that I am interested in whether Sally is tall. If I ask you "Is Sally tall?" and you answer "She is less than 5.5 feet tall", we can agree that this settles the question in a negative way. In this sense your answer is not irrelevant. However, if I am not able to infer from the fact that Sally is less than 5.5 feet tall, the fact that she is less than 6 feet tall, then I cannot settle the issue whether she is tall. Then, it can be easily accepted that you expect me to make this inference from what you said. So, it can be granted that by saying that, you expect me (the hearer) to reach the relevant conclusion, that is, you aim at communicating this conclusion. (Your intention is to make me, the hearer, believe the conclusion). Therefore, it seems that what you intent to communicate is this conclusion, rather than just the semantic content of the utterance. In that sense, it can be granted that the semantic content is in some way relevant to reach the conclusion, but what is intended to be communicated is the conclusion itself. In this sense, the semantic content lacks relevance, which explains why I make the inference and reach the conclusion.

Nevertheless, pursuing DeRose's line of thought, one might still wonder whether this explanation is satisfying. Indeed, the main problem is that, according to Rysiew, it is false that Keith knows in case B. And one might think that it is incorrect to say something false to make someone reach the right conclusion. To illustrate, let us follow DeRose and modify Sally's case. Suppose that she is more than 5.5 feet tall, but less than 6 feet tall. Is it right to falsely say "She is less than 5.5 feet tall" to make the hearer reach the right conclusion that she is not tall? I must say that this does not seem wrong to me. There are

many cases in real life where one tells white lies, and this is acceptable insofar as it is efficient.[12] As a result, it does not seem to me that DeRose's criticism is sufficient to undermine Rysiew's attempt.

However, there are still a problematic assumptions in Rysiew's view at this stage, regarding (2) and (3) above. Recall that it is not clear why the semantic meaning of a knowledge attribution would drive the hearer toward the idea that he is in a good enough epistemic position, rather than an insufficient epistemic position, to meet the epistemic standards of the context. This assumption is needed if Rysiew wants to explain why a knowledge attribution would be inappropriate, while true, in case B. Suppose for instance that Keith says "I know", and by this, literally means that his epistemic position is good enough to eliminate the knowledge-destroying alternative. Contrary to what Rysiew claims, this could drive the hearer to the right piece of information. Indeed, this may pragmatically impart that Keith is not in a position to eliminate the salient alternative. If Keith does not say "I know for sure", but merely "I know", given the rule of Quantity (assert the stronger!), the hearer may infer that Keith's epistemic position is not good enough to eliminate the salient but non knowledge-destroying alternative. The underlying idea is that one cannot just suppose, like Rysiew does, that the pragmatic meaning of an utterance such as "S knows that p" is in general that S's epistemic position with respect to p is good enough for X, where X is fixed by the context.

Rysiew seems to defend his view against this possible objection with the idea that one can base a pragmatic inference upon a mere part of the semantic content. Even if one does not have a clear view of what the truth-conditions of a knowledge attribution are (i.e. being in a good enough epistemic position to eliminate the knowledge-destroying alternatives), one can base the inference upon a part of these truth-conditions (i.e. being in a good enough epistemic position with respect to the proposition).

Thus, Rysiew writes:

> All you need is to see that if I didn't think that S did so measure up, it would be odd, indeed uncooperative, of me to say, "S knows that p"; for whatever exactly knowledge is, we know that S's knowing

[12] See Fantl & McGrath (2009, 41–42) for discussion. They insist that it is difficult to explain why a white lie is acceptable in case B. But it may seem that a white lie is efficent. They also raise another worry concerning the first-person perspective. According to them, in case B Keith will be willing to say to himself that he does not know. This cannot be explained easily by a WAM, even assuming that one easily confuses the semantic and pragmatic meanings. However, given Rysiew's fallibilist framework, one might argue that since Keith knows, he will rather be willing to say to himself: "I know that the bank is open but I should not take the risk".

> that p entails that S is in a good epistemic position with respect to p. "And why", the hearer may well ask, "would he say something that means [semantically implies] that if he didn't in fact think that S measured up to the epistemic standards that are in play, and so was able to put to rest any doubts or rule out any not-p possibilities that had just been raised?" (Rysiew 2005, 49).

There are two possible replies to this defence. On the one hand, if it it is true that "S knows that p" semantically implies that S's epistemic position is good enough in an absolute sense, then it is true that if S knows that p then S's epistemic position is good enough no matter the context. Under this interpretation, uttering "S knows that p" will semantically express that S's epistemic position is good enough for any X, where X is determined by the context. However, this cannot be what Rysiew means, because then Keith would be in a position to eliminate his wife's salient alternatives in case B.

On the other hand if Rysiew wants to mean that "S knows that p" semantically implies that S's epistemic position with respect to p is good enough *to a certain extent*, then he cannot escape the previous objection. If in a context the standards are higher than those required for knowledge, to say that someone meets these knowledge-level standards might well communicate that he does not meet those required for more than knowledge.

More generally, it does not seem that "being good enough" is a semantic part of "being good enough to a certain extent". Indeed, it does not follow from "X is good enough to a certain extent" that "X is good enough". Therefore, it does not seem that it is possible to claim that since "S knows that p" semantically imparts that "S is in a good enough epistemic position with respect to p to know p" then "S knows that p" semantically imparts that "S is in a good enough epistemic position with respect to p". As a result, (2) does not seem true.

It can also be argued in a similar way that (3) may be false. By uttering "I do not know that p" in a context with higher epistemic standards, I may convey the idea that I do not *merely* know that p. This case is similar to the one where one says that one does not believe p in order to convey that one knows p.

To recap, the idea that a knowledge attribution (or denial) lack relevance in case B is not sufficient to explain why the hearer is driven to the idea that the speaker is (or is not) in a good enough epistemic position with respect to the salient alternatives. This is one aspect in which Rysiew's account may seem insufficient.

However, my main argument will focus on (4).

My strategy against Rysiew's warranted assertability manoeuvre

In case B, the salient possibility of error is conversationally relevant because it has been raised in the conversation. One can explain why it has been raised by the fact that the stakes are high. Indeed, following Rysiew, one might think that when the stakes are high, it is natural to think about possibilities of error. Nevertheless, Rysiew's WAM does not essentialy appeal to the stakes in order to explain why the knowledge denial is warranted in case B. Indeed, there might be cases where the same manoeuvre is available, even though the possibility of error is raised in a different way.[13] Brown (2006) has put forward a slightly different view: the presence of high stakes in case B can make a difference as to the extent to which the salient possibility of error is conversationally relevant.

In what follows, I will distinguish the two approach. First I will focus on Rysiew's account, putting aside the issue of the stakes. I will come back to the second proposal in section 4.

I will develop my criticism of Rysiew with the idea that the conversational relevance of a knowledge attribution in a context must take into account two features:

1. One must take the salient possibilities of error in C
2. One must take these salient possibilities of error in function of the general goal of the conversation in C.[14]

In addition, it seems clear that an implicature cannot violate a conversational rule since it results from the need to conform to these rules. Hence, if a proposition is not fully relevant in C, that cannot be an implicature of what is said in C.

To my mind, the problem in Rysiew's explanation is that it assumes that the only way to take the salient alternative into account consists in assessing Keith's epistemic position with respect to it. However, with regard to the

[13] For instance, Hazlett (2009) uses this manoeuvre to account for knowledge denials in sceptical contexts where there are no stakes.

[14] See Grice (quoted by Rysiew 2005, 50) : "I expect a partner's contribution to be appropriate to the immediate needs at each stage of the transaction" (Grice 1989, 28). This does not imply to deny the general goal of the conversation.

general goal of the conversation, if it is true that Keith knows, it does not seem relevant for him to say that he does not know. Here is my argument.

Let us suppose with Rysiew that if Keith says that he does not know, this pragmatically prompts the true proposition q, where q is that Keith cannot eliminate his wife's salient alternative. Let us also suppose that if Keith says that he knows, this pragmatically prompts the false proposition not-q. Rysiew deduces that since q is true and not-q is false, Keith must utter what prompts q.

However, I will argue that the question whether q or not-q lacks of relevance in case B. In a rational conversation, a proposition that lacks of relevance in the context cannot be pragmatically implicated by an utterance in this context. As a result, neither q nor not-q can be implicatures in this context. If so, whether or not Keith is in a good enough epistemic position to eliminate his wife's alternative cannot explain why Keith's utterance "I do not know that the bank is open" is right, if it is true that he knows.

My main task will be to defend the idea according to which the question whether Keith is in a position to eliminate the salient alternative of his wife lacks relevance given the general course of the conversation. I will use two premises. First, I will claim that the case B is a case where the goal is to take a rational decision. Second, I will claim that in a conversational context where the goal is to take a rational decision, the question whether one is in a position to eliminate a salient but not knowledge-destroying alternative lacks relevance.

It seems to me that the second premise is the most subject to criticism. So I will mainly defend it. I will do it in the following way. First I will show that in general, when the goal of a conversation is to know whether a subject knows, the question whether one can eliminate a salient but not knowyledge-destroying alternative is irrelevant. Second, I will argue that if the question has to do with the rationality of a subject acting on a proposition, the question whether the subject knows this proposition is more relevant than the question whether he can eliminate a not knowledge-destroying alternative.

3. Ways of taking the salient possibilities of error that are not knowledge-destroying into account

Context where the question is whether a subject knows a proposition

In a context where what matters is whether or not a subject knows, it is not appropriate to say that he does not know, if he knows, even if this subject is facing a salient possibility of error that is not knowledge-destroying. Indeed, in this context, the question does not turn around the possibility for the subject to eliminate alternatives that are not knowledge-destroying, but concerns only possibility that are knowledge-destroying.

It is clear that it would be wrong for a speaker to say that the subject does not know in such cases semantically and pragmatically. From the semantic point of view, the utterance "S does not know" is literally false. From the pragmatic point of view, it seems to communicate something false: that the salient possibility of error is knowledge-destroying. Therefore, the main way of interpreting such an utterance in this context would be to think that the speaker is mistaken. He thinks that a not knowledge-destroying possibility of error is knowledge-destroying, which is not. So, this utterance is incorrect.

Thus, suppose that in such contexts someone raise a possibility of error which is not knowledge-destroying. It is clear that it is not appropriate to take it into account by saying "S does not know that p". Indeed, given the purpose of the conversation, raising this possibility was not relevant in the first place, and it would be irrational (in conversational terms) to take this possibility into account by communicating, by the utterance "S does not know", that S cannot eliminate it. Indeed, this would communicate that this possibility is knowledge-destroying and that it was right, in the first place, to consider this possibility in order to contemplate whether S knows.

Therefore, uttering "I do not know that p" in front of a salient possibility of error that is not knowledge-destroying, in a context where the point of the conversation is to consider whether I know, seems equivalent to utter "I do not know that p" in front of an epistemically and conversationally unreasonable alternative.

The upshot is simple. In a context where the question is to know whether S knows that p, it is not appropriate to take not-eliminable salient possibilities of errors that are not knowledge-destroying into account by saying "S does not know that p". On the contrary, it is appropriate to take them into account by rejecting them as not relevant.

There are at least two ways of doing it. Rysiew thinks that concessive knowledge attributions (sentences such as "I know that p but it might be that q" (where q obviously entails not-p)) are not contradictory. But then, this would be an appropriate answer. Alternatively, one can directly reject this alternative as irrelevant.

Context where the question is to decide what to do

I have argued that in a context where the question is whether a subject knows, not knowledge-destroying possibilities of error are not conversationally relevant. If one can also show that in the bank case B, the question whether Keith knows is more relevant than the question whether he can eliminate not knowledge-destroying alternatives, then the result will be that there is no implicature to the effect that Keith cannot eliminate his wife's possibility of error (assuming that it is not knowledge-destroying).

I will argue that in general, when the goal of the conversation is to decide whether one should act on a proposition, the question whether one knows this proposition is more conversationally relevant than the question whether one can eliminate a possibility of error that is not knowledge-destroying. (This is a reason to think that in raising a possibility of error, Keith's wife challenges Keith's knowledge.)

First of all, there is no denying that knowledge claims are conversationlly relevant when the purpose of the conversation is to decide what to do. For instance, discussing a case similar to the bank cases, Reed (2010, 232) writes:

> [Y]ou are trying to decide whether to check if the train stops in Foxboro because it is extremely important that you get there as quickly as possible. You have not yet decided whether it is rational to check if the train makes that stop. One of the relevant factors in your decision would presumably be an answer to the question, do you know the train will stop there?

Similarly, it is very plausible that in the bank cases, one factor that is relevant to decide what to do is an answer to the question: does Keith know that the bank is open on Saturdays? Now, the question is how relevant is this factor. Is an answer to this question more relevant than an answer to the following question: is Keith in a position to eliminate not knowledge-destroying alternatives?

If knowledge questions were sometimes not relevant, then one should expect cases where the relevance of a knowledge claim is appropriately chal-

lenged. However, it seems that in most cases, whether one knows the target proposition settle the issue in a proper way (I will claim that this is also true if the stakes are high). For instance suppose that it is right to come back tomorrow only if the bank is open. Suppose that I say that we should come back tomorrow because I know that the bank is open. One may questions whether I really know, but it does not seem appropriate to question the relevance of my contribution to the discussion. For instance, the following reply would be surprising: "Well, the question is not whether you know that the bank is open". It is highly plausible that whether one knows is always a relevant question in such a kind of case. On this assumption, it would not be surprising if Keith's wife challenged Keith's knowledge.

By contrast, it does not seem that in such contexts the question of eliminating not knowledge-destroying possibilies of error is relevant in the same way. Suppose for instance that I say to my wife: "I prefer to go to the bank tomorrow. I know it's open". Suppose that she answers: "But maybe the bank has changed its hours". It is not conversationally irrelevant for me to reply: "Of course, everything is possible, but do you really think that this will happen? What makes you think that?". I ask for a reason to consider this possibility of error, which shows that it is not obvious to me whether raising this possibility of error is relevant to the discussion. In other words, I challenge the relevance of what she has just said.

Thus, the main difference between the knowledge question and the not knowledge-destroying alternative question is the following: in general, the relevance of the first question cannot be easily challenged, while the relevance of the second one can be easily challenged. Does this show that the first question is more relevant? I think that it does, if one also notes that a challenge to a not knowledge-destroying possibility is appropriate precisely if this possibility of error is not obviously not knowledge-destroying nor obvisously knowledge-destroying.

To see this, let us compare our reactions to alternatives that are clearly not knowledge-destroying. Suppose for instance that my wife raises the possibility that Martians will destroy the bank tonight. An appropriate answer would be: "Come on! You know this won't happen". If it is obviously not knowledge-destroying, the alternative can be rejected outright. This suggests that it is appropriate to challenge the relevance of a salient possibility of error only if it might be knowledge-destroying.

It is then natural to think that challenging the relevance of an alternative aims at deciding whether or not this alternative is knowledge-destroying, or to what extent it might be knowledge-destroying. Indeed, asking for a rea-

son to think that this alternative might obtain is just asking whether it is knowledge-destroying. After all, if there is a good reason to think that an alternative to p might obtain, one's knolwedge that p is destroyed (unless one can eliminate it). And in cases where there is no good reason to think that this possibility might obtain, one's knowledge is not destroyed. These cases are precisely those in which these alternatives can be rejected as irrelevant for the purposes at hand.

To sum up, our appropriate reactions to alternatives that are raised in contexts where the question is to decide what to do seem in general governed by how these alternatives are supposed to be related to what we know. This shows that in these contexts, the conversational relevance of the question concerning possibilities of error is derived from the conversational relevance of the question concerning knowledge. As a result, the most relevant question in these contexts seems to be the knowledge question.

Applying the analysis to the bank cases

Let us apply this analysis to the bank cases. In the bank cases, what matters is to take a rational decision: either to deposit the pay cheque right now or to come back on Saturday. Even if Keith cannot eliminate his wife's possibility of error, it is also assumed that this salient possibility is not knowledge-destroying. Therefore, according to the line of thought that I have put forward, this possibility is not the most relevant to take the right decision. That does not necessarily mean that it is totally irrelevant. Although this alternative is not knowledge-destroying, it may be relevant in the sense that it is not obviously not knowledge-destroying. But in that case, Keith should challenge it. On the contrary, if one assumes that this possibility of error is obviously not-knowledge destroying, then Keith should reject it outright as irrelevant. As a result, since the fact that Keith cannot eliminate this possibility is not fully relevant (or is totally irrelevant) to the purposes at hand, Keith denial cannot pragmatically impart the proposition that he cannot eliminate this possibility.

Note that there is another independent reason to think that the case B is a case where the question is whether Keith knows. It seems clear that in raising this possibility of error, his wife is interested in whether he knows. She does not ask whether Keith can eliminate this possibility, but whether he knows that the bank is open. This suggests that she thinks that the possibility of error she has raised challenges Keith's knowledge. If so, it would be misleading for Keith to answer negatively just to communicate that he cannot elimi-

nate this alternative. Indeed, he would thereby impart that this alternative is knowledge-destroying, or would be mistaken about what his wife meant.

4. Brown's variation

One possible reply is to say that given the high stakes in case B, whether or not Keith can eliminate his wife's possibility of error is actually totally relevant to decide what to do.

I will argue that even if the stakes are high, this possibility of error is less relevant than whether Keith knows. Then, even if the stakes are high, on the assumption that it is not common knowledge that Keith knows, it is more plausible that Keith's wife raises this alternative to challenge Keith's knowledge.

Let us first develop the idea that this salient alternative has a particular relevance due to the practical situation. Brown (2006) has proposed to fill out Rysiew's view in order to explain why and how the fact the Keith's wife has mentioned a possibility of error can change the goal or the direction of the conversation, so that an utterance of "know" means "can eliminate the salient alternatives" and "not know" means "cannot eliminate the salient alternatives". According to her, the practical importance can explain this. Indeed, it seems that one can more easily resist to possibilities of error raised in contexts where it is not practically important not to be wrong, than to possibilities of error raised in contexts where it is practically important. And a relevant conversational ingredient in Keith's conversational context is the fact that Keith's wife has mentioned that a large cheque has been written.

If this analysis is correct, given that Keith's utterance is relevant, he takes this conversational aspect into account, and one must interpret his claim in light of this consideration. The fact that a large cheque has been written "makes it clear that what's relevant to the conversation is a very strong epistemic position" (Brown 2006). Therefore, our argument does not seem to work against this slightly modified WAM.

However, there are some possible troubles for this rejoinder. First, this leads to reject the idea that if you know that p, then your epistemic position with respect to p is good enough to act on p. Still, one might argue that this claim is false (Brown 2008).

Second, and more importantly, Brown's filling out rests on the idea that it is more difficult to resist the mentioning of not knowledge-destroying possibilities of error when it is important not to be wrong because when the stakes

are high, one needs to be in a strong epistemic position with respect to p to be rational to act on p.

One can grant this idea, but that does not imply that in case B Keith should impart something relatively to the salient alternative. Indeed, recall that it is assumed that it is not common knowledge that Keith knows. And it does not seem that if Keith can eliminate some specific not knowledge-destroying possibility in which not-p, he is in a stronger epistemic position than if he knows p. Then, even if a strong epistemic position is required, the more interesting issue in order to take the right decision is still whether he knows that the bank is open, rather than the question whether he is in a position to eliminate the possibility that the bank has changed its hours.

To put in another way, suppose that eliminating a specific not knowledge-destroying possibility of error in which not-p does not put you in a better epistemic position with respect to p than knowing p. Suppose also that the question is to take the right decision and the best epistemic position is required. In this case, what is conversationally relevant, given the general goal of the conversation, is to focus on what one knows rather than on one's epistemic position with respect to some specific far-fetched possibility of error.

I think that one can easily accept that the question whether one knows is more important. Suppose that p is relevant to your action, and you have a choice between two options: know whether p or eliminate a not knowledge-destroying possibility of error in which not-p. It seems clear that you will choose the first option. For instance, consider the following case:

> You are a policeman in a country where there is only one gun, and you are about to enter into a house to arrest some dangerous criminal. It is very important that the criminal is not armed. Before entering, you just have enough time to read one of the two following messages on your mobile: (a) is the criminal armed? or (b) does the criminal have a gun?

It seems clear that it is rational to choose (a) if you have to decide what to do. Indeed, suppose that you choose (b) and the answer is positive. Then it is not rational for you to go into the house. Suppose that you choose (b) and the answer is negative. Still, it is not rational for you to enter into the house because the criminal might be armed in a different way. An answer to (b) does not make any difference to what you should rationally do. On the contrary, suppose that you choose (a). If the answer is positive, then it is not rational for you to go into the house. But if the answer is negative, it does not seem problematic for you to enter.

One might object that it might not be rational to enter into the house, if one has not eliminated the possibility that the subject has a gun. Two remarks are in order. Fist it is important to note that even in high stakes situtions, saying that one knows the target proposition is in general considered as sufficient to settle the issue. In high stakes cases, one might be less willing to cite knowledge, but that does not show that citing knowledge does not settle the question.[15] Second, even if some might doubt that it is rational to enter into the house given that there is no answer to (b), it is clear that (a) is more relevant than (b).

Similarly in the bank cases. Suppose that Keith's wife does not know whether Keith knows. Does she raise this possibility of error to challenge Keith's knowledge, or Keith's ability to eliminate this salient alternative? It is implausible that she raises this possibility of error to challenge Keith's ability to eliminate it. Indeed, whether or not the bank has changed it hours is not fully relevant. Maybe the bank has not changed its hours, but actually it is closed on Saturdays except when Keith came the last time. Maybe it has changed its hours but it is still open tomorrow. On the contrary, if she raises this possibility to challenge Keith's knowledge, that seems fully relevant insofar as this possibility of error might be knowledge-destroying.

One can sum up by claiming that knowledge-destroying possibilities of error are more relevant to rational action than salient but not knowledge-destroying possibilities of error even when the stakes are high. Given that they are more relevant, if Keith knows, and on the assumption that it is not common knowledge that Keith knows, Keith should not impart that he cannot eliminate a less important possibility of error but should challenge his wife's alternative, or should say that he knows (if he knows), even if the stakes are very high.

5. Conclusion

The traditional invariantist and intellectualist account of knowledge is challenged by pairs of cases with asymmetrical stakes and different salient possibilities of error. These cases show that the correctness of knowledge attributions varies with the presence or absence of these practical factors. Still, it has been argued that these intuitions does not show that knowledge varies with practical factors, but only that the warranted assertability conditions of knowledge vary. Assertions, and knowledge attributions in particular, prompts

[15] See Fantl & McGrath (2007, 562).

conversational implicatures. However, this mechanism can be used to explain our intuition only if a general rule of conversation is operative in the context. The conversational rule that is in general put forward is the rule of relation. I have shown that this rule cannot be used to explain our intuitions about these cases.

6. References

Brown, J. (2006). Contextualism and warranted assertability manoeuvers. *Philosophical Studies*, 130:407–35.

Brown, J. (2008). Subject-sensitive invariantism and the knowledge norm for practical reasoning. *Noûs*, 42 (2):167–189.

Cohen, S. (1988). How to be a fallibilist. *Philosophical Perspectives*, 2:91–123.

DeRose, K. (2009). *The Case for Contextualism*. Oxford University Press.

Fantl, J. and McGrath, M. (2002). Evidence, pragmatic and justification. *The Philosophical Review*, 111:67–94.

Fantl, J. and McGrath, M. (2007). On pragmatic encroachment on epistemology. *Philosophy and Phenomenological Research*, 75:558–89.

Fantl, J. and McGrath, M. (2009). *Knowledge in an uncertain world*. Oxford University Press.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard Unviersity Press.

Hawthorne, J. (2004). *Knowledge and Lottery*. Oxford University Press.

Hawthorne, J. and Stanley, J. (2008). Knowledge and action. *Journal of Philosophy*, 105 (10):571–590.

Hazlett, A. (2009). Knowledge and conversation. *Philosophy and Phenomenological Research*, 78 (3):592–620.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8:339–59.

Lewis, D. (1996). Elusive knowledge. *Australian Journal of Philosophy*, 74:549–67.

Reed, B. (2010). A defense of stable invariantism. *Noûs*, 44 (2):224–244.

Rysiew, P. (2001). The context-sensitivity of knowledge attributions. *Noûs*, 35 (4):477–514.

Rysiew, P. (2005). Contesting contextualism. *Grazer Philosophische Studien*, 69 (1):51–70.

Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford University Press.

31

First person thought

FRANÇOIS RECANATI

1.

The first thing we can say about first person thoughts is that they are the sort of thought one may express by using the first person : 'I am hungry', 'I was born in Paris', 'I have crossed knees'. The second thing we can say is that they are *thoughts about the thinker of the thought*. Let us start from there, and see whether, building on these features, we can arrive at a proper characterisation of this class of thought.

Can we use the second feature – reflexivity – to define first person thoughts ? No, or at least not if we care about the first feature ; for a thought can be about the thinker of the thought *without* being being the sort of thought one would express by using the first person. One may have a thought about oneself even though one does not realize that the thought is about oneself (as when one sees oneself in the mirror, without realizing that one is self-seeing). Such a thought is not a first person thought even though, extensionally, it is about oneself.

In the other direction, however, the entailment holds : a thought that can be expressed by using the first person will be about oneself, as a matter of necessity. Let us suppose that the thinker holds a thought that is *not* about himself or herself. Could he or she express such a thought by using the first person ? Answer : No, because we use the first person to refer to ourselves. (That is a *linguistic fact* about the first person.) To be sure, it seems that we may use the first person to express a thought that is not about ourselves, in a situation in which we mistakenly *believe* that the thought is about ourselves. But let us take a closer look at such a situation. If I believe that the winner will

go to Tahiti, and I mistakenly believe that I am the winner, my thought 'The winner will go to Tahiti' will not be about myself but I will take it to be about myself. As a result, I will be in a position to use the first person in expressing my belief : 'I will go to Tahiti'. In such a case, however, there are *two* thoughts : the initial thought ('The winner will go to Tahiti') is *not* about myself, but the mistaken belief that I am the winner leads me to form another thought which *is* about myself : the thought that I will go to Tahiti. It is *that* thought — a thought about myself — which I express by using the first person.

The natural conclusion to draw from what has been said so far is that the first feature — expressibility by means of the linguistic first person — is the crucial feature one should use in characterising first person thought. The other feature follows from such a characterisation. A first person thought necessarily is a thought about the thinker because (i) it is a thought one would express by using the first person, and (ii) the first person is governed by the rule that we use it (only) to refer to ourselves.

Yet there is something deeply dissatisfying about this approach to first person thought. We appeal to the notion of a first person sentence to characterise first person thought, but when we attempt to characterise the (linguistic) first person, we appeal to the token-reflexive rule : the fact that we use the first person to refer to ourselves. Now, to follow the token-reflexive rule, we need to think of ourselves *as ourselves* — we need to think first person thoughts.

Suppose I *am* the winner and I believe, as before, that the winner will go to Tahiti. This will not lead me to say 'I will go to Tahiti' unless I also believe that *I* am the winner and (therefore) that *I* will go to Tahiti. But what is it to believe that I am the winner, or that I will go to Tahiti ? These are first person thoughts, aren't they ? What this shows is that the capacity to use the first person in language presupposes the capacity to think first person thoughts. It follows that the characterisation of first person thoughts in terms of first person utterances is ok only if it's a way of 'fixing the reference' of the phrase 'first person thought', for exposition purposes ; but it cannot be used to reach any substantive conclusion regarding first person thoughts. In particular, contrary to what one might have thought, it cannot be used to answer the question : *Why* is a first person thought necessarily a thought about the thinker ? To answer that question in a satisfactory manner, we need to abstract from the linguistic expression of first person thought : we need a way of characterising first person thoughts that is independent of issues of linguistic expression.

2.

If we abstract from linguistic expression, what can we say about first person thoughts ? I said earlier that we cannot characterise such thoughts as thoughts that are about the thinker of the thought, for many thoughts will be about the thinker of the thought by accident (as in the case of the mirror). Still, there is something to the idea that first person thoughts are thoughts about the thinker of the thought. There actually are two senses in which a thought can be about the thinker : a superficial sense in which a thought can be about the thinker without being a first person thought, and a deeper sense in which it cannot. A similar contrast arises in the language case : there too we have to draw a distinction between utterances that are accidentally about the speaker and utterances that are about the speaker in a more fundamental sense (first person utterances).

If the speaker happens to be the winner and he or she says, 'The winner will go to Tahiti', her utterance is about herself but only in the accidental sense. This is not enough to make it a first person utterance. First person utterances are about the speaker in a more fundamental sense : it is a conventional rule of the language that an utterance of the first person is about the speaker of that utterance. The relevant convention governs the reference of 'I' : it is a conventional property of the word type 'I' that a token of that word refers to the speaker. So we need to distinguish two levels in the linguistic case : the linguistic meaning of the type (Kaplan's 'character') and the referential 'content' carried by a token of the type. A token of 'The winner will go to Tahiti' uttered by the winner will be about the winner at the token level but not at the type level. Clearly, the type expression 'the winner' is not such that its tokens are supposed to refer to the speaker. In the case of 'I', however, reflexivity – the fact that a token of 'I' refers to the speaker — *is* encoded in the meaning of the type and is not merely an accidental property of the token.

Can we say analogous things about first person thoughts ? Evidently, the notion of conventional meaning does not apply in the mental realm. What we need is a property of the type that plays the same role as conventional meaning plays in the language case. Such a property must explain why the first person type in thought (the first person concept, as several philosophers call it) refers to the thinker of the thought, just as the conventional meaning of 'I' explains why a token of that word refers to the speaker.

The relevant property exists, I take it. As I argue at length elsewhere (Recanati 2012), we can think of indexical concepts in general (and the first person concept in particular) as 'mental files' whose function is to store putative

knowledge gained in virtue of standing in certain types of relation to the referent. The relevant relations are *epistemically rewarding* (ER) relations. An ER relation is a relation such that, when one stands in that relation to some object, one can gain knowledge about the object through the relation. The relation establishes a channel of information between the subject and the object.[1] The suggestion, then, is that there are mental files which are based on such relations and whose role is to store the putative knowledge gained in virtue of standing in that relation to the object. What fixes the reference of the file (what the file is about) is not the information in the file — for that can be misinformation — but the relation : the file refers to the object to which the subject stands in the relevant ER relation.

ER relations can be described as relations that are conducive to knowledge, given the (contingent) cognitive equipment of the thinker. In general, we have ways of gaining knowledge about individuals which depend upon our standing in the right relations to these individuals. Different ways correspond to different relations. First person ways of gaining information (through proprioception, kinaesthesis, introspection and so on) correspond to the relation in which one stands to an individual when one *is* that individual. So the relevant ER relation, in the first person case, is the relation of identity. That relation makes knowledge possible, given the cognitive equipment of the thinker. The knowledge one can get about an object in virtue of standing in the identity relation to that object is 'first person knowledge', or 'knowledge from inside' – the sort of knowledge one gains through proprioception and kinaesthesis. So I agree with Evans's epistemic approach to the reference of first person thoughts, and its recent defence by Morgan (Evans 1982 ; Morgan forthcoming).[2] The rule of reference – that the first person concept refers to the thinker – *follows from* the fact that an indexical file refers to the object to which the thinker stands in the ER relation which it is the function of the file to exploit. In the first person case the relation is identity, and the reference of the file is the thinker.

[1] Lewis calls such relations 'acquaintance relations'. Besides perceptual acquaintance, Lewis makes room for 'more tenuous relations of epistemic rapport' :

There are relations that someone bears to me when I get a letter from him, or I watch the swerving of a car he is driving, or I read his biography, or I hear him mentioned by name, or I investigate the clues he has left at the scene of his crime. In each case there are causal chains from him to me of a sort which would permit a flow of information. Perhaps I do get accurate information ; perhaps I get misinformation, but still the channel is there. (Lewis 1999 : 380-81).

[2] This paper started its life as a response to Morgan's paper as part of the 5th Online Consciousness Conference : see http://consciousnessonline.com/2013/02/15/a-demonstrative-model-of-first-person-thought/

On this view, the type/token distinction applies to mental files. Mental files are typed according to the type of ER relation they exploit. Thus the SELF file (my name for the first person concept) exploits the relation to oneself (viz. identity) in virtue of which one can gain knowledge about oneself in a special way, 'from inside' — a way in which one can gain knowledge about no one else (as Frege puts it). My SELF file is not the same as yours, and they refer to different persons, of course, but they belong to the same type : they are both SELF files, unified by the common ER relation it is their function to exploit. We see that the *function* of files — namely, informational exploitation of the relevant ER relation — plays the same role as the conventional meaning of indexicals : through their functional role, mental file types map to types of ER relations, just as, through their linguistic meaning (their character), indexical types map to types of contextual relation between token and referent.

One final note : there is an important asymmetry between two types of first person thought : those which correspond to (putative) items of first person knowledge gained through the relevant epistemically rewarding relations to the referent ('I am hungry', 'My knees are crossed', when these thoughts are based on the appropriate first person experience), and those which do not even putatively correspond to items of first person knowledge ('I was born in Paris' — a first person thought corresponding to a piece of knowledge that can only be gained third-personally). The epistemic relation which is conducive to first person knowledge in the former type of case is what fixes the reference of the file, on my account, but the file *also* hosts information delivered through other avenues of knowledge than the special ways we have of gaining information about ourselves. (A hallmark of information about ourselves gained third-personally is that it is vulnerable to error through misidentification, while information gained first-personallly is immune to such error.)[3]

Why is the SELF file hospitable to information gained in other ways than the first person way ? Because of the following principle governing files : *Two pieces of information (or misinformation) are stored in the same file if they are taken to be about the same object.* In this way, pieces of information which are not putative items of first person knowledge may go into the same file as first person thoughts which are putative items of first person knowledge. For example, if I believe that the winner will go to Tahiti, and hear that I won, I readily infer that I will go to Tahiti. That first person thought is based on a premise ('I am the winner') which makes it vulnerable to error through misidentification ; and that is enough to show that the information that I will go to Tahiti is not

[3] See Shoemaker 1968 and, for recent work in that area, Prosser and Recanati 2012.

gained in the first person way, from inside. In contrast, the information that I am being addressed by the person who tells me 'You won' is gained in the first person way. All these pieces of information — that I am being addressed, that I won, that I will go to Tahiti — end up in the same file, whatever their origin.

I conclude that a first person thought is a thought which deploys the first person concept, where the first person concept is construed as a mental file based on the ER relation of identity. Such thoughts are about the thinker because the self concept refers to the individual that bears that ER relation to the thinker of the thought.

3. References

Evans, G. (1982) *The Varieties of Reference* (ed. J. McDowell). Oxford : Clarendon Press.

Lewis, D. (1999) *Papers in Metaphysics and Epistemology*. Cambridge : Cambridge University Press.

Morgan, D. (forthcoming) A Demonstrative Model of First Person Thought.

Prosser, S. & Recanati, F. (eds.) (2012) *Immunity to Error through Misidentification : New Essays*. Cambridge : Cambridge University Press.

Recanati, F. (2012) *Mental Files*. Oxford : Oxford University Press.

Shoemaker, S. (1968) Self Reference and Self Awareness. *Journal of Philosophy* 65 : 555-67.

32

Against Metaphysical Disjunctivism

PASCAL LUDWIG AND EMILE THALABARD

We first met the core ideas of disjunctivism through the teaching and writing of Pascal Engel[1]. At the time, the view seemed to us as being clearly false, despite the fact that it opened new epistemological avenues, especially as far as the skeptical challenge was concerned. Today, we think that a nuanced assessment of disjunctivism is within reach. In order to defend such an assessment, we will first put forward a distinction between two aspects of the disjunctivist position, epistemological disjunctivism and metaphysical disjunctivism[2]. Epistemological disjunctivism[3] bears on the characteristics of perceptual knowledge; we will claim that it is neutral regarding the nature of perceptual experience. Metaphysical disjunctivism, on the other hand, is a view about the metaphysical nature of perceptual experience. Its main claim is that perceptual experiences are of a relational nature: the existence of conscious experiences depends on the existence of their worldly objects[4]. In order to give a first illustration of this distinction, let us consider two cases, a good case and a bad one. In the good case, a subject, let's say Mary, is seeing a red

[1] Especially through the sharp introduction to disjunctivism presented in Engel (2007).

[2] Cf. Pritchard (2012,. 23-24) for a crystal-clear recent discussion of this distinction. See also Byrne and Logue (2008) and Soteriou (2009).

[3] The main source of epistemological disjunctivism seems to be McDowell (1982). See also Byrne and Logue (2008) and Pritchard (2012).

[4] The historical sources of metaphysical disjunctivism are to be found in Hinton (1967a), Hinton (1967b), Hinton (1973), Snowdon (1981), Snowdon (1990), and Martin (2002), Martin (2004), Martin (2006). We will also rely on the presentations given by Campbell (2002), Hellie (2007) and Fish (2009). See also the papers in Byrne and Logue (2009) and Haddock and Macpherson (2008). See Crane (2006) for a comparison between metaphysical disjunctivism and its main competitor, intentionalism.

rose and forming the belief that this rose is red on the basis of her experience. In the bad case, Mary is not in optimal viewing conditions. For the sake of the discussion, we will even assume that she is having a mere hallucination of a red rose, and that she is forming a belief about a rose she thinks she is seeing on the basis of her mental condition. Let us also assume that Mary cannot distinguish, from her subjective perspective, between what it is like being in the good case and what it is like being in the bad case: for her, both situations are introspectively indistinguishable on the basis of experience. According to epistemological disjunctivism, Mary has two very different kinds of reasons for her beliefs in the good vs. the bad case. In the good case, she has a reason to believe that is both factive and reflectively accessible: because she is seeing that the rose is red, she has access to a reason that gives her a rational guarantee for the truth of the proposition that the rose is red. In the bad case, on the other hand, Mary does not have access to such a factive reason, and therefore is not in a position to gain knowledge. In this paper, we will assume the truth of epistemological disjuncgivism, because we want to focus our discussion on the related, but much more radical, *relational conception of experience*[5]. According to this conception, that we also call "metaphysical disjunctivism", there is no common, fundamental nature at all in Mary's veridical experience of the rose in the good case and her hallucinatory mental condition in the bad case, despite the fact that Mary cannot subjectively distinguish between the good case and the bad one. Indeed, the metaphysical disjunctivist claims that veridical experiences are *essentially different* from the mental conditions involved in bad cases: in the good case, what it is like to see the red rose for Mary is essentially constituted, in her view, by a relation to the fact that the rose is red. Since such a relation cannot exist in the bad case, where there is no such fact to be related to, she infers that hallucinations are of a very different nature from veridical experiences. The relational conception of experience radically departs from more standard conceptions in rejecting the claim that being subjectively indistinguishable, for two mental states A and B, is enough for being typed as identical, or at least as very similar, experiences[6]. As we will see, the metaphysical disjunctivist claims that state A can essentially differ from state B even though what it is like to be in A is the same as what it is like to be in B. We will argue that epistemological and metaphysical disjunctivism should be sharply distinguished: one can reject the relational

[5] We borrow the terms "relational conception of experience" and "Relational View" to Campbell (2002). See also Crane (2006).

[6] Cf. Martin (2004)

conception of experience while embracing the view that perception provides reasons that are both factive and reflectively accessible. We will also argue that an explanatory argument can be leveled against the Relational View, and as a consequence that it should be rejected.

1. Epistemological disjunctivism

The core thesis of epistemological disjunctivism

We will borrow the exact definition of epistemological disjunctivism to Duncan Pritchard[7]:

> *Epistemological Disjunctivism: The Core Thesis*
>
> In paradigmatic cases of perceptual knowledge an agent, S, has perceptual knowledge that Phi in virtue of being in possession of rational support, R, for her belief that Phi which is both factive (i.e. R's obtaining entails Phi) and reflectively accessible to S. (Pritchard 2012, p. 13).

In good, paradigmatic, cases of perceptual knowledge, the agent has access to a defeasible and factive reason. By seeing a red rose, Mary has access to the content of her visual experience, which presents her a certain rose being red. She is thereby in a good position to acquire the knowledge that the rose is red. The reason given by the visual experience is factive, because in a paradigmatic case of visual perception, one cannot see a rose as being red if it is not the case that it is red. At the same time, this perceptual reason is defeasible: if Mary gains evidence to the effect that the context of perception is not normal — for instance, to the effect that she might be hallucinating — it is rational for her to reconsider her belief that there is a red rose just before her. Pritchard's definition is consistent with the traditional definition of knowledge as justified true belief: having a perceptual factive reason to believe that P because one enjoys a veridical experience does not involve, in itself, possessing knowledge, but merely being in a good position to acquire it, even if one does not eventually exploit this possibility[8]. Let us imagine again that while she is visually

[7] Cf. Pritchard (2012).

[8] Cf. Pritchard (2012, 26). For the opposing view that perception directly amounts to the acquisition of knowledge, hence that there can be knowledge without justified belief, see Williamson (2000).

presented with a red rose, Mary also has good evidence that she may be hallucinating. In such a situation, she sees that the rose is red, she has a (defeasible) factive reason to believe that the rose is red, hence she is in a good position to gain knowledge of the fact that the rose is red. Nevertheless, it is rational for her in this context to suspend her judgment, not exploiting her good position to gain knowledge. As Pritchard emphasizes, it would be a mistake, in such a situation, to say that Mary knows that the rose is red: she does not know this fact since she refrained from acquiring the belief that the rose is red. It seems important, in order to leave open the possibility of such epistemic situations, to insist that having access to perceptual factive reasons does not directly provide knowledge to the perceiving subject, but only opportunities for knowledge.

What makes epistemological disjunctivism a type of *disjunctivism* is its treatment of perceptual reasons. In this framework, paradigmatic cases of perception provide factive reasons. To this extent, they essentially differ from other subjectively indiscernible mental conditions, like illusions or hallucinations. This does not mean, however, that the epistemological disjunctivist should deny that non-paradigmatic perceptual-like experiences do not confer subjective reasons. To see why this important point is true, let us say more about reasons and their role in belief acquisition. We do not think that having a reason to judge a content that *P* requires entertaining a correct argument to the effect that *P* is true. Mary's visual experience of the red rose, for instance, can be a reason both for her judging that the rose before her is red, and for the introspective judgment that she is seeing a red rose. This does not entail that she could grasp an argument to the effect that those contents are true. Rather, we take it that she has a reason to judge according to these contents because these judgments are likely to be true from her point of view, and because she has access to the content of the experience. Because it perceptually seems to Mary as though the rose is red, it rationally makes sense from her point of view to judge that the rose is red: considering the content of the experience, the truth conditions of this proposition are likely to be satisfied. In such a perspective, reasons, seen as considerations accessible for the subject and according to which certain contents are likely to be true, have two important aspects. From an objective perspective, something counts as a reason for judging that P if there is a truth-connection between its obtaining and the satisfaction of P's truth-conditions. A visual experience is an objective reason because there is a truth-connection between having a visual experience presenting the fact that P, and its being the case that P. From a subjective perspective, we think that a reason is accessible to a subject to the extent that she is sensitive to it, even

though she is not capable of making explicit the connexion between the obtaining of the reason and the obtaining of the content for which it is a reason. If the conception of reasons that we have put forward is on the right track, it should be clear that a subject can be sensitive to a certain kind of reasons, have access to these reasons while forming beliefs or making judgements, without having a full and explicit grasp of the truth-connexions that confer a warrant role to those reasons. It follows that having access to a reason does not imply knowing all its rational characteristics. This is important because typically, a subject reflecting upon the rational role of a factive reason will not know, just because she can access it, that it is factive: having access to a factive reason does not imply being able to discriminate it from a non-factive one.

If it seems to Mary that she is seeing a red rose because she is having an hallucination, her sensitivity to this state, leading her to judging that the rose is red, cannot be blamed from a rational point of view. John McDowell acknowledges this point in the following passage: "it might be rational (doxastically blameless) for the subject— who only seems to see a candle in front of her—to claim that there is a candle in front of her"[9]. Mary's doxastic behavior is not unintelligible or irrational when she judges that the rose is red on this basis, because from her point of view the hallucinatory experience is not discriminable from a factive reason. To conclude on this point: even in the bad case, an epistemological disjunctivist may accept that a non-veridical experience confers a reason to believe, despite this reason not being truth-conducive.

Epistemological disjunctivism and internalism

Epistemological disjunctivism is inconsistent, to some extent, with internalism, and it is important to understand exactly to what extent. According to both positions, a subject has a perceptual reason to judge that *P* if and only if she has access to a mental state, an experience that counts as an internalist epistemic support for *P*. Epistemological disjunctivists, however, insist that some mental states, when considered as reasons, have to be typed in a relational way. Let us consider again the contrast between a paradigmatic, truth-conducive, visual experience—the good case—, and a subjectively indiscriminable hallucination—the bad case. This means that what it is like, for the subject, to be in the good case, is identical to what it is like to be in the bad case, or at least that the subject cannot discriminate from the inside between the good case and the bad one. Nevertheless, according to epistemological

[9] McDowell (2002, 99).

disjunctivism, the subject has access to very different reasons in the good and bad cases: in the good case, but not in the bad one, she has access to a factive reason. This should not be surprising. Our folk psychology itself contrasts factive and non-factive senses of verbs like "to see". Seeing that the rose is red, in a factive interpretation, entails that the rose is red; so it makes sense to claim that a subject, by seeing (in a factive sense) that a rose is red, has access to a factive reason to believe that this rose is red. An epistemological disjunctivist, we think, should not be committed to the claim that the subject having access to a factive reason can know by reflexion alone that the reason is factive. Nor should she be committed to the claim that she cannot know such properties of reasons by reflexion alone: she should just remain neutral on this question. The only essential assumption she should be committed to, we contend, is that in accessing a factive reason in a normal case, a subject has access to a mental state that is distinct in kind[10] from the non-factive reasons she has access to in non-normal cases, even though she cannot discriminate between having access to a factive reason and having access to a non-factive one. This should not be very controversial. In the good case, a visual experience is a (truth-conducive) bearer of information, and as such accessing it gives an opportunity to gain knowledge. The fact that factive and non-factive reasons differ with respect to this epistemological (or informational) property is enough to justify the claim that they differ in kind.

What *would* be controversial would be the different claim that the subject accesses different kinds of reasons in the good and bad cases in virtue of having experiences of a different metaphysical nature. But why would an epistemological disjunctivist be committed to this? The property of being a bearer of information is analyzed, in the current theories of information, as a *relational* property[11]. So if one does not think that experiences have a relational nature, one is not committed to the claim that experiences having distinct relational properties also have, for this very reason, distinct natures.

So the kind of internalism that is inconsistent with epistemological disjunctivism is a quite strong claim. Following Duncan Pritchard[12], we will describe it by using Putnam's thought experiment of a recently envatted duplicate of a normally perceiving subject. Let us assume that Mary is having a paradigmatic, normal, veridical visual experience of a red rose, and that her brain has

[10] Let us emphasize that being distinct in kind from a non-factive reason does not imply being of a different metaphysical nature. In our terminology, two states may differ in kind because one is a bearer of information but not the other, even if they share a common metaphysical nature.

[11] See for instance Dretske (1995).

[12] Cf. Pritchard (2012).

just been duplicated and envatted. We will also suppose that Mary and Twin Mary's brains are synchronized: the patterns of activations in Twin Mary's brain are exactly the same as the patterns in Mary's brain. Let us also assume that Mary's envatted duplicate has conscious experiences, and that these conscious experiences are qualitatively indistinguishable from Mary's[13]. As we have seen, epistemological disjunctivism implies that Mary and Twin Mary *do not have access to the same kinds of reasons*. Mary's experiences have relational properties with her environment that endow them with the property of being factive, so she has, contrary to Twin Mary, access to factive reasons. This is precisely here that epistemological disjunctivism diverges from classical internalism. According to Pritchard, a widely held core thesis of epistemic internalism is the following "New Evil Genius Thesis"[14]:

> *The New Evil Genius Thesis*
>
> Mary's internalist epistemic support for believing that P is constituted solely by properties that Mary has in common with Twin Mary.

The New Evil Genius Thesis is not consistent with epistemological disjunctivism, since according to this view, the reasons Mary has access to differ in their properties from the reasons Twin Mary has access to. Let us consider Mary's visual experience of the red rose. This experience has the relational property of conveying information upon the fact that the rose Mary is seeing is red. Let us consider now the qualitatively identical twin mental state Twin Mary is in when Mary is seeing the red rose. Even if we grant that what it is like for Twin Mary while she is enjoying the experience is identical to what it is like for Mary to see a red rose, and for this reason that both experiences, having the same phenomenal character, are intrinsically alike, we do not have to accept the internalist view according to which both experiences have also exactly the same epistemological properties: Mary's and Twin Mary's experiences differ with regard to their relational properties, and these relational properties might very well be essential to their epistemological standing.

[13] This assumption, as we will see later, is controversial.

[14] This is a slightly modified version of Pritchard's own rendering of the thesis, cf. Pritchard (2012, 38).

The local supervenience thesis

Let us take stock. Epistemological disjunctivism is inconsistent with epistemological internalism in so far as it rejects the New Evil Genius Thesis. It is consistent, however, with the claim that the intrinsic properties of experiences remain the same for Mary and Twin Mary. This claim, that many metaphysicians of mind find plausible, is a consequence of the local supervenience principle. In order to be able to give a statement of this principle, let us first clarify our terminology. First, we will define the *phenomenal character* of an experience as that property of the experience that enables a subject to classify it according to what it is like to have it[15]. As a consequence, experience *E1* and experience *E2* differ in their phenomenal character exactly to the extent that what it' i like to have E1 differs from what it is like to have *E2*. Two experiences that differ in their total phenomenal character can be phenomenally similar with respect to certain dimensions. It is useful to introduce the concept of a phenomenal property to capture such similarities. Talking about the phenomenal properties of experiences is a way of typing the similarities between them. Thus, Mary's visual experience of a red rose differs qualitatively from her visual experience of a red tomato; nevertheless, the two experiences share a phenomenal property, which explains their qualitative similarity. We can now formulate the Local Supervenience Principle[16] :

> *Local Supervenience Principle*:
>
> Phenomenal properties and phenomenal characters supervene on brain properties. That is: two organisms that do not differ in their brain properties will differ neither in the phenomenal characters of the experiences they have, nor in the phenomenal properties of those experiences.

Let us assume that it is possible, in principle at least, to artificially reproduce the neural activity of a brain in a laboratory context, in the absence of the stimuli which would normally cause this neural activity. Let us also assume that for a given subject, an experience having phenomenal character P is normally correlated with the occurrence of neural activity A. The Local Supervenience

[15] Note that according to this definition, the phenomenal character of an experience is an objective feature of this experience that does not depend on the introspective capacities of the subject. It does not follow *a priori* from this definition that indiscriminable experiences should have the same phenomenal characters.

[16] We borrow the expression "local supervenience principle" to William Fish. Cf. Fish (2009, chap. 2).

Principle implies that it should be possible to replicate an experience having phenomenal character P just by reproducing the neural activity A, even in the absence of the normal objects of the experience. This means that according to the Local Supervenience Principle, Mary's and Twin Mary's experiences have the same phenomenal character: they share all their phenomenal properties. If we also assume that the metaphysical nature of experiences is essentially phenomenal — that is, that a given experience having a phenomenal character P could not instantiate a different phenomenal character in any possible world —, it follows that Mary's and Twin Mary's experiences share a common metaphysical nature if the Local Supervenience Principle is true — presumably, a common neural basis.

Again, this consequence is not inconsistent with the core thesis of epistemological disjunctivism. "Being factive" can be a property of Mary's red rose experience without being one of its *essential* properties. In the informational framework we favor, experiences carry information about the world and they do so in virtue of informational relations with the objects and properties that are instantiated in it. To this extent, an experience can be compared with a map of an environment. The shapes and colors on the map — the analogue of the phenomenal properties instantiated by the experience — do denote places and environmental characteristics in normal paradigmatic situations of use, and in such normal uses the map will give factive reasons to believe that the denoted characteristics are instantiated by the denoted places. By looking at a map, we have an opportunity to gain knowledge precisely because the map carries (factive) information in normal contexts. The factive character of the map, however, crucially depends on the existence of certain contextual relations to the environment. If we move the map in a radically different environment, for instance if we try to use it on another planet, it will of course afford no opportunity to gain knowledge. So it is because the map has certain relational properties that it carries information. These properties are not essential, as witnessed by the fact that we can use the map to navigate in a wrong environment. The map has a potential to deliver knowledge, but this potential can be expressed only if it is properly used in the right environment.

In a similar way, it can be claimed that the factive aspects of conscious perceptual experiences depend upon their relational, non-essential, properties. Such a claim makes sense in a representational framework. However, many authors have defended a metaphysically very ambitious interpretation of the main thesis of epistemological disjunctivism, that rejects representationalism and is inconsistent with the Local Supervenience Principle. We now turn to this interpretation.

2. Metaphysical disjunctivism and the relational conception of experience

The conception of the epistemic role of experience that we have sketched in the first part of our paper is disjunctivist in a very modest way: it claims that veridical perceptual experiences are factive reasons to believe, and that they should be typed apart from illusions and hallucinations at least to this extent. This does not imply that there is nothing mental in common between veridical and non-veridical experiences: two mental states may differ relative to their epistemological standings, one being a factive reason contrary to the other, but still have a common mental nature. This epistemological difference may lead one to classify them in different categories — after all, they have distinct epistemological properties, since veridical experiences reveal the world as it is to the subject, whereas illusions and hallucinations do not — while remaining neutral upon whether they have a common mental nature or not.

Metaphysical disjunctivism and the rejection of the common, fundamental kind thesis

Many disjunctivists are more ambitious, and claim that veridical states and hallucinations are of different fundamental kinds. Note that nobody claims that these states have *absolutely nothing in common*, since both a veridical experience and a hallucination may at least share the property of being subjectively indiscriminable from a perception of an F. The interesting and controversial claim is that they do not share any fundamental property:

> *Metaphysical Disjunctivism: the Core Thesis*
>
> Veridical perceptual experiences do not share any essential, fundamental, nature with non-veridical experiences (like hallucinations or illusions).

One finds a clear statement of this thesis in M. G. Martin's writings, who characterizes disjunctivism as the rejection of the Common Kind Assumption, thus formulated: "whatever kind of mental event occurs when one is veridically perceiving some scene, such as the street scene outside my window, that kind of event can occur whether or not one is perceiving"[17].

It should be clear that the core thesis of epistemological disjunctivism does not logically imply the core thesis of metaphysical disjunctivism. As Duncan

[17] Cf. Martin (2004), in Byrne and Logue (2009, 273).

Pritchard emphasizes, "that the rational standing available to the agent in normal veridical perceptual experiences and corresponding to (introspectively indistinguishable) cases of illusion and hallucination are radically different does not in itself entail that there is no common metaphysical essence to the experience of the agent in both cases"[18]. So, metaphysical disjunctivism does not follow from epistemological disjunctivism.

Naïve realism and the relational conception of experience

What are the motivations for rejecting the common kind assumption, then? It is difficult to give a completely systematic answer since the core thesis of metaphysical disjunctivism is negative. However the most interesting motivation has to do with a simple and attractive conception of conscious experience, that Martin calls "naïve realism": "the prime reason for endorsing disjunctivism, he writes, is to block the rejection of a view of perception I'll label *Naïve Realism*. The Naïve Realist thinks that some at least of our sensory episodes are presentations of an experience-independent reality"[19]. The notion of presentation, in this quote, should be interpreted in the following way: objects and their properties are *constitutive* of the phenomenal character of our conscious experiences. In order for there to be a conscious experience for a subject, she has to be presented with certain facts. If the facts did not exist, they could not be presented, and as a consequence the experience would not exist. Naïve Realism, as Martin understands it, considers any perceptual experience as a relational structure existentially dependent upon its *relata*. For this reason, following John Campbell, we will also call it the "relational conception of experience". As Campbell puts it:

> On a Relational View, the phenomenal character of your experience, as you look around the room, is constituted by the actual layout of the room itself: which particular objects there are, their intrinsic properties, such as colour and shape, and how they are arranged in relation to one another and to you. On this Relational View, two ordinary observers standing in roughly the same place, looking at the same scene, are bound to have experiences with the same phenomenal character. Campbell (2002, 116).

[18] Cf. Pritchard (2012, 24).
[19] Martin (2004), in Byrne and Logue (2009, 272).

In order to have a good understanding of the relational conception of experience, it is convenient to follow Campbell and to contrast it with its main contender, the Representationalist View. According to Campbell's own characterization:

> On (...) a Representationalist analysis, in contrast, perception involves being in representational states, and the phenomenal character of your experience is constituted not by the way your surroundings are, but by the contents of your representational states. Campbell (2002, p. 116).

According to this definition, experiences have representational properties which determine their representational content, and their phenomenal characters are constituted by these contents. This is not the only way to characterize the Representationalist View, nor maybe the best, but we will grant it for the sake of discussion.

The Relational and Representationalist views of experience give a very different analysis of what being consciously aware of an object (or an instantiated property) amounts to. According to the Relational View, conscious awareness is a (perceptual) relation to the objects present in the perceived scene and to their properties. That is the reason why, as John Campbell puts it, "we have to think of the external object, in cases of veridical perception, as a constituent of the experience. (...) We have to think of cognitive processes as 'revealing' the world to the subject, as making it possible for the subject to experience particular external objects" (Campbell, 2002), p. 118[20].

A very close relative of the Relational View that is worth mentioning is the view that the phenomenal characters of veridical experiences are factive and purely mental properties, a view that Benj Hellie calls "Phenomenal Naivete"[21]. Strictly speaking, Campbell's Relational View does not imply Phenomenal Naivete, because he construes the phenomenal characters of experiences as acquaintance relations to particulars and instantiated properties in the world, not as acquaintance relations to facts. The subtle distinction between the Relational View and Phenomenal Naivete is of no importance in the context of the

[20] One finds a similar formulation in Martin's writings when he claims that «some of the objects of perception—the concrete individuals, their properties, the events that partake in it—are constituents of the experience. No experience like this, no experience of fundamentally the same kind, could have occurred had no appropriate candidate for awareness existed» Martin (2004), in Byrne and Logue (2009, 273). Martin, however, is not committed to the idea that conscious experience is existentially dependent on worldly objects.

[21] Cf Hellie (2007, 264-265).

present paper, so we will sometime speak as if phenomenal characters were factive according to the Relational View.

According to representationalism, now, one is consciously aware of an object O being P if and only if one is having an experience representing O as being P. Conscious awareness, in this view, is a relation between the subject and a represented object. One sometimes reads that the represented object is a constituent of the representational content of the state, but this is contentious, since on some views contents are unstructured (for instance when they are construed as sets of possible worlds). Besides, the representational relation is intentional. This means that in a representationalist framework, a subject may be consciously aware of an entity that is not really present in the perceptual scene. The Representationalist View implies that normative conditions of satisfaction are associated with experiences: being a representation, a given experience is correct in some contexts, and incorrect in other contexts. This is enough to draw a distinction between the Relational View and the Representationalist View, since the former is not committed to the claim that experiences have conditions of satisfaction.

The contrast between the two positions is especially striking when one considers situations in which perceptual experiences occur in an abnormal way, for instance situations of hallucination. The Representationalist View can explain why Mary's hallucinatory visual experience of a red rose is indiscernible from a veridical experience: in the bad case as in the good one, the experience is nothing but a visual representation of a rose being red[22]. Since a state can represent another state in its absence, the existence of the representation does not depend upon the actual presence of its intentional objects in the scene of perception. The representational properties of the perceptual state and its representational content may be exactly the same in the good case and in the bad one. It follows that on a representationalist view, one may assume that there is a fundamental mental nature in common between the good case and the bad one, namely, a certain perceptual representation.

An advocate of the Relational View is bound to disagree. On this view, the perceived object is a constituent of the conscious experience in the good, paradigmatic case of perception. In the bad case, where no real object is to be perceived, nothing can enter into the experience as such a constituent. How are we to understand that the visual experience, in the bad case, subjectively feels just like its veridical counterpart? According to Campbell, éthe experience is quite different in the case of the hallucination, since there is no object

[22] See Smith (2002).

to be a constituent of your experienceé[23]. This is quite an understatement, though. By his own admission, the phenomenal characters of conscious experiences are metaphysically *constituted* by the real objects of these experiences. This logically entails that an experience without object cannot have any phenomenal character at all. There isn't anything it is like to hallucinate a red rose, since such a mental state doesn't disclose any fact in the world that could serve as its object. A "mere" hallucinatory or illusory state cannot be an experience in the full sense, since it is hard to see how it could have a phenomenal character. It follows that a disjunctive analysis of the concept of experience is inevitable: an experience is either a perceptual relation to the world, or a state of a very different kind. The problem that remains, and that we will address later in the paper, is to understand how a state devoid of any qualitative character may be subjectively indistinguishable from a conscious perceptual experience.

Some motivations for the Relational View

In this section we will present and discuss two important motivations for the Relational View.

Transparency

The first motivation is phenomenological. According to the Relational View, one in only aware of the real objects present in a perceptual scene and of their properties in an episode of veridical perception. To this extent, the Relational View seems to be in line with what the phenomenology of such episodes reveals in introspection. When we introspectively reflect upon the characteristics of our perceptual experiences, we do not gain knowledge on anything internal to the mind or on anything having to do with representational vehicles or with representational properties. Let's take Mary who, while perceiving a red rose, focuses her attention not directly on the rose, but rather on her experience of it. What will she learn through introspection? She will self-ascribe a perception of a red rose, a knowledge she would thus express:

(1) I am seeing a rose, and the rose I am seeing is red.

Such a self-ascription does not characterize the visual experience by referring to any internal object, but rather by directly referring to the object seen.

[23] Cf. Campbell (2002,117).

This reflects the transparency of experience: attending to the "reddish" phenomenal quality of the experience, it seems, is phenomenologically nothing else than attending to the color quality of the rose—a worldly property of a worldly object. Let us borrow the formulation of the transparency thesis to Christopher Hill:

> *Transparency Thesis*: when one tries to attend introspectively to a perceptual experience, (...) one is aware only of what it is an experience of (...).[24]

Let us emphasize that the Transparency Thesis is an epistemological claim, not a metaphysical claim. Accepting the Transparency Thesis does only imply that we gain knowledge about the phenomenal properties of our experiences by attending to the objects of these experiences. The thesis is utterly silent on the nature of those objects and on the nature of those phenomenal properties[25]. It does not imply, for instance, that the phenomenal properties of experiences are supervenient on the properties of their objects: it only implies that those phenomenal properties that can be known by introspection supervene on properties of the perceived objects. Thus, the thesis does not imply that phenomenal properties are essentially object-dependent, but only that we get information about them by attending to objects. Transparency is a phenomenological fact that a good theory of consciousness should explain; it should not count as a decisive argument in favour of any theory.

The folk psychology of appearances and the Relational View as the default position

In view of the above, an inference to the best explanation could be drawn to the effect that the Relational View is true, along the following lines:

1. trough introspective reflection, conscious sense perception seems to us to be nothing else that a direct contact with the perceived objects and their properties;

[24] Cf Hill (2009, 57). Hill characterizes the Transparency Thesis further, by saying that in introspecting one is aware of "what the experience represents or signifies". This reflects his commitment to an intensionalist theory of perception. We leave this out of our definition of transparency, because we want to define it in a neutral way wit respect to bpth the Relational and the Representationalist Views.

[25] Cf. Kind (2003).

2. the Relational View, which construes experiences as an acquaintance relation between the subject and the objects of experience, is the best explanation of this observation;

3. so the Relational View is probably true.

Some authors think that this reasoning can be strengthened by appealing to experts. We are not convinced that it really makes sense to refer to expertise in a domain like introspection, but let us assume, at least for the sake of the discussion, that there are indeed experts in phenomenology. Benj Hellie borrows the following five quotes from such experts, whose convergent testimonies are supposed to bring support to the Relational View[26]:

> In its purely phenomenological aspects seeing is (...) ostensibly prehensive of the surfaces of distant bodies as coloured and extended. It is a natural, if paradoxical, way of speaking to say that seeing seems to "bring one into direct contact with remote objects" and to reveal their shapes and colours. (Broad, 1952, 32-33);

> Mature sensible experience (in general) presents itself as [...] an immediate consciousness of the existence of things outside of us. (Strawson, 1979, 97);

> When someone has a fact made manifest to him, [. . .] the obtaining of the fact is precisely not blankly external to his subjectivity. (McDowell, 1982, 390–1)

> Visual phenomenology makes it for a subject as if a scene is simply presented. Veridical perception, illusion, and hallucination seem to place objects and their features directly before the mind. (Sturgeon, 2000, 9)

> The ripe tomato seems immediately present to me in experience. I am not in any way aware of any cognitive distance between me and the scene in front of me; the fact that what I'm doing is representing the world is clearly not itself part of the experience. The world is just there. (Levine, 2006, 179)

[26] Hellie (2007, 266). Cf also Fish, (2009, chap. 1), who seems to agree with Hellie that this list brings support to the Relational View.

We agree with the advocates of the Relational View that these "experts" give a faithful rendering of the phenomenology of visual experience. It seems to us, however, that these testimonies do not give any strong support to the Relational View. What seems to be coming out therefrom is that visual experience is conceived as an immediate relation to the objects we are seeing. We concur, and we even think that folk psychology typically conceives perceptual experience as being relational. This does not tell much in favor of the Relational View, however, because the Relational View bears on the *metaphysical nature* of perceptual experiences, not on the way they are typically conceived. Let us develop this further.

3. Representationalism as an alternative explanation

Our strategy in this paper is to grant to the disjunctivist that perceptual experiences are factive reasons, and that they are conceived as such by ordinary people. Ordinary people seem to think, along with the "experts", (i) that we are related, through our visual experiences, to objects in the world and to their properties (ii) that this relation is immediate, and that as a consequence the objects are "presented" to us in perception. By "immediate", it seems we just need to understand that the relation is not inferentially based: looking at objects enable us to gain veridical information about them in a non-inferential way. Apart from that, folk psychology is not committed to any particular conception of the perceptual relation and the perceptual states. As a consequence, there does not seem to be any inconsistency between the judgements of the experts and of the folk on the one hand, and representationalism on the other hand, at least insofar as the natures of the perceptual relation and of the perceptual states are concerned. According to the representationalist view, the function of perceptual-representational systems is to track ecologically relevant objects in the world, in order to enable the cognitive agent to accumulate information about them and to act upon them. It follows that in normal cases of veridical perception, perception can indeed be seen as relational in such a framework, since representational states are related to their objects by informational channels. Campbell emphasizes that on a Relational View of perception, we have to think of cognitive processes as 'revealing' the world to the subject Campbell (2002, 118). It is hard to see, however, why the revelation metaphor could not be applied to the Representationalist View as well as to the Relational View: as we have insisted in the first part of this paper, the Representationalist View can incorporate the idea that perceptual

experiences are factive reasons. In normal contexts, the occurrence of a perceptual representation is linked to the existence of an informational channel relating the subject to the perceived scene: the experience would simply not occur if the informational channel did not exist, and if it did not allow a flow of information. Following David Lewis, let us call "acquaintance relations" the informational channels through which we gain information about the objects we perceive and their features[27].

The Representationalist View implies that subjects are normally acquainted with the objects of perception, and that this acquaintance relation is direct and immediate, in the sense that it does not rely on any inference[28]. It is also consistent with the transparency of experience. The function of representational systems is to collect information about ecologically relevant, objective features of the organism's environment. The states of those systems represent objective environmental states. To this extent, they are about objects in the perceptual scene, not about mental objects. As a matter of historical fact, some of the first and foremost advocates of the transparency of experience are also advocates of representationalism. For instance, Gilbert Harman claims that our experience of the world is not mediated in any way by a prior and more fundamental awareness of intrinsic mental features:

> When you see a tree, you do not experience any features as intrinsic features of your experience. Look at a tree and try to turn your attention to the intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree (...). (Harman, 1990, 667).

Harman also contends that we are only aware of the represented intentional objects of our experiences, not of their intrinsic non-intentional aspects. This is the point he makes in the following text:

> In the case of a painting, Eloise can be aware of those features of the painting that are responsible for its being a painting of a unicorn. That is, she can turn her attention to the pattern of the paint on the canvas by virtue of which the painting represents a unicorn.

[27] Perceptual relations are the paradigm of acquaintance relations according to Lewis; they are based on "channels" or "causal chains" from the object to the cognitive system which "permit a flow of information». Cf. Lewis (1999, 380-381).

[28] This is of course consistent with the popular idea that visual representations are constructed by the brain through algorithmic processes. Such processes are sub-personal, hence non-inferential.

> But in the case of her visual experience of a tree, I want to say that she is not aware of, as it were, the mental paint by virtue of which her experience is experience of seeing a tree. She is aware only of the intentional or relational features of her experience, not of its intrinsic nonintentional features. (Harman, 1990, *ibid.*).

In light of the above discussions, we can say that both the Relational View and the Representationalist View can explain the same range of phenomenological facts. Both views conceive perception as relational[29].

4. An explanatory argument against the Relational View

We have argued that the Representationalist View has the resources to explain the phenomenology of perceptual experiences. We have also argued that there is no reason why a representationalist could not endorse epistemological disjunctivism. What remains to be demonstrated, now, is that the Representationalist View provides a better overall explanation of the phenomenological *data*.

Let us start with the following methodological principle, that we think should not be controversial:

> *Explanatory Constraint*: A good theory of conscious experience and its phenomenal properties should be able to explain the phenomenal similarities and dissimilarities mong experiences.

Now, the Relational View implies that the phenomenal character of perceptual experiences metaphysically depends on the objects and properties the subject is related to when she perceives. As Campbell writes:

[29] This is contested by some authors. For instance, Tim Crane writes that "the intentionalist view (...) comes with a price. For it must deny that perceptual experience is a relation. When one does succeed in perceiving an object, one is related to it, of course; but this relation is not essential to the perceptual experience being of the fundamental kind that it is"». (Crane, 2006, 141). This statement might first strike us as blatantly contradictory, since Crane describes the intentionalists both as denying that "perceptual experience is a relation", and as claiming that "when one does succeed in perceiving an object, one is related to it". Crane does not deny that perception *can* be interpreted as relational by the Representational View: as we have insisted on before, in paradigmatic contexts of veridical perception, representational states are bearers of factive information about the perceptual scene. What he denies is that perception, that is, the first-order representational state brought about by perception, is *essentially* relational on the Representationalist View.

> (...) thousands of people might visit the very same spot and enjoy the same external objects. You characterize the experience they are having by saying which view they are enjoying. On the Relational View, this is the same thing as describing the phenomenal character of their experiences. (Campbell, 2002, 116).

This leads to a precise prediction that the advocates of the Relational View should endorse:

> *Similarity of Objects Principle*: the similarities between conscious perceptual experiences should always be explainable by appealing to similarities in the objects and properties perceived in these experiences: similarities between sensations are due to similarities in their real objective correlates.

It follows that any two similar experiences with respect to their phenomenal properties but dissimilar in their objects would constitute a counter-example to the Relational View.

Let us emphasize at the outset that the explanatory constraint that we have put forward is consistent with the very modest conception of introspection that is advocated by the Relational View[30]. It does indeed not imply that subjectively indiscernible experiences should have identical phenomenal characters, but only, much more modestly, that their subjective indiscriminability should be explainable by only referring to the properties of their objects. In this regard, we do not see the existence of subjectively indiscriminable experiences having different objects as a problem for the Relational View, as long as it can explain the subjectively felt resemblance between those experiences[31]. Let us consider Dretske's example of two subjectively indiscriminable black-horse experiences, E1 and E2, having two distinct horses H1 and H2 as objects[32]. It is true that according to the Relational View, E1 and E2 have distinct phenomenal characters, since H1 and H2 are numerically distinct. This is not

[30] Cf. Martin (2004).

[31] It seems to us that if one accepts the intransitivity of indiscriminability, one should also accept that there should be indiscriminable perceptual experiences of the world having distinct phenomenal characters. We are not committed to this claim, but we do not consider it to be blatantly implausible either. For a very different view, see Smith (2002), who claims that as a matter of definition, subjectively indiscriminable experiences should have identical phenomenal characters.

[32] Cf. Dretske (1995).

as implausible as it might seem: if one endorses the modest account of introspection favored by the Relational View, one should abandon the idea that indiscernible experiences are necessarily type-identical. As Martin makes clear, this very common presupposition could be questioned:

> Many have supposed that what we mean by the phenomenal character of an experience is just that aspect of it which is introspectible, and hence that any two experiences which are introspectively indiscriminable must share their phenomenal characters, even if they differ in other ways. Now, while some such complaints may have widespread support in discussions of phenomenal consciousness, it is not clear whether is should be taken as a primitive claim which is somehow obvious, and the rejection of which is incredible. (Martin, 2006, 366-367).

The important point, as far as the Explanatory Constraint is concerned, is that the Relational View can explain the phenomenal similarity between the indiscriminable experiences E1 and E2: E1 and E2 have phenomenal characters that are metaphysically distinct; nevertheless, since H1 and H2 share many of their properties—we may assume that they share all their intrinsic properties—, it is easy to explain the phenomenal similarity between E1 and E2 by appealing to the similarity between H1 and H2.

The Similarity of Objects Principle is also compatible with the recognition that dissimilarities in the phenomenal contents of experiences should sometimes be explained, at least in part, by referring to characteristics of the cognitive systems of the subjects having the experiences. It is sometimes claimed that the Relational View wrongly "attribute[s] all of the distinguishing features of every fact of perceptual consciousness to the entities that count as the objects of consciousness"[33]. We do not agree with this claim. Of course, it is true that "how an object of consciousness appears to us sometimes depends, at least in part, on factors that lie on the subject side of the subject/object divide"[34]. But this is not inconsistent with the Relational View: two subjects facing the same object may be presented with different experiential contents simply because their cognitive system does not respond to the same subset of properties among the set of all the properties instantiated in the object. We think that this is what Campbell alludes to in the following passage[35]:

[33] Cf. Hill (2009, 83). Cf. also Rey (2005).

[34] Cf. Hill (2009, 83).

[35] Rey also quotes this passage, and concludes that Campbell is contradicting himself: "per-

> After all, two people could be seeing the very same object, and yet the intrinsic character of their experience be quite different. This in itself is undeniable. It is the next step that leads to rejection of the Relational View. The next step is to say that the way in which the object is given is independent of whether the object exists, and independent of whether the subject is experiencing one or many similar objects. (Campbell, 2002, 126).

Perception cannot reveal to the subjects all the properties that are instantiated in a given context of perception: the facts that are seen depend on the subject's perspective, but also on the perceptual-recognitional abilities that are actualized by the subject in the context[36]. For instance, a subject may be unable to visually recognize a given color instantiated by an object. In such a case, her visual experience will differ from the visual experience of another subject endowed with a more sensitive recognition ability, even though both subjects are presented with the same object[37].

So let us now turn our attention to cases that are really problematic for the Relational View, i.e., to cases in which similarities between sensory experiences cannot be explained by similarities in their objects. Hallucinations are a *prima facie* clear counter-example to the *Similarity of Objects Principle*: in an hallucinatory episode of a red rose, Mary enjoys an experience that she would describe as very similar to a veridical visual experience of a red rose,

haps I am missing something here, he writes, but it's hard not to construe these passages as flatly contradictory, and as a *reductio* of the Relational View". Cf. Rey (2005, 138). We do not agree with Rey, because we do not think that the Relational View logically implies that two persons seeing the same objects instantiating the same properties will be presented with the same phenomenal contents.

[36] Block (2010) suggests an argument against direct realism by appealing to the phenomenal effects of attention: two perceptual experiences of the same worldly objects and properties may exhibit different phenomenologies because according to the distribution and focalization of attention, some features of experience will be more or less salient. Since these very objects and properties are constitutive of the phenomenology of perceptual experience, it seems that a naive realist is at a loss when having to explain why these two experiences differ. Block's argument takes the form of a dilemma: either the naive realist tries to explain away the phenomenal difference, or he bites the bullet and considers that one of the two experiences, differing only by the distribution and focalization of attention, is illusory. The latter explanation is unsatisfactory, since it would make illusion too widespread. However, there seems to be no explanation available, following the first strategy, that wouldn't appeal to mental properties in order to account for the difference in phenomenal characters. Given our remarks concerning the influence of the subject's cognitive system, it seems we can sidestep Block's objection: attention has a role in how visual information is picked up, and hence, on how worldy objects and properties contribute to a subject's phenomenology.

[37] On this point, see Fish (2009, chap. 3).

in the absence of any seen object. Mary's testimony that she had a conscious visual experience very similar to a veridical experience of a red rose during the episode is hard to reconcile with the Relational View. In an hallucinatory episode, no real fact is revealed to the subject. There is nothing real in the scene that could constitute the phenomenal character of the experience. So it is even hard to understand, on the Relational View, how Mary can claim that she had a conscious experience endowed with a phenomenal character[38].

The more radical way to address this difficulty is to bite the bullet and claim that a hallucination, and more generally any non-factive experience, is only "conscious" in a derivative sense, because it simply *does not have any phenomenal character*. William Fish puts forward such a bold approach[39]. He advocates an error-theory of hallucinations as conscious experiences. On this approach, when Mary hallucinates a red rose, the mental state she is in during the episode entirely lacks any phenomenal character. All that is happening in her mind is that she wrongly forms the same introspective beliefs and behaviors that she would acquire in the context of a veridical visual experience. In particular, she acquires the (false) belief that she had a visual experience with the phenomenal character of a red-rose perception.

Fish is opposing the majority view in the philosophy of phenomenal consciousness, according to which there cannot be any distinction between it seeming to a subject as if she is having a phenomenal experience and her really having this experience. Most philosophers are strongly inclined to think that there is no room for the appearance/reality distinction in our introspective grasp of phenomenal states.

This leads to a first argument against Fish's radical position, the argument from the authority of the subject on her self-ascription of phenomenal contents: it just seems inappropriate to raise doubts about self-ascriptions of phenomenal contents, even in non-veridical contexts of perception, and this seems to stem from the very meaning of our concept of a sensory conscious experience. Subjects seem to have a special kind of authority upon these self-ascriptions. There is a contrast, in this regard, between the following dialogues:

(H) a. Mary: This is a red rose.
b. Pierre: You are wrong. There is no rose at all in front of you, you are hallucinating.

[38] Cf. Smith (2002).
[39] Cf. Fish (2009).

(E) a. Mary: It now visually appears to me as if there is a red rose in front of me.
b. Pierre: You are wrong. Nothing visually appears to you, you are hallucinating.

(H) is OK: in abnormal circumstances, one can raise doubt about the rational justification provided to a subject by one of her visual experiences. (E) seems not only odd but, according to our folk psychology of visual hallucinations and of visual appearances, contradictory. Fish argues that if our folk psychology considers that there is no appearance/reality distinction in the domain of conscious experiences, then our folk psychology is systematically mistaken: hallucinations appear to have a phenomenal character, despite the fact that there is literally nothing it is like to having an hallucination. He borrows to David Rosenthal's higher order thought theory of consciousness the idea that a subject may have a higher-order thought that she is in a first-order mental state of a given kind even in the absence of this first-order thought. Again, this claim has very counter-intuitive consequences. Consider an amputated patient feeling pain in her phantom limb. On Fish's view, such phantom pains cannot share the phenomenal properties of veridical episodes of nociception, since they simply do not have any phenomenal character. If follows that the subject reporting a painful experience in a phantom limb is wrong: the non-veridical sensory state cannot be painful, since it is devoid of phenomenal properties. She only has the higher-order thoughts that accompany normal, veridical, experiences of pain, but these states, not being strictly speaking "phenomenal", do not exemplify the phenomenal property of painfulness. It is very hard to believe, however, that the existence of higher-order conceptual thoughts could account for the painfulness of the subject's phantom limb.

According to our folk psychology, conscious experiences have a dual role. First, they have an *explanatory role*. The occurrence of experiences can typically cause motor responses and can lead to the acquisition of beliefs. This seems to be true in hallucinatory context as well as in normal contexts of veridical perception: thus, Macbeth's hallucination of a dagger before him causes him to grasp for something. His conscious visual experience explains his reaching behavior. According to Fish's theory however, Macbeth doesn't have any conscious experience we could refer to when explaining his behavior. So how are we to explain it? An obvious answer is to mention the higher-order thoughts acquired by Macbeth during the episode, in particular the belief that it visually appears to him that there is a dagger before him. This is unsatisfactory, however, because Macbeth's acquiring the non-veridical higher-order belief

about his visual experience is left completely unexplained. How is this belief acquired? It is not caused by any conscious visual experience, since Fish denies the existence of such experiences in hallucinatory contexts. So we must suppose that the belief is caused by unconscious mental states, presumably by unconscious states of Macbeth'svisual cortex. This is a very unwelcome consequence: to our knowledge, unconscious visual states are not apt to directly cause beliefs. Cognitive neuroscientists of vision postulate numerous types of unconscious representations and of unconscious processes, but none of these representations are supposed to directly give rise to beliefs, precisely because they are unconscious, hence not accessible to the subject.

So we see that it is hard to explain why Macbeth has self-ascribed a visual content of a dagger in front of him in the absence of any conscious experience that could have caused this introspective belief. This is the second problem that Fish's theory has to face: it cannot explain how and why the higher-order thoughts that play, according to his view, such a prominent role in accounting for the subject's linguistic behavior in hallucinatory contexts are acquired. Let us emphasize that we do not deny that higher-order thoughts may have an important role to play in understanding some hallucinations. According to the metacognitive belief model of hallucinatory experience, for instance, hallucinatory episodes arise from the externalization of intrusive thoughts—typically, of unintentionally occurring sounds or visual images[40]. On this view, however, the occurrence of higher-order thoughts is explained by the occurrence of first-order states which are themselves endowed with phenomenal content. Fish cannot appeal to such states, since he claims that non-factive states are devoid of any phenomenology[41].

[40] Cf. Filippo Varese and Frank Laroi (2012).

[41] Against Fish's "reflexive account" of the metacognitive view on hallucination, Jérôme Dokic and Jean-Rémy Martin (2012) endorse a "monitoring account" according to which hallucinations are mistaken for veridical perceptions because of low-level metacognitive mechanisms, responsible for the monitoring of the quality of first-order experiential states. These "metaperceptual" mechanisms usually detect whether an experiential state has been generated internally or externally and are sensitive to its source (be it perception, imagination or what have you). On this account, hallucinations are simply states which have been wrongly tagged by this low-level monitoring system *as perceptual states* and which, as a result, produce the same cognitive effect as a genuine perception, without sharing its sensory phenomenology: indeed, no sensory phenomenology is at play in such a case – what these wrongly tagged states share with genuine perceptions is a "feeling of reality", which is no part of the sensory content of perception. We do not think that this view escapes the explanatory problem we have raised. Indeed, it has to face the following dilemma. Either the states tagged as perceptual states are endowed with a phenomenology of their own, differing in kind from the phenomenology of factive states. But then, the main advantage of Fish's approach is lost, since the very existence of this phenomenology and

There is also a third problem, which is related to the rationality of introspective beliefs, and to their relation to knowledge. We have commented above on the *explanatory role* of conscious experiences. These states also have a *justificatory role*: conscious experiences give reasons to act and believe. This seems to be also true for hallucinations. Macbeth's hallucinatory vision of a dagger not only causally explains, but also rationally justifies, his decision to try to reach a dagger in front of him. As Pautz notes[42], even philosophers who endorse a radically externalist conception of perceptual evidence, as for instance Timothy Williamson does, typically agree that an illusory or an hallucinatory experience provides a justification[43]: the visual appearing of a dagger in front of Macbeth is a reason for him to form a belief and to act on the basis of this belief.

On Fish's view, by contrast, an hallucinatory state is metaphysically constituted by a set of non-veridical higher-order thoughts. This entails that hallucinatory states cannot be reasons in any sense or justify actions or beliefs—not even introspective beliefs. They can play no rational role in thought. So clear cases in which an hallucinatory state would play a rational role in motivating a conclusion would be a decisive argument against Fish's view. Let us discuss two such cases:

> *Case 1: The lucid hallucinator*
>
> Jean-Paul S. is an expert in phenomenology. He ingests drugs on a regular basis, in order to study what he takes to be the phenomenology of visual hallucinations. These drugs give rise to episodes that are difficult to discriminate from veridical perceptions. Along these similarities, there are also some subtle differences that he is able to notice, so he is able to discriminate hallucinations from veridical perceptions when he concentrates. One morning however, as he wakes up, he happens to have forgotten whether or not he has ingested his drug. As a consequence, he concentrates on

its nature would have to be explained. Or they are devoid of any phenomenal character. But then, again, why do they give rise to higher-order cognitive states such as beliefs, episodic memories, ...etc? As far as we know, only conscious states give rise to beliefs or to other cognitives states.

[42] Cf. Pautz (2013).

[43] « In unfavourable circumstances, one can fail to gain perceptual knowledge, perhaps because things are not the way they appear to be. (...) Nevertheless, one still has perceptual evidence, even if the propositions it supports are false. True propositions can make a false proposition probable (...). If perceptual evidence in the case of illusions consists of true propositions, what are they? The obvious answer is: the proposition that things appear to be that way » (Williamson, 2000, 198).

> the phenomenology of the rich visual experience he is enjoying in order to decide whether or not this experience is veridical. After a short a while, because he has carefully taken note of some relevant characteristics of his experience, he concludes that he is enjoying an hallucination, which is true.

Jean-Paul's conclusion seems to be not only true, but justified: Jean-Paul *knows* that he has gone through an hallucinatory episode. Noticing the specific characteristics of his visual experience, he has rationally come to the conclusion that this experience is not veridical. It seems difficult, however, to make sense of this case on Fish's approach: how could Jean-Paul get knowledge about the phenomenology of his hallucination if hallucinations do not have any phenomenology to begin with?

> *Case 2: Psychedelic Mary*
>
> Like Jackson's Mary, Psychedelic Mary has never seen any color. One day, however, she discovers that some drugs can systematically produce visual hallucinations of colors. Because she has (again, like Jackson's Mary) total knowledge of the working of her brain, she can predict which drug is going to produce which hallucination. Hence, she can describe the colors that she hallucinates as being red, orange, rose, green, etc. Despite never having seen any red object, it seems that Mary knows what it is like to have an experience of red. To this extent, it seems that Mary has gained knowledge about the phenomenal character of a "reddish" experience. This is also reflected in her ability to correctly describe similarities among colors. For instance, she knows that an experience of orange is more phenomenally similar to an experience of red than to an experience of green and that an experience of violet is more phenomenally similar to an experience of blue than to an experience of green[44].

Fish's theory entails that Psychedelic Mary has not acquired any knowledge about color visual experience. How could she have acquired such knowledge if, as Fish claims, hallucinations do not have any phenomenal charac-

[44] We borrow the idea of this thought experiment to Johnston (2004). Johnston claims, as we do, that "Mary could come to know what red is like by hallucinating ... [Even in hallucination] one comes to know what certain qualities are like, and ... so [one] is able to place them in a [ressemblance-order] with other qualities of the same family" (Johnston, 2004, 130-131). Cf. also Hawthorne and Kovakovich (2006, 178).

ter? Mary's inferential and linguistic behavior would therefore be difficult to explain. We must remember that according to Fish's approach, Psychedelic Mary never had any conscious experience of colors. It follows, we may presume, that the color concepts she seems to be using, for instance when she claims that orange is phenomenally more similar to red than to green, do not denote anything. So Mary seems to have acquired knowledge, but she doesn't know anything about color experiences; she seems to be able to recognize colors, but she does not master any color concepts; she seems to make true statements about the phenomenal relations between color experiences, but these statements are just devoid of any content. This is not credible; for this reason, the conceivability of Psychedelic Mary's case is inconsistent with Fish's theory.

Let us conclude on Fish's radical view. This view rests on a revisionary conception of consciousness, according to which the subjects do not always have authority upon the contents of their conscious experiences. We argued to the effect that it has to face two serious objections: it can neither give any convincing explanation of the higher-order thoughts it appeals to, nor a correct account of the justificatory role of hallucinatory episodes. It seems very implausible, to deny that subjects enjoy a kind of conscious experience when they are hallucinating. This does not imply, however, that the Relational View is false: an advocate of the Relational View may grant that hallucinations (and other non-veridical experiences) have phenomenal characters, while insisting that these phenomenal characters differ in kind from the phenomenal characters of veridical experiences[45]. This seems to be Mike Martin's position. Indeed, Martin endorses the following claims:

1. certain visual experiences, namely, "causally matching" hallucinations[46], are introspectively indiscriminable from veridical perceptions[47];

2. these experiences are phenomenally conscious: "Surely the condition of introspective indiscriminability guarantees that phenomenal conscious-

[45] This is also Hinton's position: "In the first place, there must be indistinguishable, or at least closely similar, subjective events; though not at all in the way that the doctrine of visual experiences requires, not ones that I can tell you about. We have touched on this already: it would be absurd not to posit, not to hypothesize, similar going-on in me when I see a flash of light and when I have that illusion." (1967b, 226)

[46] A "causally matching hallucination", in Martin's terminology, is an hallucination that is "brought about through the same proximal causal conditions as a veridical perception" (Martin, 2006, 368).

[47] Cf. Martin (2006, 369).

ness is present"[48]; we may assume, since there is on Martin's view "something it is like" to have these experiences, that they have phenomenal characters—in this regard, Martin's approach differs from Fish's;

3. the phenomenal properties of these hallucinations should be typed by indiscriminability properties, that is, by negative epistemological properties. In other words, there is nothing more to the phenomenal character of a causally matching hallucination than the negative epistemic property of being introspectively indiscriminable from a veridical perception: "why did James shriek like that? He was in a situation indiscriminable from the veridical perception of a spider... With no detectable difference between this situation and such a perception, it must seem to him as if a spider is there and so reacts in the same way"[49].

Martin contends that a metaphysical disjunctivist should not search for a more substantive characterization of hallucinations' phenomenal characters than (iii). His negative approach is tailored to eschew what he calls the "screening off" concern:

> Suppose we do get a further specification of the kind of mental event that occurs in the non-privileged circumstances. If what marks these cases out in the first place is just that they involve the absence of perception, then one may worry that whatever fixes what they have in common with each other will apply equally to any case of perception (...). Now if the common element is sufficient to explain all the relevant phenomena in the various cases of illusion and hallucination, one may also worry that it must be sufficient in the case of perception as well. In that case, disjunctivism is threatened with viewing its favored conception of perception as explanatory redundant. Martin (2004, 46).

We see that Martin's motivation is that he wants to avoid the introduction of a common factor that could explain both the phenomenal characteristics of hallucinations and the phenomenal characteristics of veridical perception: "if one allows that there is a more substantive characterization available across a wide range of cases of what it is for mere appearance to occur, the question arises whether such a state can also be present in the case of veridical perception"[50].

[48] Cf. Martin (2006, 375).
[49] Cf. Martin (2004, 68).
[50] Cf. Martin (2006, 370).

Such a common element could exert a preemptive role and "screen off" the relational aspect of perceptual states in the explanations in which these states are mentioned.

If "non-privileged" states are typed according to indiscriminability properties, it is clear that the preemption threat is averted. Let us consider, for instance, the explanation of James' shrieking while hallucinating a spider. On Martin's view, that is only because there is no subjectively detectable difference between James' seeing a spider and James' hallucinating one that James reacts as if a spider was present: nothing beyond the phenomenal properties of "privileged" states needs to be mentioned in the explanation. This is fine as far as the "screening off" problem is concerned, but this leaves some important phenomenological facts unexplained. If we define a class of hallucinations as a class of states introspectively indiscriminable from veridical perceptions, it will be *a priori* true of these episodes that they will be phenomenally similar to veridical experiences. This does not mean, however, that the explanatory challenge we have raised in the beginning of this paragraph has been met: it remains a complete mystery, on Martin's modest approach, why there are states that are introspectively indiscriminable from veridical perceptual states, and why those states seem to be phenomenally similar to veridical states even though they are of a different metaphysical nature.

More generally, Martin's view fails to account for phenomenal similarities that exist between relational and non-relational phenomenal states. Let us consider the case of pictorial experience. Pictures, as Wollheim emphasizes, allow us to enjoy visual experiences of "things that are not present to the senses". Pictorial experience has indeed a dual aspect. When we look at a picture, we see its surface and its properties, but we also see the objects that are depicted. To borrow Wollheim's terminology, we see the depicted objects "in" the picture, in the sense that the visual experience we have of these objects while looking at the picture is very similar to the experience we would have if we were directly seeing the things themselves. Here is the passage in which the dual aspect of seeing-in is explained:

> Seeing-in is a distinct kind of perception, and it is triggered by the presence within the field of vision of a differentiated surface. (...) When the surface is right, then an experience with a certain phenomenology will occur, and it is this phenomenology that is distinctive about seeing-in (...) The distinctive phenomenological feature I call "twofoldness" because, when seeing-in occurs, two things happen: I am visually aware of the surface I look at, and

> I discern something standing out in front of, or (in certain cases) receding behind, something else. So, for instance, I follow the famous advice of Leonardo da Vinci to an aspirant painter and I look at a stained wall, or let my eyes wander over a frosty pane of glass, and at one and the same time I am visually aware of the wall, or of the glass, and I recognize a naked boy, or dancers in mysterious gauze dresses, in front of (in each case) a darker ground. In virtue of this experience I can be said to see the boy in the wall, the dancers in the frosty glass.

We will not commit ourselves to Wollheim's project of defining depiction in terms of seeing-in, but only to the claim that pictorial vision has the twofold phenomenological nature that he identified. The existence of this dual aspect raises a serious difficulty for all versions of the Relational View of experience: it seems that we cannot explain all the phenomenal properties of pictorial experiences by mentioning only the objects present in the surroundings of the subject and their properties.

Let us suppose that you are looking at Chardin's still life with glass flask and fruit. According to the Relational View, only instantiated properties of objects present in the perceived scene can account for the phenomenal properties of your experience. But referring to features of the scene will not be enough to explain the phenomenal character of your experience. For instance, you are having a visual experience of a pear standing on a table and instantiating a certain visual shape. It would be natural to type the corresponding phenomenal property by referring to the specific shape of the pear. Nevertheless, there is no pear in the context of your visual experience: in fact, no three-dimensional object does instantiate the visual shape you are looking at.

You cannot say either that your visual experience of seeing a pear in the picture is indiscriminable from a veridical perception of a pear. This would be plainly false: Chardin's picture is not a trompe-l'oeil, and as a consequence your pictorial experience of the painting does not just replicate the ordinary experience of seeing a pear. According to Wollheim's view on seeing-in, seeing-in involves a simultaneous, conscious awareness of a picture's design and of its representational content[51]. So when you see a pear in Chardin's picture, you both see the picture's surface—two-dimensional shapes, colors, ...—*and* you visually recognize the depicted object as a pear, that is, as a three-dimensional object. So we certainly cannot type the phenomenal character of your expe-

[51] In these regards, it differs from Gombrich's illusion theory of pictorial experience.

rience by indiscriminability properties. Nonetheless, it seems to be a phenomenological fact that this experience is very similar to the experience of directly seeing a pear, and this phenomenological fact has to be explained.

A representational explanation of this phenomenological fact is easy to put forward: it can be assumed that in a visual experience of seeing-in a picture, the visual system *both* registers the properties of the picture's surface and the properties of the depicted objects. According to Mohan Matthen, a picture provides two sets of conflicting cues to the visual system: cues about visual properties of the picture itself—its texture, color, etc...— and cues that are similar to the cues that the depicted object would have provided if it were present. The visual treatment of these cues " lead to two different visual representations that exist side-by-side, though they cannot be attended to simultaneously"[52]. Some recent empirical findings show, consistently with the representational view, that looking at pictures of things puts the visual systems in states that are very much like the states that we are in while seeing the real things[53]. Thus, as Matthen notes, Koenderink and van Doorn have developed an experimental technique that shows that a subject looking at a flat surface depicting a three-dimensional object is able to map the three-dimensional aspects of the object seen in the picture[54]. This entails that the pictorial experience of the subject is an experience as of an object in a three-dimensional space, even if the surface that is directly seen is flat.

According to the Relational View, on the other hand, one obviously cannot account for the phenomenal similarities between pictorial experience and direct vision by mentioning representational properties, nor can one account for the twofoldness of pictorial seeing by appealing to the co-existence of two different kinds of representations. An advocate of the Relational View has therefore to face the following dilemma: either depicted objects are not really seen into pictures; or pictures really instantiate the properties of these objects. Jérôme Dokic, in a recent paper, embraces the second horn of this dilemma:

> When I see Richter's Candle (1982), I do not have any feeling that a candle is present. This in turn has been analysed as entailing that no candle is presented as being located in egocentric space (even

[52] Cf. Matthen (2005, 390).

[53] One can already find this idea in Descartes' *Optics*, where it is stated that pictures "enable the soul to have sensory perceptions of all the various qualities of the objects to which they correspond" and that a picture "causes our sensory perception of these objects". Descartes, R., *Philosophical Writings*. Tr. by J. Cottingham et al., 2 vols. Cambridge: Cambridge University Press, 1985., vol. 1 p. 166.

[54] Cf. Koenderinkand van Doorn (2003, 255).

> though egocentric-spatial notions are relevant to specifying the depicted scene). Still, my visual-recognitional abilities related to candles are actualized in the same way as when I see a real candle. What I want to suggest is that the actualization of these abilities is factive in both the pictorial and the ordinary cases. In the pictorial case, I see part of the picture itself as having the appearance of a candle. More precisely, I see the picture as having the appearance of a candle on its surface, or perhaps in it. There is no illusion here, since the picture really has this appearance, which is perceptually accessible only from a selected set of points of view. (Dokic, 2012, 404).

We see that according to Dokic, a picture of a candle really has the appearance of a candle, where such an appearance has to be constructed as an objective perspectival property. This is not plausible however. Koenderink and van Doorn's experiment shows that subjects have an experience of the three-dimensional properties of depicted objects while looking at their pictures. Does this really imply that those pictures instantiate three-dimensional shapes properties? Where would those three-dimensional properties be instantiated? Let us suppose, for instance, that you are looking at Chardin's still life, and that you visually recognize the three-dimensional shape of a pear. Dokic claims that the appearance of the pear is really instantiated by the painting, but it is not clear to us how this could make sense, at least if we agree that the depicted pear appears to you as a three-dimensional object, and that a picture of a pear is typically not pear-shaped. We can conclude that pictorial visual experience is a counter-example to the Similarity of Objects Principle.

There may be other counter-examples to the Similarity of Objects Principle. Let us consider speech-perception. When we attend to the content of a speech, we consciously perceive phonological structures in the stream of discourse. We are sensitive to perceived similarities, which enable us to group the linguistic sounds into the distinctive units that are known as phonemes. These similarity classes are language-relative. In English, for example, the aspirated "p" in "pen" is perceived as sufficiently similar to the unaspirated "p" in "spun" to be categorized in the same linguistic unit. In other languages, like Thai, Indi or Kechua, there are what linguists call "minimal pairs" of words that are phonologically differentiated only by aspiration. Now on the Relational View, we should find features of the acoustic wave corresponding to those phonological contrasts. This is a consequence of the Similarity of Objects Principle: phenomenal similarities should always be explainable by sim-

ilarities in the objects perceived. As Georges Rey has emphasized in several publications, this prediction of the Relational View is at odds with the findings of contemporary phonology. Here is a typical textbook statement to the effect that there is no correspondance between phonological structures and acoustic structures:

> The stream of speech within a single utterance is a continuum. There are only a few points in this stream which constitute natural breaks, or which show an articulatory, auditory or acoustically steady state being momentarily preserved, and which could therefore serve as the basis for analytical segmentation of the continuum into 'real' phonetic units... The view that such a segmentation is mostly an imposed analysis, and not the outcome of discovering natural time-boundaries in the speech continuum, is a view that deserves the strongest insistence. (Laver, 1993, 101).

According to mainstream phonology, a hearer will typically represent phonological structures in a linguistic sound to which it is not clear at all, as far as we know, that anything real does correspond in the sound wave[55]. Of course this does not make sense on the Relational View. The issue here, it seems to us, is not so much that the predictions of the Relation View contradict our best scientific theories. After all, these theories might be wrong, and we might find structures in the acoustic wave in the future that could be identified with the sound properties presented in hearing linguistic utterances. The problem is rather that on the Relational View, it is *a priori* impossible for conscious sensory states to be systematically illusory: on this view, we can be assured *a priori* that there are real acoustic correlates of phonematic distinctions, even though we have been unable to identify them until now, and even if we have very strong empirical reasons to doubt that such correlates exist. This does not sound plausible at all.

5. Conclusion.

Our strategy in this paper has been to concede that, indeed, perception and hallucination differ in kind. However, we suggest that this difference should be understood as a difference between factive and non-factive first-order experiential states. Subjects of genuine perceptions and subjects of hallucina-

[55] For more on this topic, see (Fodor et al., 1972, 279-313) and (Jackendorff, 1987, 57).

tions are not in the same epistemic position regarding the objects and properties of their environments. As we have argued, this epistemological distinction between perception and hallucination does not entail metaphysical disjunctivism. Our modest disjunctive account is compatible with a representationalist understanding of phenomenal character - be it a weak one - according to which phenomenal properties supervene on a subject's brain state. In turn, this causal thesis concerning phenomenology is incompatible with the thrust of the metaphysical disjunctivist approach (Snowdon, 1981; Nudds, 2009). Representationalism is everything but as capable as disjunctivism to account for the phenomenology of conscious perception. However, it fares better when it comes to giving a positive account of the introspective indiscriminability of perception and hallucination. This explanatory advantage should count as a decisive point in favor of a representational account of phenomenal consciousness.

6. References

Block, Ned (2010) "Attention and Mental Paint", *Philosophical Issues*, vol.20, 23-63

Broad, Charlie Dunbar (1952) "Some Elementary Reflexions on Sense-Perception", Philosophy 27, 3-17. Reprinted in Schwartz, R. J. (ed.) (1965) *Perceiving, Sensing and Knowing*, Berkeley, University of California Press.

Byrne, Alex and Logue, Heather (2008) "Either/Or", in Haddock, A. and Macpherson, F., (2008) 57-94.

Byrne, Alex and Logue, Heather (eds.) (2009) *Disjunctivism. Contemporary Readings*, Cambridge, Ma., MIT Press.

Campbell, John (2002) *Reference and Consciousness*, Oxford, Clarendon Press.

Crane, Tim (2006) "Is There a Perceptual Relation", in *Perceptual Experience*, Gendler T. S. and J. Hawthorne (eds.), Oxford and New York, Oxford University Press, 126-146.

Dokic, Jérôme (2012) "Pictures in the Flesh: Presence and Appearances in Pictorial Experiences", *British Journal of Aesthetics*, vol.52 (4), 391-405.

Dokic, Jérôme & Martin, Jean-Rémy (2012) "Disjunctivism, Hallucinations, and Metacognition", *WIREs Cognitive Sciences*, vol.3, 533-543.

Dretske, Fred (1995) *Naturalizing the Mind*, Cambridge, Mas., MIT Press.

Engel, Pascal (2007) *Va savoir ! De la connaissance en général*, Paris, Hermann.

Fish, William (2009) *Perception, Hallucination, and Illusion*, New York, Oxford University Press.

Fish, William (2004) "Disjunctivism and Non-Disjunctivism: Making Sense of the Debate", *Proceedings of the Aristotelian Society*, vol.105, 119-127.

Haddock, Adrian and Macpherson, Fiona (eds.) (2008) *Disjunctivism: Perception, Action, Knowledge*, New York, Oxford University Press.

Harman, Gilbert (1990) "The Intrinsic Quality of Experience", *Philosophical Perspectives*, vol.4, p31-52.

Hawthorne John and Kovakovich Karson (2006) "Disjunctivism", Proceedings of the Aristotelian Society, supp. vol. 80, 145-183.

Hellie, Benj (2007) "Factive Phenomenal Characters", *Philosophical Perspectives*, 21, Philosophy of Mind, 259-306.

Hill, Christopher S. (2009) *Consciousness*, Cambridge, Cambridge University Press.

Hinton, John (1967a) "Experiences", *Philosophical Quarterly*, 17, 1-13.

Hinton, John (1967b) "Visual Experiences", *Mind*, vol.76, 217-227.

Hinton, John (1973) *Experiences: An Inquiry into Some Ambiguities*, Oxford, Clarendon Press.

Johnston, Mark (2004) "The Obscure Object of Hallucination", Philosophical Studies 120, 113-183.

Kind, Amy (2003) "What's so Transparent about Transparency", *Philosophical Studies*, vol.115(3), 225-244.

Koenderink, Jan J. & Van Doom, Andrea (2003) "Pictorial Space", in Hecht, H., Schwartz, R. & Atherton, M., *Looking into Pictures: An Interdisciplinary Approach to Pictorial Space*, 239-403.

Laver, John (1993) *Principles of Phonetics*, Cambridge: Cambridge University Press.

Levine, Joseph (2006) "Conscious Awareness and Self-Representation", in Kriegel, U., and Williford, K. (eds.) *Self-Representational Approaches to Consciousness*, Cambridge, Ma., MIT Press.

Lewis, David (1999) *Papers in Metaphysics and Epistemology*, Cambridge, Cambridge University Press.

Martin, Michael (2002) "The Transparency of Experience", *Mind and Language*, vol.17(4), 376-425.

Martin, Michael (2004) "The Limits of Self-Awareness", *Philosophical Studies*, vol.120, 37-89.

Martin, Michael (2006) "On Being Alienated", in Gendler & Hawthorne, (Eds.) *Perceptual Experiences*, Oxford, Oxford University Press, p354-410.

Mathen, Mohan (2005) *Seeing, Doing and Knowing: a Philosophical Theory of Sense Perception*, New York, Oxford University Press, 2005.

McDowell, John (1982) "Criteria Defeasibility and Knowledge", *Proceedings of the British Academy*, vol.68, 369-394.

McDowell, John (2002) "Knowledge and the Internal Revisited", *Philosophy and Phenomenological Research* 64, 97-105.

Nudds, Matthew (2009) "Recent Works in Perception: Naive Realism and its Opponents", Analysis Reviews, vol.69(2), 334-346.

Pautz, Adam (2013) "Do the Benefits of Naïve Realism Outweigh the Costs? Comments on Fish, Perception, Hallucination and Illusion", Philosophical Studies 163, 25-36.

Putnam, Hillary (1999) *The Threefold Chord: Mind, Body and World*, New York, Columbia University Press.

Pritchard, Duncan (2012) *Epistemological Disjunctivism*, Oxford, Oxford University Press.

Smith (2002) *The Problem of Perception*, Cambridge, Ma, Harvard University Press.

Snowdon, Paul (1981) "Perception, Vision and Causation", *Proceedings of the Aristotelian Society* (new series) 81, 175-192.

Snowdon, Paul (1990), "The Object of Perceptual Experience", *Proceedings of the Aristotelian Society* (suppl. vol.) 64, 121-150.

Soteriou, Matthew (2009) "The Disjunctive Theory of Perception", *Stanford Encyclopedia of Philosophy*, ed. E. Zalta, http://plato.stanford.edu/entries/perception-disjunctive/.

Strawson, Peter F. (1979) "Perception and its Objects", in Graham Macdonald (ed.), *Perception and Identity: Essays Presented to A. J. Ayer with His Replies,* London. Macmillan.

Sturgeon, Scott (2000) *Matters of Mind*, London, Routledge.

Tye, Michael (2009) *Consciousness Revisited*, Cambridge, Ma., MIT Press.

Varese, Filippo & Laroi, Frank (2012) "Misattributions Models (I): Metacognitive Beliefs and Hallucinations", in Jardi, R. & Cachia A. (Eds.), *The Neurosciences of Hallucinations*, Springer.

Williamson, Timothy (2000) *Knowledge and its Limits*, New York, Oxford University Press.

33

Explaining Reference: A Plea for Semantic Psychologism[1]

SANTIAGO ECHEVERRI

There is a traditional opposition between two camps in the theory of reference: descriptivism *vs.* referentialism or Fregeanism *vs.* Russellianism. Before one chooses which view—if any—one favors, a more fundamental issue should be addressed: Can we *explain* reference? Whereas 'full-blooded' theorists answer 'yes,' so-called modest theorists answer 'no.'[2]

Since the notion of *explanation* can be understood in different ways, there is no single full-blooded/modest contrast that describes all possible views. One could be a full-blooded theorist in relation to one of these concepts of explanation but a modest one in relation to the others.

One might mean by 'explanation' some form of reductive *conceptual analysis.* So the explanatory task would consist in offering a non-circular analysis of key concepts like reference, denotation or truth (Dummett 1975, 1976, 1991; Engel 1989, 1994).

The notion of explanation might also amount to an *ontological reduction* of semantic properties to non-semantic properties like causation, information, biological functions, and so on (Field 1972, 1978; Dretske 1981; Millikan 1984).

[1] Jérôme Dokic read a previous version of this paper, and made some insightful comments. Work on this article was funded by research grant no. 100015_131794 of the Swiss National Science Foundation.

[2] Dummett (1975, 1976, 1991) first defended (a version of) the full-blooded theory of meaning. 'Modest' theories are associated with Davidson (1984) and McDowell (1977, 1987, 1997), among others.

One might also drop reductionism, and conceive of the explanatory program as the enrichment of a purely extensional semantic framework with intensions, characters, possible worlds, etc. (Engel 1989).

Still, one might take the explanatory program as a meta-semantic one seeking to uncover the *historical* and *social facts* that explain why some expressions have the meaning they presently have (Devitt 1981; Kaplan 1989; Almog 2005).

Some of these ways of cashing out the full-blooded program bear interesting relations to each other. If you are a conceptual reductionist about reference, you will probably reject historical accounts of reference such as Kripke's (1980), which uses notions like meaning intentions to account for the transmission of reference (Dummett 1973: 148-9; McKinsey 2009). It is however desirable to keep these different concepts of explanation separate before one makes up one's mind on whether one should try to explain reference.

Pascal Engel has rejected various forms of modesty during his prolific career. Yet, given the different ways in which one can construe the notion of explanation, it would be misleading to qualify him as a full-blooded theorist. He has been sympathetic to Davidson's anomalism of the mental, which implies that intentional properties cannot be reduced to physical properties, and has also expressed serious doubts on teleosemantic accounts of semantic content (Engel 1996). Although he has sometimes presented the full-blooded program as some version of conceptual analysis (Engel 1989, Ms.), I would be reluctant to describe him as favoring some form of non-circular factorization of the concept of reference. Still, Engel has been a critic of the sort of modesty championed by philosophers like McDowell and some Wittgensteinian philosophers (Engel 1996, 2001). So, if Engel is a full-blooded theorist, his 'full-bloodedness' does not apply to all the notions of explanation that have been used to characterize full-blooded positions.

An important strand of Engel's critique of modesty can be found in his works on the philosophy of logic. In his seminal book *La norme du vrai: Philosophie de la logique* (1989), he used the debate between Davidson and Dummett on the nature and scope of a theory of meaning as the bedrock to introduce a number of topics in the philosophy of logic. In that book, he granted some of Davidson's constraints on a good theory of meaning, and defended a weak form of psychologism. In the next years, he pursued these two issues in more depth. In his thèse d'état *Davidson et la philosophie du langage* (1994), Engel offered a systematic reconstruction of Davidson's program, arguing that it is compatible with a psychological account of language mastery in the spirit of Chomsky's theory of tacit knowledge. Two years later, he published *Philoso-*

phie et psychologie (1996), a sustained attack on the dogma that psychologism must be avoided at any cost. This led him to defend a 'healthy psychologism,' and reject the pervasive idea that the realm of norms is wholly disconnected from the realm of causes.

All these works constitute an attempt at finding a middle way between some form of modesty and what I would call 'psychological full-bloodedness.' Engel offers a nice formulation of his proposal within Frege's distinction between the world of psychological facts (world 2) and the world of objective truths (world 3). According to Engel (1996: 121-ff.), it makes good sense to introduce a world 2½, i.e. a place where normative and natural properties are in contact. The notion of a world 2½ expresses his conviction that any theory of norms should accommodate two seemingly conflicting data. On the one hand, the psychology of reasoning has shown that subjects may reason in ways that contradict the rules of classical logic. On the other, it is clear that the same subjects who seem to reason in a non-classical way developed classical logical systems, and used them to evaluate and criticize their own patterns of reasoning (Engel 1989: Chapter XIII; 2006). Hence, world 2½ expresses the idea that, although logic does not *describe* the actual procedures used by ordinary subjects in normal reasoning, it is not part of a disconnected 'third realm' of objective truths. Engel sees world 2½ as enabling us to specify the kinds of processes that ought to govern the mental life of humans. But the idea goes beyond that: It suggests that the links between the psychological and the logical are intimate but less direct than some forms of psychologism might take them to be. Hence the label 'healthy psychologism.'

The topic of this paper is not the philosophy of logic but the theory of reference. Yet, my project is tightly related to Engel's lucid defense of world 2½. I do not have any reductionist agenda, nor do I believe in the program of conceptual analysis. Moreover, I will not give arguments for the enrichment of semantics with intensions or possible worlds, nor for a historical account of reference. Still, I will defend a healthy psychologism in the theory of reference. Thus, I might be taken to be pursuing a full-blooded program. I call this brand of the full-blooded program 'semantic psychologism.' Semantic psychologism is the conjunction of a negative and a positive claim:

> **Negative claim:** Pursuing the program of truth-conditional semantics is insufficient to account for semantic competence, for it does not provide a sufficiently illuminating account of referential abilities.
>
> **Positive claim:** There are good reasons to supplement the program

> of truth-conditional semantics with a *psychological* account of referential abilities, i.e. an account that is framed at an intermediary level of description between the personal level and the explanations provided by neuroscience.

As I understand it, semantic psychologism is orthogonal to any historical or social account of reference. Whereas the latter deal with reference *qua* property of expressions in a public language, semantic psychologism seeks to provide an account of the abilities needed to make use of a language. In other words, one does not need to see historical or social accounts of language as *competitors* to semantic psychologism. Its real opponent is the anti-psychologism that dominates various forms of philosophy of language that construe it as a branch of the philosophy of logic or reject any conception of psychology as an explanatory discipline.

My defense of semantic psychologism will have two parts. First, I will present some of the main reasons why the program of truth-conditional semantics is modest in a relevant sense: it does not offer the explanations we want when we are interested in offering an account of referential abilities (sections 1-2). Second, I will respond to some influential considerations against a psychological explanation of reference (sections 3-8).

1. The Explanatory Limits of Truth-Conditional Semantics

Frege's doctrine of *Sinn* is the first formulation of a truth-conditional semantics. Some philosophers reject the notion of *Sinn*, however, because of its dubious ontological pedigree. Yet, one can partially circumvent this problem by pursuing a 'functional' approach. On a functional view, the notion of *Sinn* is introduced by its *theoretical role*. The *Sinn* of an expression *E* is construed as a *solution* to some of the key questions that arise in the study of *E*.[3]

The central role of a theory of *Sinn* is to provide a truth-conditional account of the meaning of whole sentences. On this approach, the *Sinn* of a whole sentence *S* (a *Gedanke*) expresses its truth-conditions, and the *Sinn* of its constituent expressions $E_1, E_2, \ldots E_{n-1}, E_n$ is their contribution to the truth-conditions of *S*.

How does the truth-conditional role of *Sinn* relate to the more general debate that opposes full-blooded to modest conceptions of meaning? My first claim is that, even if the truth-conditional aspect is a necessary ingredient of a

[3] This approach is implicit in Burge (1977: 59) and McDowell (2005).

full-blooded theory, it is not sufficient to explain reference. There are at least two reasons why offering a truth-conditional account of language is not sufficient to explain reference. First, a theory that only specifies truth-conditions would leave out some central *explananda* in the theory of meaning. Second, truth-conditional semantics exploits a very thin notion of explanation, i.e. a notion that is not sufficiently illuminating for philosophical purposes. I develop these two ideas in the next sub-sections.

The Determination of Reference

If one takes the notion of *Sinn* as spelling out the contribution of any expression-type E_i to the truth-conditions of sentences in which E_i may occur, it seems possible to frame Frege's theory of *Sinn* as an axiomatic theory.[4] The main textual support for this reading can be found in § 32 of *Grundgesetze der Arithmetik*, where Frege makes clear that his stipulations of the meanings of names convey their *Sinn*.[5] Building on this minimal conception of *Sinn*, one can specify the contribution of singular terms and predicate expressions by means of a list of axioms. Let us consider the following two examples:

(A1) 'Hesperus' stands for (or denotes) Hesperus.

(A2) '*x* is agile' is true of something if and only if it is agile.[6]

These axioms satisfy the first characterization of the notion of *Sinn.* Yet, there is a sense in which they are not explanatory. Given that 'stands for' and 'is true of' are *used* in the axioms, these specifications do not explain *how* 'Hesperus' refers to Hesperus, nor *why* the predicate '*x* is agile' is true of some things but not of others. Dummett (1975) offers an influential argument in favor of this claim: one could imagine a subject who understands quotation devices but is unable to *use* 'Hesperus' to refer to Venus, and could not *identify* Venus in the sky. So, even if the first axiom articulates the contribution of the name 'Hesperus' to the truth-conditions of the sentences in which it may occur, it is not sufficient to explain what it is to *understand* the word 'Hesperus.' (See also Engel 1989)

[4] See McDowell (1977), Evans (1981a, 1981b), and Sainsbury (2005). This notion fits Lewis' (1980) characterization of 'semantic value': "If they [semantic values] don't obey the compositional principle, they are not what I call semantic values." (35)

[5] See Frege (1893: § 32, 50-1). For a recent reading in this sense, see Kripke (2008: 182-ff.).

[6] See McDowell (1977, 1997). I set aside some complications with predicates. One should allow their argument places to be saturated either by names or variables. I also omit the qualification that these axioms are relative to a particular language, in this case, English.

A highly influential way of understanding the contrast between modest and full-blooded theories of meaning is therefore to claim that, for modesty, it is not possible to provide a more fundamental characterization of reference than the one provided by axioms that use semantic expressions like 'denote,' 'refer,' 'stand for,' and so on. If one thinks that truth-conditional specifications are all there is to explain the reference of the primitive vocabulary of a language, one holds a modest view. By contrast, if one thinks that one can provide an informative account of the notions of denotation, reference, satisfaction, etc., one will hold a full-blooded view.

A popular way of cashing out this distinction is to say that Frege's notion of *Sinn* not only plays a truth-conditional role but also more substantial roles. One of these roles is to offer a *method* by which speakers can determine the semantic value of linguistic expressions.[7] Thus, the notion of *Sinn* can be seen in the service of another question: What connects the name with the referent? Or, as Almog puts it: What is the chemistry of the bond between a singular expression and the referent?[8] If one thinks these questions are meaningful, one must find the axioms stated above too austere. In order to provide explanations, one cannot just *state* the axioms of the semantic theory; one has to *explain* how the names mentioned and used in these axioms connect to entities in the world.[9]

Although this approach to the non-sufficiency claim is widespread, it leaves a number of questions open. First, the equation of the notion of *Sinn* with a method has some verificationist connotations, and might convey an unfaithful idea of the referential use of words. When I use a telescope to watch the craters in the moon, I certainly use a method to single out the craters. But it is unclear whether language or concepts work as a method in this instrumental sense. Second, the equation of the notion of *Sinn* with a method raises questions concerning the links between language and thought. If one assumes that the axiomatic theory is a model of *public* language, one shall probably need to distinguish the *public Sinn* articulated in the axiomatic reconstruction of a language from the *private means* speakers use to determine the semantic values of words. As far as proper names are concerned, some theorists have held that speakers use *different routes* to determine their semantic values, even though

[7] See Dummett (1973: 93, 95; 1978: 119-20). For a critical discussion, see Evans (1982: 17-ff.).

[8] See Almog (2005: 498) and Wettstein (2004: 110).

[9] I am assuming that the notion of explanation is psychological. As indicated in the introduction, this is not the only relevant notion of explanation. One might require reductive, historical or social explanations. For reasons of space, I will not deal with these other options here.

there would be a 'shared' or 'public' *Sinn* of those words.[10] I won't explore these possibilities here.[11]

Truth-Conditional Semantics as Instrumental Explanation

Another way of showing that the truth-conditional approach is not sufficient to explain reference is to focus on its underlying notion of explanation, and show that it is not sufficiently illuminating from a philosophical perspective. When one lays down the axioms specifying the semantic value of an expression-type E_i, one is assuming that the axioms have 'projectible predicates' in Goodman's (1965) sense. These predicates occur in inductive statements, so they express counterfactual-supporting generalizations.[12] Thus, contrary to what might be initially thought, when one lays down the axioms of a fragment of a language, one is not merely *describing* it. If the axioms are compositional, they constitute a system of interrelated principles (Davies 1987).

The explanations provided by the axioms can be controlled for their correctness. On the one hand, theorists can check whether their assignments *cohere* with speakers' intuitions on the truth-conditions of sentences. On the other, those assignments enable theorists to make some *predictions* on the behavior of expression-types. Hence, one can compare competing axioms by their 'explanatory potential'; one has only to examine their intuitive adequacy and the predictions they make. This approach is instrumentalist because semantic theories work as predictive devices. The axioms capture projectible *regularities* in the behavior of expression-types.

Normally, issues of validity provide the main tests of these investigations: One asks which inferences are made valid under a semantic assignment, and

[10] As a result, Burge (1990) distinguishes objective sense from the subject's grasp of sense. Since a speaker might have a partial (or mistaken) grasp of sense, this would lead to a further distinction between the idiosyncratic and the public aspects of language. For a related distinction, see Higginbotham (1998). In a different perspective, Millikan (2005) distinguishes the conventional meaning of words from subjects' conceptions, which correspond to the methods a speaker uses to track the referent.

[11] See Frege (1918-1919). On the basis of this text, Kripke (1979, 2008) urges that, as far as proper names are concerned, Frege is committed to idiolects (not public languages) as the main units of analysis.

[12] There is a stronger conception of *projectible* predicates as those that enter into *laws*. But it would be controversial to assume that truth-conditional semantics formulates laws in any non-trivial sense of that term. It is less controversial, however, that truth-conditional semantics treats expression-types as *natural kinds*.

which turn out to be invalid.[13] Crucially, one can reject a semantic proposal if it fails to accommodate intuitively valid patterns of reasoning or if it counts as valid some pieces of reasoning most competent speakers intuitively count as invalid. The following inference illustrates this strategy:

P1 Jones believes that all men are created equal.

P2 Smith doubts that all men are created equal.

$\therefore$ Thus, there is something that Jones believes but Smith doubts.

The instrumentalist theorist is interested in laying down semantic rules capable of specifying the behavior of 'that'-clauses in such a way that inferences of this sort go through. Crucially, one could discard some accounts because they cannot accommodate this intuitively valid inference. An influential solution is to treat the 'that'-clause as a sentential operator. When it is conjoined with a sentence, it forms a singular term that designates a proposition. This view yields a valid inference. Let us use 'S' as a variable ranging over persons, 'B' as a belief operator, 'D' as a doubt operator, and 'p' as a variable for propositions. So we have:

P1 B(s, p)

P2 D(s, p)

$\therefore$ $\exists x$: x is a proposition p & (B(s, x) & D(s, x))[14]

If this sort of explanation is all there is to explain reference, one may count as a modest theorist. I assume Engel would agree with this. In a recent review of LePore & Ludwig's *Davidson's Truth-Theoretic Semantics*, he praises the development of specific proposals within Davidson's semantic program. Nevertheless, he also expresses "some nostalgia for the pioneering efforts of the 1970s." (Engel 2007) As philosophers, we may want to understand how such concepts as reference and truth fit within a natural world. If we are interested in linguistic competence, we may also want to tell a story about the referential abilities underlying the use of linguistic expressions. In the next section, I develop this point in some detail.

[13] Other authors have drawn similar distinctions. Rorty (1979: Chapter 6) calls 'pure semantics' a project similar to what I am calling instrumentalism, and contrasts it with what he calls 'impure semantics.' Barwise & Perry (1983: lxxv) term 'thin semantics' a project similar to instrumentalism. Recently, Predelli (2005: Chapters 1 and 4) mounted a defense of *thin* semantics.

[14] I have borrowed the example from Salmon (1986: 6). This kind of reasoning is pervasive in contemporary philosophy of language.

2. The Psychological Program

Let us take stock. Spelling out the contribution of an expression-type *E* to the truth-conditions of sentences in which *E* may occur might be a necessary component of a theory of reference.[15] Yet, it is insufficient. First, the truth-conditional account would leave out the determination of reference and, second, it would rely on a very 'thin' conception of explanation as prediction. If we are interested in linguistic competence, we might be interested in elucidating the psychological abilities underlying the mastery of a language.

In this section, I focus on the expansion of the instrumentalist program by means of a psychological account of referential abilities. I will remain neutral on the exact form of such a psychological program. I will simply show that the instrumentalist notion of explanation at work in truth-conditional semantics presupposes the possession of more fundamental referential abilities that may require a psychological account.

One can clarify the notion of explanation proper to semantic psychologism by examining the requirements to *determine* the referent of any expression-type. In the context of a truth-conditional reconstruction of a fragment of a language, there is a sense in which the semanticist determines the semantic value of an expression-type such as a proper name 'NN.' In this case, to determine the semantic value of 'NN' is to furnish a mathematical function that maps occurrences of 'NN' onto its semantic value, e.g. a referent. This sense of reference determination is implicit in some of Frege's remarks on the notion of *Sinn*. Consider the following excerpt from "Funktion und Begriff":

> It is thus necessary to lay down rules from which it follows, e.g., what '☉ + 1' stands for, if '☉' should (*soll*) stand for the Sun. What stipulations we lay down is indifferent; but it is essential that we should do so—that 'a + b' should always get a *Bedeutung*, whatever signs for determinate objects may be inserted in the place of 'a' and 'b.' (Frege 1891: 19-20; translation modified)

According to Frege's example, reference determination can be a stipulation. In a formal system, one can determine the referent of '[F080?]' by providing a rule that assigns to it one and only one referent. To determine a referent of a sign is an *obligation* any theorist incurs when she is laying down the axioms

[15] It would take us too far afield to examine the reasons why most theorists take the truth-conditional program as a necessary component of a systematic theory of meaning. My main concern here is with the non-sufficiency claim.

of the system (Frege 1893: XII; see also Heck 2002: 4). This obligation is also incurred when one tries to formalize a fragment of a natural language. The stipulative sense also captures some of the ways in which Kripke illustrates his notion of 'reference fixing.' He grants that, at least in some cases, one can perform some ceremonies in which one stipulates the meaning of a proper name, as when a policeman in London declares: "By 'Jack the Ripper' I mean the man, whoever he is, who committed all *these* murders, or most of them." (Kripke 1980: 79)[16]

In order to stipulate the referent of an expression, one needs an *independent way* of referring to the target referent. Thus, when Frege stipulates that '[F080?]' stands for the sun, he assumes that the audience has independent means to single out the sun, e.g. iconic memories of the sun and the word 'sun.' Similarly, in order to stipulate that 'Jack the Ripper' refers to the man, whoever he is, who committed all these murders or most of them, the policeman must have an independent way of referring to that murderer. This clearly shows that explaining the determination of reference does not stop when the semanticist (or the policeman) 'fixes' the referent of a name in the stipulative sense. We still need an account of the *prior* referential abilities that enable them to make some stipulations, and the audience to understand those stipulations. This asks for a different kind of explanation.

This remark suggests that something along Engel's lines could be correct in the theory of reference. Recall that Engel advocates a healthy psychologism, and believes that the notion of tacit knowledge employed in cognitive science is more respectable than its Wittgensteinian critics assume. Yet, one might wonder whether the previous remark would be sufficient to *vindicate* the project of providing a *psychological* account of referential abilities. After all, from the fact that semantic stipulations presuppose prior referential abilities it does not follow that one can *explain* those abilities. Whereas a semantic psychologist might be optimistic about the prospects of the explanatory task, the semantic anti-psychologist might be inclined to take those referential abilities as primitive. In what follows, I defend semantic psychologism. To this end, I respond to an influential series of considerations against it. I examine the arguments in order of increasing relevance and plausibility.

[16] See also Evans' (1982: 50) example of Julius, the inventor of the zipper, and Kripke's (1980: 55) remarks on the standard meter in Paris.

3. Our Present State of Ignorance

In his influential paper "Index, Context, and Content," Lewis describes a grammar "as part of a systematic restatement of our common knowledge about our practices of linguistic communication." (1980: 21) Later on, he makes clear that he does not envisage providing a psycholinguistic theory:

> The subject might be differently delineated, and more stringent conditions of adequacy might be demanded. You might insist that a good grammar should be suited to fit a psycholinguistic theory that goes beyond our common knowledge and explains the inner mechanisms that make our practice possible. There is nothing wrong in principle with this ambitious goal, but I doubt that it is worthwhile to pursue it in our present state of knowledge. (Lewis 1980: 24)

Lewis' remark suggests that our future state of knowledge might enable us to offer an explanation of the inner mechanisms "that make our practice possible." Still, he also thinks that our present state of knowledge does not warrant any psychologically oriented formulation of a theory of a language.

There is a sense in which Lewis' remark is right: we lack the required *empirical* evidence to formulate a detailed psychological account of referential abilities. Nevertheless, this is not sufficient reason to set aside semantic psychologism. After all, any explanatory program requires prior conceptual ground clearing, and philosophers are particularly good at that task. Moreover, science requires the formulation of theories capable of organizing the available findings and generating new predictions. So, it would be a mistake to 'wait' until we gather more data before we decide to pursue semantic psychologism. If our goal is to offer a psychological account of referential abilities, it is already necessary to make some conceptual work to guide empirical research.

The philosophically interesting question is: What general form could a future cognitive theory of referential abilities take? There are at least two ways of answering this question. According to optimism, there will be a *psychological* theory of reference. According to pessimism, there will be no such thing as a psychological theory of reference. The only cognitive account we will get, if we get anything properly called 'cognitive,' will be provided by neuroscience.

It is the latter possibility that threatens semantic psychologism, for it undermines it on principled grounds. And, if there cannot be a psychological account of referential abilities, psychological modesty is not only reasonable

but also mandatory. In what follows, I respond to some in-principle arguments against semantic psychologism.[17]

4. The Quietist Stance

A prominent line of attack derives from a quietist conception of philosophy. On this view, providing a psychological account of reference is a bad idea because philosophy is not a theoretical enterprise. This outlook is implicit in Kripke's version of the historical account of reference. He insists that reference is maintained if speakers *intend* to use names with the same referent. Hence, his picture does not explain the required referential intentions. Crucially, this is a problem only if one evaluates the historical view as a theory. He dismisses this reading, though: "You may suspect me of proposing another theory in [the] place [of the cluster theory of names]; but I hope not, because I'm sure it's wrong too if it is a theory." (Kripke 1980: 64)[18]

I am unable to see how I could grant a conclusion based on overarching premises on what philosophy is (or ought to be). But maybe some philosophical issues cannot be adequately tackled by constructing theories. What is special about reference that might preclude the formulation of psychological theories thereof? An extreme view says that the problem is not related to reference *per se* but to the project of formulating a theory of the *mind*. On this view, psychology is not a science of the mind in the same way as astrology is not a science of destiny. There is a version of this claim in some of McDowell's seminal remarks in favor of a modest account of reference:

> There is no merit in a conception of the mind that permits us to speculate about its states, conceived as states of a hypothesized mechanism, with a breezy lack of concern for facts about explicit awareness. Postulation of implicit knowledge for such allegedly explanatory purposes sheds not scientific light but philosophical darkness. (McDowell 1977: 180)

A less extreme view holds that, although there is a psychological science of the mind, there is no *mechanistic explanation* of referential abilities within a science

[17] The following discussion is congenial to Engel's (1996) lucid defence of a healthy psychologism.

[18] I owe this point to McDowell (1977: 198). In his paper "Speaking of Nothing," Donnellan (1974: n 3) is quite clear that he "wants to avoid a seeming commitment to all the links in the referential chain being causal." As Wettstein points out, we should distinguish the idea of a chain of communication from the more committal idea of a causal theory of reference.

of the mind. Interestingly, a prominent cognitive scientist, Zenon Pylyshyn, holds this view. He thinks that the mind requires some form of *direct reference* analogous to an index. Yet, he also thinks that there is no cognitive account of how indices refer. This leads him to hypothesize that a correct account of how indices work falls "under an architectural or neuroscience vocabulary." (Pylyshyn 2007: 39, 82)

Unfortunately, Pylyshyn's view does not lend support to McDowell's more radical statement, which impugns the very idea of hypothesizing mental mechanisms. McDowell (and many others) has a general picture of the mind that prevents him from approving the development of psychological accounts of reference. The next sections explore the most prominent ways of defending this form of semantic anti-psychologism.

5. Reference and the Vehicle-Content Distinction

In subsequent work, McDowell presents a more specific attack on the program of explaining reference. The attack does not have the form of a rigorous argument but is offered as a collection of suggestive considerations.

Here is the main line of thought. In order to formulate the problem of reference, one has to introduce a dichotomy between two aspects of meaningful entities: a *physical* and a *semantic* aspect. Once a sharp line between these two aspects is drawn, we create a gap that cannot be bridged. Since the gap cannot be bridged, we are left with the feeling that reference is a very deep problem that lacks any intelligible solution. If we reject the underlying dichotomy, however, the problem of reference should not arise. Thus, instead of trying to *solve* the problem of reference, we should try to *dissolve* it by rejecting the dichotomy presupposed in its formulation. We find these considerations in McDowell's comments on Putnam:

> Putnam has often expressed suspicion of the idea that there is good philosophy to be done by grappling with questions like 'How does language hook on to the world?' It ought to be similar with questions like 'How does thinking hook on to the world?' Such a question looks like a pressing one if we saddle ourselves with a conception of what thinking is, considered in itself, that deprives thinking of its characteristic bearing on the world—its bearing about this or that object in the world, and its being to the effect that this or that state of affairs obtains in the world. If we start from a conception of thinking as in itself without referential bearing on the world, we

> shall seem to be confronted with a genuine and urgent task, that of reinstating into our picture the way thinking is directed at the world. But if we do not accept the assumption that what thinking is, considered in itself, is a mental manipulation of representations in Putnam's sense [as vehicles], no such task confronts us. The need to construct a theoretical 'hook' to link thinking to the world does not arise, because if it is thinking that we have in view at all—say being struck by the thought that one hears the sound of water dripping—then what we have in view is *already* hooked on to the world; it is already in view as possessing referential directedness at reality. (McDowell 1992: 288)

The problem of explaining reference arises from a *dualistic* understanding of the distinction between the vehicles of representation and their contents. On this view, one could eventually have a physical vehicle, let us say the ink mark **'Aristotle,'** which could fail to refer to the famous philosopher.[19] This makes the problem of reference puzzling. If that ink mark is *just* a physical pattern with no referential power, how can it refer to Aristotle? If one understands the relation between vehicle and content as intrinsic, however, the question appears to be empty.

Consider an analogy. One could use a piece of bronze to make a statue but also to create many other things. Still, there seems to be a difference between one's having a piece of bronze and one's having a statue. How is it possible that a piece of bronze can also be a statue? This sort of problem seems to rest upon a mistaken assumption on how the piece of bronze is related to the statue. Although the latter is made of bronze, the constitution of the statue *qua* statue does not depend on its matter alone but also on its design. Since a description of the piece of bronze as a statue belongs to a specific level of description, it makes little sense to ask a question about the constitution of the statue by remaining at the lower level of description that considers the statue as a mere piece of bronze. By parity of reasoning, if you focus on the purely physical side of any sign, you will be unable to explain how it manages to refer to something in the world. If you focus on the vehicle *qua* bearer of meaning, however, the problem of reference should not arise.

Howard Wettstein develops a similar line of thought. He compares the problem of reference to Descartes' puzzlement about the locomotive capacities of some bodies:

[19] I am using **'boldface'** quotation to refer to types of physical entities.

> Descartes says that he found it amazing that bodies, mere pieces of nature, could move themselves. If locomotion can seem miraculous, what about reference? That mere pieces of nature can mean, or symbolize, or stand for something really seems extraordinary. (Wettstein 2004: 104)

According to Wettstein, Descartes' puzzlement is based on a dualistic assumption. It is certainly extraordinary that animals are capable of locomotion. Still, if we describe their bodies as *merely extended entities* (like stones or pieces of clay), there is nothing we could 'add' to explain how they self-move. That is probably why one might be led to posit a mysterious soul to explain their locomotive capacities. But positing a soul merely explains the obscure by the more obscure.

Similarly, if we describe sound patterns as *mere* physical vehicles, how can we explain reference without adding something similar to a soul? Only by adding something intrinsically significant, such as a Fregean sense or a description, can we explain their reference. But this strategy will be explanatory only if we already understand what it is for an immaterial entity to be endowed with *intrinsic meaning.* And the same will occur if we try to derive reference from mental states. Our account will only work if we are prepared to swallow the idea of an *intrinsic* intentionality (Searle 1983). But very few people seem to understand what it means to be intrinsically intentional (see, e.g., Clark 2005). Wettstein exploits these considerations to promote philosophical modesty:

> Why not leave things where we found them? This suggests—and this might seem at least mildly depressing (but not to worry, it grows on one)—that perhaps the best we can do is to describe our ways with language, making no attempt to go beyond or behind. (Wettstein 2004: 106)

One might wonder whether this argument *shows* that semantic psychologism is hopeless. I do not think so. It merely *suggests* that, given a particular metaphysics of mind and language, one cannot provide an intelligible account of reference. But this falls short of undermining semantic psychologism, or so I shall argue.

Most contemporary philosophers certainly think that there is a vehicle-content distinction in the mental realm (but see Sedivy 2004, for criticism). This might lead one to think that this distinction is a necessary presupposition

of the problem of reference. Nevertheless, one can still formulate the problem of reference without presupposing a conception of the mind that takes that distinction for granted. Consider first the case of public language. It is an uncontroversial fact that public signs are *conventional.* Because signs are conventional, it is legitimate to draw the vehicle-content distinction to characterize them. One can use different physical marks such as **'red,' 'rot,' 'rouge,' 'rojo'** to denote one and the same property: REDNESS. This clearly suggests that signs are *arbitrarily related* to their content. Meaning is not an intrinsic property of any physical shape. Hence, trying to explain this relation, as full-blooded theorists try to do, is legitimate. For any word 'W' one may ask: How did the physical pattern **'W'** come to denote the entity it denotes (or express the property it expresses)? What conditions must obtain for a physical pattern to be meaningful?

Similar remarks apply in the cognitive realm. Even if one rejects the application of the vehicle-content distinction to the mental realm, there is something analogous to the arbitrariness of signs in the psychological structures that underlie our mastery of a language. Imagine a pair of biological twins, one of them raised in England, and the other in Spain. Even though they may be physical duplicates, one of them will learn to use **'red'** to denote RED, while the other will learn to use **'rojo'** to denote the same property. Which specific referential abilities each of them acquired is, to some extent, accidental. One and the same physical substratum could be used to realize different referential abilities. It is therefore legitimate to ask: How can the same types of physical substrata embedded in different environments come to realize different referential abilities?[20]

Moreover, from a developmental perspective, it makes sense to describe infants' first encounters with words like 'red' as their bare recognition of sound patterns. Even though they might have expectations that **'red'** be meaningful, they might fail to experience **'red'** as having a determinate content, as occurs when a monolingual English speaker listens to Chinese or Russian. So there are psychological questions that look perfectly adequate even for those who may be reluctant to apply the vehicle-content distinction in the mental realm: How do children identify the content conveyed by physical vehicles like **'red'**? How should we characterize children's incorporation of new words in their

[20] To be sure, one might claim that the acquisition of a language produces changes in brain structures. So twins raised in different environments will not be physically identical. As far as I can see, this does not affect the main point of the argument: that there is something conventional in referential abilities.

cognitive life? How do these words interact with children's prior conceptual capacities?

The intuitive plausibility of these questions can be invoked in favor of semantic psychologism. McDowell and Wettstein might be right when they try to undermine some dualistic assumptions. After all, sharp dichotomies usually create unbridgeable gaps. Still, their remarks do not establish the truth of semantic anti-psychologism. It might be that the vehicle-content distinction is not adequate to theorize on the mental. It is however a good policy to introduce the vehicle-content distinction to mark the conventionality of natural languages. Besides, there is a similar conventionality in referential abilities. And none of the considerations introduced above suggests that one can explain reference only by adding a mysterious entity (i.e., intrinsic meaning or intentionality).[21]

6. The Argument from Confinement

The previous reply suggests that semantic anti-psychologism is based on a more substantial view of the mind and its place in the world. In the next sections, I explore some arguments that spell out this view. The first one, the 'argument from confinement,' traces back to early modern philosophy.

Here is a familiar line of reasoning. One starts by depicting minds as having ideas, intentionality, representations, concepts, a language, etc. Next, one assumes that these descriptions pick out some *constitutive* properties of minds. As a result, the question: 'How are ideas, intentionality, representations, concepts, a language, etc. 'hooked' to the world?' becomes problematic. After all, in order to solve the question, one has to assume that the relation between minds and those properties is not constitutive.

This problem has an epistemic counterpart. As theorists, we have minds. In order to answer questions about the relation between our constitutive properties and the world, we should be able to abstract from those properties. When we formulate the problem of reference, it is as if we had to get 'outside' ourselves or 'verify' how the reference relation between our constitutive properties and the world obtains. But, given that the link is constitutive, we cannot perform those feats.

[21] The vehicle-content distinction is one of the most widely *used* in cognitive science. There is, however, little *discussion* on the ways of drawing it. As far as I know, the clearest and most recurrent version of the distinction is based on the model of public language. Marr's (1982) computational theory of vision introduces the distinction by means of the notion of a code. This has remained a common practice among cognitive scientists and philosophers (see Block 1995).

This line of argument can be seen as a reaction to our previous reply. As the prior example of the twins suggested, it is not a constitutive fact about the twins that they use **'red'** or **'rojo'** to refer to RED. Those are conventional facts in Lewis' (1969) sense: for each twin, there are alternative ways of picking out RED that would have been equally effective. Still, this does not undermine the intuition that drives the modest philosopher. After all, before each twin learned how to use **'red'** or **'rojo**,' she was already able to refer to the world. So providing an account of how they managed to incorporate **'red'** or **'rojo'** in their linguistic repertoire does not really explain their more basic referential abilities. Thus, there is a sense in which we are 'confined' to the realm of intentionality even when we tell a story about the twins' acquisition of the words 'red' or 'rojo.'

There are classical versions of this problem in Berkeley's (1710) arguments for the claim that *esse est percipii* and also in Kant's (1781/1787) arguments for the unknowability of things in themselves. And their force remains intact in some circles. So Rorty (1979) rejects the tendency of analytic philosophers to believe that they could occupy a neutral point of view on nature, i.e. a standpoint independent of any empirical theory. Similarly, Searle (1983) justifies the modesty of his semantic analysis of intentionality on the ground that it is impossible to get 'outside' the intentional circle:

> In my view it is not possible to give a logical analysis of the Intentionality of the mental in terms of simpler notions, since Intentionality is, so to speak, a ground floor property of the mind, not a logically complex feature built up by combining simpler elements. There is no neutral standpoint from which we can survey the relations between Intentional states and the world and then describe them in non-Intentionalistic terms. *Any explanation of Intentionality, therefore, takes place within the circle of Intentional concepts.* (Searle 1983: 26; see also: 79)

It is difficult to resist this rhetoric of confinement. Human beings cannot occupy any neutral, external or transcendent point of view to survey the way natural language expressions, mental representations, or intentional states are related to the world. But we should be suspicious of this imagery. If we had skeptical proclivities, we should take very seriously the idea that we are confined to contemplate our ideas, representations or language. When we are theorizing on reference, however, we are already assuming that skepticism is not a live option. We are assuming that we *can* refer to the world. Crucially,

intentionality, representations, and languages are *parts* of the world. Having the ability to refer to intentional states, representations, and languages is the only thing we need to theorize on them!

The semantic psychologist is not forced to occupy any point of view external to intentionality, representations or language. What she rejects is to stay at the level of a *commonsense* understanding of mind and language. Granting that we must see the world through the prism of intentionality or language, we can still provide explanations *from* that perspective. This is what scientists do when they build models. Certainly, nobody has observed elementary particles with the naked eye. One can use, however, macroscopic objects as models of their structure. By parity of reasoning, even though we cannot place ourselves outside intentionality or language, we can use models available to language users to explain reference.[22]

The prior reply might look simplistic to some readers. If it does, it is likely that there are more substantial commitments in the argument from confinement. In what follows, I show that *some* of the arguments one might use to rebut the previous line of reply do not undermine semantic psychologism.

7. The Challenge of Internal Realism

When one asks questions concerning how a representation (public or linguistic) refers to an object, one is tacitly assuming that objects are self-standing entities that are intelligible independently of the relation of reference. In other words, one is assuming that there are two different realms: the referring realm (constituted by minds, ideas or representations) and the referred realm (constituted by objects, events, properties, etc.). The problem is to explain how these two realms relate to each other in the asymmetric and normative way proper to reference. Some philosophers have challenged this assumption, though. Examples include transcendental idealism (Kant 1781/1787) and Putnam's (1988, 1990) more recent defense of internal realism.

One of Putnam's arguments goes as follows. The theory of reference presupposes an epistemic-free notion of an object. But there are reasons not to hold an epistemic-free notion of an object. So, in order to engage in a theory of reference, one has to presuppose something we have good reasons not to presuppose.

I will not examine Putnam's defense of internal realism. Instead, I will argue that, even if internal realism is correct, it does not undermine the program

[22] I am indebted to Sellars (1956: 94-6; 1964: Chapter 1). See also Engel (1996: 239-ff.).

of providing an informative psychological account of referential abilities. To this end, let me start with the skeptical challenge he formulates on the way we *count* objects:

> Suppose I take someone into a room with a chair, a table on which there are a lamp and a notebook and a ballpoint pen, and nothing else, and ask, "How many objects are there in this room?" My companion answers, let us suppose, "Five." "What are they?" I ask. "A chair, a table, a lamp, a notebook, and a ballpoint pen." How about you and me? Aren't we in the room?" My companion might chuckle. "I didn't think you meant I was to count people as objects. Alright, then, seven." "How about the pages of the notebook?" (Putnam 1988: 110-1)

Putnam's argument has the form of a skeptical challenge. For any portion of reality R (e.g. a room), there is a *salient* answer to the question: How many objects are there in R? Still, one can always force one's opponent to grant that this salient answer is not compulsory, for it tacitly presupposes an arbitrary classificatory principle. Just by modifying the relevant classificatory principle, one can shift the estimation of the cardinality of objects. If the classificatory principle is 'non-living material object,' the response is 'five' but this answer is inadequate if one shifts to 'material object' as a classificatory principle. If classificatory principles are expressed by sortals $F_1, F_2, \ldots, F_{n-1}, F_n$, for any portion of reality R, the number of objects it contains is relative to a sortal F_i. Since sortals are relative to our minds, epistemic equipment, conceptual schemes, languages, etc., there is no epistemic-free notion of an object we can rely on in order to formulate the problem of reference.

There are a number of replies available to the defender of the epistemic-free notion of an object (or 'metaphysical realist' for short). Nevertheless, for the sake of the argument, I will grant Putnam's claim that the notion of an object is not epistemic-free. I will also grant that *many* theorists of reference seem to presuppose some form of metaphysical realism. Nevertheless, it does not follow that the problem of reference will disappear just by pointing out that the notion of an object is epistemic-dependent.

We can defend this point by examining an extreme version of the claim that the notion of an object is epistemic: the idealist contention that objects are not merely *relative to* a conceptual scheme but also *mental constructions*. Even in this extremely constructivist framework there is an intuitive difference between what one might call *purely subjective* states like pains and *seemingly objective* states like visual or tactile perceptual experiences. If one holds

that objects are mental constructions, one has to offer an account of the intuitive difference between these two sorts of states. After all, pains *do not seem* to refer to items in the world, at least in the same way as visual and tactile experiences do. The situation becomes even more delicate in the presence of linguistic devices such as proper names, which do *seem* to have referential functions that go beyond our mental life. This clearly shows that the problem of reference is not only a problem for the metaphysical realist. It is a problem for everybody, including the idealist who grants that there is a distinction between seemingly subjective and seemingly objective mental states. Whereas the metaphysical realist may be puzzled by the gap between minds and a realm of epistemic-free objects, the idealist should be puzzled by the difference between seemingly subjective and seemingly objective mental states. To be sure, the idealist cannot formulate the problem of reference as the problem of relating mental states to mind-independent items in the world. Yet, she will face the complementary problem of *internal differentiation:* How do conceptual schemes or languages generate the difference between seemingly subjective and seemingly objective mental states? What cognitive abilities underwrite that contrast? The problem of internal differentiation is the anti-realist counterpart of the more familiar problem of reference.

8. The Autonomy of the Intentional

Some modest philosophers might concede that we can study intentionality by building models available from our intentional perspective. They might also grant that even the anti-realist faces a problem that is very similar to the problem of reference. Nevertheless, they might reply that psychological concepts are not well suited to provide scientific explanations of reference or intentionality. And, when such explanations are provided, they always leave out a crucial aspect of reference or intentionality. This line of thought is implicit in McDowell's writings on this topic:

> An account given from outside is an account that denies itself the only descriptions under which we know that linguistic actions make rational sense, and we have been given no reason to suppose we can still see the activity of a speaker as hanging together rationally if we are required to describe it in other terms. (McDowell 1997: 113; see also Wettstein 2004)

The intuition behind this remark is that explanations at the sub-personal level fall short of providing a *full* understanding of reference, which is a personal-level notion. After all, it is *people* who refer to Aristotle by means of 'Aristotle,' not their parts or organs. So, cognitive accounts of reference leave out a central dimension of reference. This dimension includes the social and normative aspects of language use (see also Wettstein 2004: 108-9).

Consider an analogy. Someone asks: 'How does this machine work?' An engineer could try to find an answer by opening the machine, identifying the different parts, and seeing how they are related to each other. This would enable her to identify the function of the parts and the principles by which their interplay enables the machine to perform a complex task. The modest philosopher might insist, however, that this identification of smaller parts leaves something out. In the case of persons, the individual parts that compose them are not endowed with intentionality, for intentionality is the resultant of all the parts working in concert within a broader social and historical context. When you look for the smaller parts that make up the machine, you are not talking about intentionality anymore; the concepts of intentionality are designed to understand minds as situated in the world but not the mechanisms underlying them:

> [F]olk-psychological concepts can express a kind of understanding of a person that seems to have little or no relation to predictive power. [...] If the understanding that common-sense psychology yields is *sui generis*, there is no reason to regard it as a primitive version of the understanding promised by a theory of inner mechanisms. The two sorts of understanding need not compete for room to occupy. (McDowell 1995: 413; see also McDowell 1994)[23]

These remarks do not offer a clear argument against the semantic psychologist. Yet, they clearly delineate a possible view that, if true, would undermine the program of delivering a psychological account of reference. We can formulate the challenge as the conjunction of two claims:

> **(A) Autonomy:**
>
> (A1) Metaphysical: The personal level is independent from sub-personal levels.

[23] For a critical discussion of this sort of view, see Engel (1996), Bermúdez (2005), and Burge (2005).

(A2) Epistemological: Our understanding of ourselves as minds is confined to the personal level.

(I) Irreducibility:

The mind is irreducible to any sub-personal mechanism. In other words, the mind is not like an organ or a part of the person. Minds *are* persons.

If these claims were true, it would not be possible to understand the sub-personal level by employing or adapting aspects of the vocabulary of intentionality. Moreover, it would not be possible to shed light on personal-level notions like reference by engaging in sub-personal psychological theorizing. This approach enables us to see the debate between modest and full-blooded theorists in a new light. Psychological modesty is not merely a 'defeatist' attitude; it relies on a very specific view on the relation between personal and sub-personal levels.[24] Modesty hinges on the substantial claim that we cannot frame any full-blooded account of semantic competence in *sub-personal* terms.[25]

I propose to conclude this paper by providing a consideration against the modest outlook. According to modesty, in order to elucidate language, intentionality, etc., we must remain at the personal level. If we move to a lower level, we cannot find anything that may be recognized as genuinely intentional. When we theorize on lower levels, we are just changing the topic. Brain processes are too distant from personal experience to tell us anything interesting about philosophical muddles.

I take this challenge as a serious one. We cannot predict whether we will be able to provide adequate *psychological* explanations of reference. Despite this difficulty, there is at least one reason that militates in favor of a psychological account of reference.[26] This reason exploits the notion of *intuitive understanding*. Imagine that neuroscience has progressed so much, that it enables us to correlate types of brain activation with uses of types of words like 'not,' 'but,' 'Aristotle,' and so on. Whenever a person uttered 'Aristotle,' we would observe an activation pattern *A*; whenever she uttered 'Plato,' we would observe an activation pattern *P*, etc. Similarly, neuroscience would manage to

[24] *Pace* Dummett (1987) and Smith (2006: 951).

[25] For a well-documented discussion of the autonomy claim, see Bermúdez (2005: 3.1-3.2).

[26] The following considerations are inspired by Cussins (1987, 1992).

provide accurate descriptions of the use of word-types in the context of sentences. Thus, the sentence-type: 'Aristotle is a Great Philosopher' would correspond to the activation pattern *AGP*. Even in this happy scenario, where a perfect correlation would obtain between types of neural events and types of linguistic behavior, we would not think that *this* correlation provided everything we needed to *understand* the referential abilities underlying our uses of 'Aristotle,' 'Plato,' etc. Even if we were ready to accept the existence of such type-type correlations, they would remain *too distant* from *our* personal-level understanding of our use of proper names.

These considerations can be accepted by both modest and full-blooded theorists because they display an explanatory insufficiency of type-type correlations. Now, their reasons for being unsatisfied would be different. The semantic anti-psychologist would reject psychophysical type-identity because she wants to stress the autonomy of the personal over the sub-personal level. The semantic psychologist, by contrast, would probably exploit the prior intuition on the epistemic insufficiency of such perfect type-type correlations to introduce intermediary explanatory levels to account for referential abilities. Even after having established detailed correlations between types of linguistic behavior and types of brain activity, an *explanatory gap* would remain.

I think this consideration provides a *slight* advantage to semantic psychologists over semantic anti-psychologists. If the autonomy picture is right, there is nothing we can do to bridge the gap between our commonsense understanding of reference and the sort of understanding that could be provided by neuroscience. If semantic psychologists are right, however, there is something we can do to bridge the gap: develop *intermediary explanations* of reference couched in a vocabulary enabling us to get *an intuitive* understanding thereof. This is precisely what cognitive science and AI do when they idealize over some details of brain activity, and elaborate a theoretical apparatus that borrows some concepts from commonsense psychology. This theoretical apparatus is sufficiently similar to commonsense to offer an intuitive understanding of a number of phenomena but also sufficiently different therefrom to advance our understanding of our mental life beyond the confines of our commonsense view.

Notice that the argument is based on epistemological considerations. The limits of some forms of reductionism and pluralism are not merely ontological. They are epistemological as well (Cussins 1992; Kim 2010). Even if psychological phenomena turned out to be physics in the long run, we would still mind the gap. We would want to understand why things that look so different turned out to be identical. This is, I think, the force of the prior argument:

It does not beg the question against the modest theorist who insists on the autonomy of our personal-level view of the world. It shows how uncomfortable that view is for *understanding* our own place in a world that is—I take it—fundamentally physical.

Given the epistemological orientation of this argument, however, some theorists might declare it unstable. Consider a well-known episode in the history of astronomy. The geocentric theory provided *the* 'intuitive' understanding of the motions of the stars. According to that doctrine, the Earth was a stationary body at the center of the universe, and the celestial bodies moved around it. It seems natural to think that this theory provided a more intuitive understanding than the subsequent heliocentric theory. After all, the geocentric theory could be easily mapped onto our everyday experience of the sky, while the heliocentric view contradicted that ordinary experience. Still, the intuitive view was rejected in favor of a more counterintuitive explanation. At the end of the day, people had to learn to think in a new way that contradicted their most entrenched intuitions. Why should semantic competence and intentionality be different? Why persist in accounting for reference by multiplying levels of explanation?

I have no definite response to these questions. For the time being, I would justify the relevance of sub-personal explanations by stressing a potential difference between the geocentric case and the hypothetical case of a perfect type-type correlation between types of brain activity and types of uses of words. When one engages in sub-personal theorizing, one is tacitly assuming that the gap between the higher level of language use and intentionality and the lower level of neurological mechanisms is so deep that intermediary explanatory levels are necessary. What we do is look at our personal-level concepts and try to exploit them as models to bridge the gap. But there is nothing analogous to this explanatory strategy in the geocentric example. *The hypothesis that the stars turn around the earth is not a step toward a better understanding of the heliocentric hypothesis but an obstacle thereto.* If we introduce sub-personal explanations of reference, it is because we mind the gap between explanations in neural terms and the semantic vocabulary we use to characterize language and intentionality. Bridging the gap requires the establishment of conceptual bridges, not a sharp personal/neural dichotomy.

9. Concluding Remarks

The contrast between full-blooded and modest theories of meaning articulates an opposition between an optimistic and a pessimistic attitude toward the explanation of reference. Since the relevant notion of explanation can be understood in different ways, there are many different ways of drawing the line. In this paper, I focused on a strand of the full-blooded/modest divide that has played a central role in Engel's work: whether one could provide a psychological account of referential abilities. I defended this program by displaying the insufficiency of truth-conditional semantics, and responding to some influential arguments against what I called 'semantic psychologism.'

The discussion of these arguments showed that, far from being a defeatist view, modesty hinges on substantial claims about the nature of mind and its relation to the world. My response to these arguments purported to show that, even if one grants some of these substantial claims, the problem of reference would not disappear. And our last argument led us to a more fundamental issue: the contrast between a view of the mind as an autonomous realm and a more interactive picture that seeks to spell out the relations that various levels bear to each other. Since the latter picture is not committed to reductionism, it bears some similarities to Engel's idea of world 2½: semantic psychologism may be seen as the study of how psychology and semantics are related to each other. This sort of inquiry requires a revision of the usual view of the philosophy of language as a branch of logic, and a rejection of anti-psychologism as our default view of normativity.

10. References

Almog, J. (2005) Is A Unified Description of Language-and-Thought Possible? *The Journal of Philosophy*, 102 (10): 493-531

Barwise, J. & J. Perry (1983) *Situations and Attitudes*, Cambridge (MA.), MIT Press / Bradford Books

Berkeley, G. (1710) *A Treatise Concerning the Principles of Human Knowledge*, in his *Philosophical Writings*, edited by Desmond M. Clarke, Cambridge, Cambridge University Press, 2008

Bermúdez, J. L. (2005) *Philosophy of Psychology: A Contemporary Introduction*, New York, Routledge

Block, N. (1995) The Mind as the Software of the Brain, in E. E. Smith & D. N. Osherson, eds., *An Invitation to Cognitive Science*, 2 nd edition, vol. 3: Thinking, Cambridge (MA.), MIT Press: 377-425

Burge, T. (1977) Belief *De Re*, *The Journal of Philosophy*, 74 (6): 338-62. Reprinted with a postscript in his *Foundations of Mind*, Oxford, Clarendon Press, 2007: 44-81

Burge, T. (1990) Frege on Sense and Linguistic Meaning, in David Bell & Neil Cooper, eds., *The Analytic Tradition: Meaning, Thought, and Knowledge*, Oxford, Blackwell: 30-60. Reprinted in his *Truth, Thought, Reason: Essays on Frege*, Oxford, Clarendon Press, 2005: 242-69

Burge, T. (2005) Disjunctivism and Perceptual Psychology, *Philosophical Topics*, 33 (1): 1-78

Clark, A. (2005) Intrinsic Content, Active Memory, and the Extended Mind, *Analysis*, 65 (285): 1-11

Cussins, A. (1987) Varieties of Psychologism, *Synthese*, 70 (1): 123-54

Cussins, A. (1992) The Limitations of Pluralism, in D. Charles & K. Lennon, eds., *Reduction, Explanation, and Realism*, Oxford, Clarendon Press: 179-223

Davidson, D. (1984) *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press

Davies, M. (1987) Tacit Knowledge and Semantic Theory: Can a Five Per Cent Difference Matter? *Mind*, 96 (384): 441-62

Devitt, M. (1981) *Designation*, New York, Columbia University Press

Donnellan, K. S. (1974) Speaking of Nothing, *The Philosophical Review*, 83 (1): 3-31

Dretske, F. (1981) *Knowledge and the Flow of Information*, Cambridge (MA.), MIT Press

Dummett, M. (1973) *Frege: Philosophy of Language*, 2 nd edition, Cambridge (MA.), Harvard University Press

Dummett, M. (1975) What Is a Theory of Meaning? In S. Guttenplan, ed., *Mind and Language*, Oxford, Oxford University Press. Reprinted with an Appendix in his *The Seas of Language*, Oxford, Clarendon Press: 1-33

Dummett, M. (1976) What Is a Theory of Meaning? (II) In G. Evans & J. McDowell, eds., *Truth and Meaning: Essays in Semantics*, Oxford / New York, Oxford University Press. Reprinted in his *The Seas of Language*, Oxford, Clarendon Press: 34-93

Dummett, M. (1978) What Do I Know When I Know a Language? First published in his *The Seas of Language*, Oxford, Clarendon Press: 94-105

Dummett, M. (1987) Reply to John McDowell, in B. Taylor, ed., *Michael Dummett: Contributions to Philosophy*, Dordrecht, Martinus Nijhoff: 253-68

Dummett, M. (1991) *The Logical Basis of Metaphysics*, London, Duckworth

Engel, P. (1989) *La norme du vrai : Philosophie de la logique*, Paris, Gallimard

Engel, P. (1994) *Davidson et la philosophie du langage*, Paris, PUF

Engel, P. (1996) *Philosophie et psychologie*, Paris, Gallimard

Engel, P. (2001) The False Modesty of the Identity Theory of Truth, *International Journal of Philosophical Studies*, 9 (4): 441-58

Engel, P. (2006) Logic, Reasoning and the Logical Constants, *Croatian Journal of Philosophy*, 6 (17): 219-35

Engel, P. (2007) Review of Ernest Lepore, Kirk Ludwig, *Donald Davidson's Truth-Theoretic Semantics*, *Notre Dame Philosophical Reviews*, 8

Engel, P. (Ms.) Une théorie de la signification peut-elle être autre que modeste ?

Evans, G. (1981a) Understanding Demonstratives, in H. Parret & J. Bouveresse, eds., *Meaning and Understanding*, Berlin, New York, Walter de Gruyter: 280-303. Reprinted in his *Collected Papers,* Oxford, Clarendon Press: 291-321

Evans, G. (1981b) Semantic Theory and Tacit Knowledge, in S. Holtzman & C. Leich, eds., *Wittgenstein: To Follow a Rule*, London, Routledge & Kegan Paul. Reprinted with a new section in his *Collected Papers*, Oxford, Clarendon Press: 322-42

Evans, G. (1982) *The Varieties of Reference*, edited by John McDowell, Oxford, Clarendon Press

Field, H. (1972) Tarski's Theory of Truth, *The Journal of Philosophy*, 69: 347-75

Field, H. (1978) Mental Representation, *Erkenntnis*, 13: 9-61. Reprinted with a postscript in N. Block, ed., *Readings in Philosophy of Psychology*, vol. 2, Cambridge (MA.), Harvard University Press, 1981: 78-114

Frege, G. (1891) Funktion und Begriff, *Vortrag, gehalten in der Sitzung vom 9. Januar 1891 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft*, Jena, H. Pohle, II, 31 pp. Reprinted in G. Frege, *Funktion – Begriff – Bedeutung*, edited by M. Textor, Göttingen, Vandenhoeck & Ruprecht, 2002: 1-22

Frege, G. (1893) *Grundgesetze der Arithmetik, begriffschriftlich abgeleitet*, vol. 1, Jena, Pohle. Reprinted: Hildesheim, Olms, 1962

Frege, G. (1918-1919) Der Gedanke – eine logische Untersuchung, *Beiträge zur Philosophie des deutschen Idealismus*, 1: 58-77. Reprinted in G. Frege, *Logische Untersuchungen*, edited by G. Patzig, Göttingen, Vandenhoeck & Ruprecht, 2003: 35-62

Heck, R. (2002) Do Demonstratives Have Senses? *Philosophers' Imprint*, 2 (2): 1-33

Higginbotham, J. (1998) Conceptual Competence, *Philosophical Issues,* vol. 9: Concepts: 149-162

Kant, I. (1781/1787) *Kritik der reinen Vernunft*, Hamburg, Felix Meiner, 1998

Kaplan, D. (1989) Afterthoughts, in J. Almog, J. Perry & H. Wettstein, eds., *Themes from Kaplan*, New York, Oxford University Press: 565-614

Kim, J. (2010) *Essays in the Metaphysics of Mind*, Oxford, Oxford University Press

Kripke, S. A. (1979) A Puzzle About Belief, in A. Margalit, ed., *Meaning and Use*, Dordrecht, Reidel: 239-83. Reprinted in P. Ludlow, ed., *Readings in the Philosophy of Language*, Cambridge (MA.) / London, MIT Press / Bradford Books: 875-920

Kripke, S. A. (1980) *Naming and Necessity*, Cambridge (MA.), Harvard University Press

Kripke, S. A. (2008) Frege's Theory of Sense and Reference: Some Exegetical Notes, *Theoria*, 74: 181-218

Lewis, D. (1969) *Convention: A Philosophical Study*, Cambridge (MA.), Harvard University Press

Lewis, D. (1980) Index, Context, and Content, in S. Kanger & S. Öhman, eds., *Philosophy and Grammar,* Dordrecht, Reidel: 79-100. Reprinted in his *Papers in Philosophical Logic*, Cambridge, Cambridge University Press, 1998: 21-44

McDowell, J. (1977) On the Sense and Reference of a Proper Name, *Mind*, 86: 159-85. Reprinted in his *Meaning, Knowledge, and Reality*, Cambridge (MA.) / London, Harvard University Press: 171-98

McDowell, J. (1987) In Defense of Modesty, in B. Taylor, ed., *Michael Dummett: Contributions to Philosophy*, Dordrecht, Martinus Nijhoff: 59-80. Reprinted in his *Meaning, Knowledge, and Reality*, Cambridge (MA.) / London, Harvard University Press: 87-107

McDowell, J. (1992) Putnam on Mind and Meaning, *Philosophical Topics*, 20: 35-48. Reprinted in his *Meaning, Knowledge, and Reality*, Cambridge (MA.) / London, Harvard University Press: 275-91

McDowell, J. (1994) The Content of Perceptual Experience, *The Philosophical Quarterly*, 44: 190-205. Reprinted in his *Mind, Value, and Reality*, Cambridge (MA.) / London, Harvard University Press: 341-58

McDowell, J. (1995) Knowledge and the Internal, *Philosophy and Phenomenological Research*, 55: 877-93. Reprinted in his *Meaning, Knowledge, and Reality*, Cambridge (MA.) / London, Harvard University Press: 395-413

McDowell, J. (1996) *Mind and World*, 2 nd edition, with a New Introduction, Cambridge (MA.), Harvard University Press

McDowell, J. (1997) Another Plea for Modesty, in R. G. Heck, ed., *Language, Thought, and Logic: Essays in Honour of Michael Dummett*, Oxford / New York, Oxford University Press: 105-29. Reprinted in his *Meaning, Knowledge, and Reality*, Cambridge (MA.) / London, Harvard University Press: 108-31

McKinsey, M. (2009) Thought by Description, *Philosophy and Phenomenological Research*, 78 (1): 83-102

Marr, D. (1982) *Vision*, San Francisco (CA.), W. H. Freeman

Millikan, R. G. (1984) *Language, Thought, and Other Biological Categories*, Cambridge (MA.) / London, MIT Press / Bradford Books

Millikan, R. G. (2005) *Language: A Biological Model*, Oxford, Oxford University Press

Predelli, S. (2005) *Contexts: Meaning, Truth, and the Use of Language*, Oxford, Clarendon Press

Pylyshyn, Z. W. (2007) *Things and Places: How the Mind Connects with the World*, Cambridge (MA.) / London, MIT Press / Bradford Books

Putnam, H. (1988) *Representation and Reality*, Cambridge (MA.) / London, MIT Press / Bradford Books

Putnam, H. (1990) *Realism With a Human Face*, Cambridge (MA.), Harvard University Press

Rorty, R. (1979) *Philosophy and the Mirror of Nature*, Princeton / Oxford, Princeton University Press

Sainsbury, R. M. (2005) *Reference Without Referents*, Oxford / New York, Oxford University Press

Salmon, N. (1986) *Frege's Puzzle*, Cambridge (MA.) / London, MIT Press / Bradford Books

Searle, J. R. (1983) *Intentionality*, Cambridge, Cambridge University Press

Sedivy, S. (2004) Minds: Contents Without Vehicles, *Philosophical Psychology*, 17 (2): 149-81

Sellars, W. (1956) Empiricism and the Philosophy of Mind, *Minnesota Studies in Philosophy of Science*, vol. 1, edited by H. Feigl & M. Scriven, Minneapolis, University of Minnesota Press. Reprinted as *Empiricism and the Philosophy of Mind*, with an Introduction by R. Rorty and a Study Guide by R. Brandom, Cambridge (MA) / London, Harvard University Press

Smith, B. C. (2006) What I Know When I Know A Language, in E. LePore & B. C. Smith, eds., *The Oxford Handbook of Philosophy of Language*, Oxford, Oxford University Press: 941-82

Wettstein, H. (2004) *The Magic Prism: An Essay in the Philosophy of Language*, Oxford / New York, Oxford University Press

34

Éléments d'un contextualisme dialectique *

PAUL FRANCESCHI

Résumé Dans ce qui suit, je m'attache à présenter les éléments d'une doctrine philosophique, qui peut être définie comme un *contextualisme dialectique*. Je m'efforce tout d'abord de définir les éléments constitutifs de cette doctrine, à travers les dualités et pôles duaux, le principe d'indifférence dialectique et le biais d'uni-polarisation. Je m'attache ensuite à souligner l'intérêt spécifique de cette doctrine au sein d'un domaine particulier de la méta-philosophie : la méthodologie utilisée pour la résolution des paradoxes philosophiques. Je décris enfin une application de cette dernière aux paradoxes suivants : le paradoxe de Hempel, le paradoxe de l'examen-surprise et l'argument de l'Apocalypse.

Abstract In what follows, I strive to present the elements of a philosophical doctrine, which can be defined as *dialectical contextualism*. I proceed first to define the elements of this doctrine, through dualities and polar contraries, the principle of dialectical indifference and the one-sidedness bias. I emphasize then the special importance of this doctrine in a specific field of metaphilosophy : the methodology for solving philosophical paradoxes. Finally, I describe several applications of this methodology on the following paradoxes : Hempel's paradox, the surprise examination paradox and the Doomsday Argument.

*Ce texte constitue une version rédigée à partir d'éléments entièrement remaniés de mon mémoire d'habilitation à diriger les recherches, présenté en 2006. Les modifications introduites dans le texte, comportant notamment la correction d'une erreur conceptuelle, suivent en cela les commentaires et les recommandations que Pascal Engel m'avait faits à l'époque.

Mots-clés contextualisme dialectique, contextualisme, dialectique, biais d'uni-polarisation, distorsion cognitive méta-philosophie, pôles duaux

Keywords dialectical contextualism, contextualism, dialectics, one-sidedness bias, cognitive distortion, metaphilosophy, polar contraries

Dans ce qui suit, je m'attacherai à présenter les éléments d'une doctrine philosophique spécifique, qui peut être définie comme un *contextualisme dialectique*. Je m'efforcerai tout d'abord de préciser les éléments qui caractérisent cette doctrine, en particulier les dualités et pôles duaux, le principe d'indifférence dialectique et le biais d'uni-polarisation. Je m'attacherai ensuite à en décrire l'intérêt au niveau méta-philosophique, notamment en tant que méthodologie pour aider à la résolution des paradoxes philosophiques. Je décrirai enfin une application de cette méthodologie à l'analyse des paradoxes philosophiques suivants : le paradoxe de Hempel, le paradoxe de l'examen-surprise et l'argument de l'Apocalypse.

Le contextualisme dialectique décrit ici est fondé sur un certain nombre d'élément constitutifs qui présentent une nature spécifique. Au nombre de ces derniers figurent : les dualités et pôles duaux, le principe d'indifférence dialectique et le sophisme d'uni-polarisation. Il convient d'analyser tour à tour chacun de ces éléments.

1. Dualités et pôles duaux

Nous nous attacherons tout d'abord à définir la notion de *pôles duaux* (*polar opposites*)[1]. Bien qu'intuitive, une telle notion nécessite d'être précisée. Des exemples de pôles duaux sont ainsi *statique/dynamique*, *interne/externe*, *qualitatif /quantitatif*, etc. Nous pouvons définir les pôles duaux comme des concepts (que nous pouvons dénommer A et $\bar{A}$) qui se présentent par paires, et qui sont tels que chacun d'eux est défini comme le contraire de l'autre. Par exemple, *interne* peut être défini comme le contraire d'*externe*, et de manière symétrique, *externe* est défini comme le contraire d'*interne*. En un certain sens, il n'y a pas ici de notion primitive et aucun des deux pôles duaux A et $\bar{A}$ ne peut

[1]Une telle notion se trouve au coeur du concept de *matrice de concepts* introduit dans Franceschi (2002), dont on peut considérer qu'elle constitue le noyau, ou une forme simplifiée. Pour le présent exposé portant spécifiquement sur les éléments du contextualisme dialectique et leur application pour la résolution de paradoxes philosophiques, la présentation des pôles duaux se révèle suffisante.

être considéré comme la notion primitive. Considérons tout d'abord une *dualité* donnée, que nous pouvons dénoter par A/Ā, où A et Ā constituent des concepts *duaux*. Une telle dualité est représentée sur la figure ci-dessous :

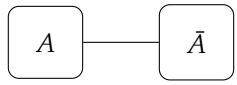


Figure 34.1: Les pôles duaux A et $\bar{A}$

À ce stade, nous pouvons donner également une énumération (qui présente nécessairement un caractère partiel) des dualités :

> Interne/Externe, Quantitatif/Qualitatif, Visible/Invisible, Absolu/Relatif, Abstrait/Concret, Statique/Dynamique, Diachronique/Synchronique, Unique/Multiple, Extension/Restriction, Esthétique/Pratique, Précis/Vague, Fini/Infini, Simple/Composé, Individuel/Collectif, Analytique/Synthétique, Implicite/Explicite, Volontaire/Involontaire

Afin de caractériser les pôles duaux avec davantage de précision, il convient de s'attacher à les distinguer par rapport à d'autres concepts. Nous présenterons ainsi plusieurs propriétés des pôles duaux, qui permettent de les différencier d'autres concepts voisins. Les pôles duaux sont ainsi des concepts neutres, de même que des qualités simples ; en outre, ils se distinguent des notions vagues. En premier lieu, deux pôles duaux A et Ā constituent des concepts *neutres*. Ils peuvent ainsi être dénotés par A^0 et $\bar{A}^0$. Ceci conduit à représenter les deux concepts A^0 et $\bar{A}^0$ de la manière suivante :

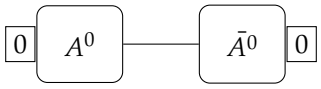


Figure 34.2: Les pôles duaux neutres A^0 et $\bar{A}^0$

Les pôles duaux constituent des concepts neutres, c'est-à-dire des concepts qui ne présentent aucune nuance méliorative ou péjorative. En ce sens, *externe*, *interne*, *concret*, *abstrait*, etc., constituent des pôles duaux, à la différence de concepts tels que *beau*, *laid*, *courageux*, qui présentent une nuance soit méliorative soit péjorative, et qui sont donc non-neutres. Le fait que les pôles duaux soient neutres possède son importance, car cela permet de les distinguer de

concepts qui possèdent une connotation *positive* ou *négative*. Ainsi, la paire de concepts *beau/laid* ne constitue pas une dualité et *beau* et *laid* ne constituent donc pas des pôles duaux, au sens de la présente construction. En effet, *beau* possède une connotation positive et *laid* présente une nuance péjorative. Dans ce contexte, nous pouvons les dénoter par $beau^+$ et *laid*.

Il convient de souligner, en second lieu, que les deux pôles duaux d'une même dualité correspondent à des *qualités simples*, par opposition aux *qualités composées*. La distinction entre qualités simples et composées peut s'effectuer de la manière suivante. Soient A_1 et A_2 des qualités simples. Dans ce cas, $A_1 \wedge A_2$, de même que $A_1 \vee A_2$ sont des qualités composées. Pour prendre un exemple, *statique*, *qualitatif*, *externe* sont des qualités simples, alors que *statique et qualitatif*, *statique et externe*, *qualitatif et externe*, sont des qualités composées. Une définition plus générale est ainsi la suivante : soient B_1 et B_2 des qualités simples ou composées, dans ce cas $B_1 \wedge B_2$, de même que $B_1 \vee B_2$ sont des qualités composées. De manière incidente, ceci met également en lumière pourquoi les paires de concepts *rouge/non-rouge*, *bleu/non-bleu* ne peuvent pas être considérés comme des pôles duaux. En effet, *non-rouge* peut ainsi être défini en tant que qualité composée de la manière suivante : *violet* $\vee$ *indigo* $\vee$ *bleu* $\vee$ *vert* $\vee$ *jaune* $\vee$ orange $\vee$ *blanc* $\vee$ *noir*. Dans ce contexte, on peut assimiler *non-bleu* à la *négation-complément* de *bleu*, une telle négation-complément étant définie à l'aide de qualités composées.

Compte tenu de la définition précédente, nous sommes également en mesure de distinguer les pôles duaux des objets *vagues*. Nous pouvons observer tout d'abord que les pôles duaux et les objets vagues possèdent en commun certaines propriétés. En effet, les objets vagues se présentent par paires, de la même manière que les pôles duaux. De plus, les concepts vagues sont considérés classiquement comme possédant une extension et une anti-extension, qui sont mutuellement exclusives. Une telle caractéristique est également partagée par les pôles duaux. À titre d'exemple, *qualitatif* et *quantitatif* s'assimilent à une extension et à une anti-extension, qui présentent la propriété d'être mutuellement exclusives ; il en va de même pour *statique* et *dynamique*, etc. Cependant, il convient de souligner les différences existant entre les deux catégories de concepts. Une première différence (a) réside ainsi dans le fait que l'union de l'extension et l'anti-extension des concepts vagues n'est pas exhaustive, en ce sens qu'elles admettent des cas-limites (et aussi des cas-limites de cas-limites, etc. donnant ainsi naissance à une hiérarchie du vague d'ordre n), qui constitue une zone de pénombre. À l'inverse, les pôles duaux ne possèdent pas nécessairement une telle caractéristique. En effet, l'union des pôles duaux peut être soit exhaustive, soit non-exhaustive. Par exemple, la dualité

abstrait/concret est, de manière intuitive, exhaustive, car il ne semble pas exister d'objets qui ne sont ni abstraits ni concrets. Il en va de même pour la dualité *vague/précis* : intuitivement, il n'existe pas en effet d'objets qui ne sont ni vagues ni précis, et qui appartiendraient à une catégorie intermédiaire. Ainsi, il existe des pôles duaux dont l'extension et l'anti-extension se révèle exhaustive, tels les deux pôles de la dualité *abstrait/concret*, à la différence des notions vagues. Il convient de mentionner, en second lieu, une autre différence (b) entre les pôles duaux et les objets vagues. En effet, les pôles duaux constituent des qualités simples, alors que les objets vagues peuvent consister en des qualités simples ou composées. Il existe en effet des concepts dénommés objets vagues multi-dimensionnels, tels que la notion de *véhicule*, de *machine*, etc. Enfin, une dernière différence entre les deux catégories d'objets (c) réside dans le fait que certains pôles duaux présentent une nature intrinsèquement précise. Tel est notamment le cas de la dualité *individuel/collectif*, qui est susceptible de donner lieu à une définition tout à fait précise.

2. Le principe d'indifférence dialectique

À partir des notions de dualité et de pôles duaux qui viennent d'être définis, nous sommes en mesure de définir également une notion de *point de vue*, relatif à une dualité ou un pôle dual donné. Ainsi, nous avons tout d'abord la notion de point de vue correspondant à une *dualité* donnée $A/\bar{A}$: ceci correspond par exemple au point de vue de la dualité *extension/restriction*, celui de la dualité *qualitatif/quantitatif*, ou de la dualité *diachronique/synchronique*, etc. Il en résulte également la notion de point de vue relatif à un *pôle* donné d'une dualité $A/\bar{A}$: on a par exemple (au niveau de la dualité *extension/restriction*) le point de vue par *extension*, de même que le point de vue par *restriction*. De même, il en résulte le point de vue ou angle *qualitatif*, ainsi que le point de vue ou angle *quantitatif*, etc. (au niveau de la dualité *qualitatif/quantitatif*). Ainsi, lorsqu'on considère un objet donné *o* (que ce soit un objet concret ou bien un objet abstrait telle que par exemple une proposition ou un raisonnement), on est susceptible d'envisager ce dernier par rapport à différentes dualités, et au niveau de ces dernières, par rapport à chacun de ses deux pôles duaux.

L'idée sous-jacente inhérente aux points de vue relatifs à une dualité donnée, ou à un pôle donné d'une dualité, est que chacun des deux pôles d'une même dualité, *toutes choses étant par ailleurs égales*, possède une égale légitimité. En ce sens, si on considère un objet *o* du point de vue d'une dualité $A/\bar{A}$, il convient de ne pas privilégier l'un des pôles par rapport à l'autre. Afin d'ob-

tenir un point de vue objectif par rapport à une dualité A/Ā, il convient de se placer tout à tour du point de vue du pôle A, puis de celui du pôle Ā. Car une approche qui n'aborderait que le point de vue de l'un des deux pôles se révélerait partielle et tronquée. Le fait de considérer tour à tour le point de vue des deux pôles, lors de l'étude d'un objet *o* et de la classe de référence qui lui est associée, permet d'éviter une démarche subjective et de satisfaire, autant que possible, les besoins de l'objectivité.

On le voit, l'idée qui sous-tend la notion de point de vue peut être formalisée en un *principe d'indifférence dialectique*, de la manière suivante :

> (PRINCIPE D'INDIFFERENCE DIALECTIQUE) Lorsqu'on considère un objet donné *o* et la classe de référence E qui lui est associée, sous l'angle de la dualité A/Ā, toutes choses étant par ailleurs égales, il convient d'accorder une égale importance au point de vue du pôle A et au point de vue du pôle Ā.

Ce principe est formulé en terme de *principe d'indifférence* : si l'on considère un objet *o* sous l'angle d'une dualité A/Ā, il n'y a pas lieu de privilégier le point de vue A par rapport au point de vue Ā, et sauf élément contraire résultant du contexte, on doit placer à égalité les points de vue A et Ā. Une conséquence directe de ce principe est que si l'on considère le point de vue du pôle A, il est nécessaire de prendre également en considération le point de vue du pôle opposé Ā (et réciproquement). La nécessité de prendre en considération les deux points de vue, celui résultant du pôle A et celui associé au pôle Ā, répond au souci d'analyser l'objet *o* et la classe de référence qui lui est associée d'un point de vue objectif. Cette objectivité est atteinte, autant que faire se peut, par la prise en considération des points de vue complémentaires qui sont ceux des pôles A et Ā. Chacun de ces points de vue possède en effet, eu égard à la dualité A/Ā, un droit égal à la pertinence. Dans de telles circonstances, lorsque seul le pôle A ou (exclusivement) le pôle Ā est pris en considération, il s'agit alors d'un point de vue *uni-polarisé*. À l'inverse, le point de vue qui réalise la synthèse des points de vue correspondants aux pôles A et Ā, est par nature *bi-polarisé*. Fondamentalement, une telle démarche se révèle d'essence dialectique. En effet, l'étape d'analyse successive des points de vue complémentaires par rapport à une classe de référence donnée, est destinée à permettre, dans une étape ultérieure, une synthèse finale, qui résulte de la prise en compte conjointe des points de vue correspondant à la fois aux pôles A et Ā. Dans la présente construction, le processus de confrontation des différents points de vue pertinents par rapport à une dualité A/Ā est destiné à

construire, cumulativement, un point de vue plus objectif et exhaustif que celui, nécessairement partiel, qui résulte de la prise en compte des données qui résultent d'un seul des deux pôles.

La définition du principe d'indifférence dialectique qui est proposée ici se réfère à une *classe de référence* E, qui se trouve associée à l'objet *o*. La classe de référence[2] est constituée par un ensemble de phénomènes ou d'objets. Plusieurs exemples peuvent en être donnés : la classe des êtres humains ayant jamais existé, la classe des événements futurs de la vie d'une personne, la classe des parties du corps d'une personne, la classe des corbeaux, etc. Nous examinerons, dans ce qui suit, un certain nombre d'exemples. La mention d'une telle classe de référence possède son importance, car sa définition-même se trouve associée à la dualité $A/\bar{A}$ précitée. En effet, la classe de référence peut être définie du point de vue de A ou bien du point de vue de $\bar{A}$. Une telle particularité nécessite d'être soulignée et nous sera utile lors de la définition du biais qui se trouve associé à la définition-même du principe d'indifférence dialectique : le biais d'uni-polarisation.

3. Caractérisation du biais d'uni-polarisation

La formulation précédente du principe d'indifférence dialectique suggère, de manière directe, une erreur de raisonnement d'un certain type. De manière informelle, une telle erreur de raisonnement consiste à privilégier un point de vue lorsqu'on s'intéresse à un objet donné, et à négliger le point de vue opposé. De manière plus formelle, dans le contexte qui vient d'être décrit, une telle erreur de raisonnement consiste, lorsqu'on considère un objet *o* et la classe de référence qui lui est associée, à ne prendre en considération que le point de vue du pôle A (respectivement $\bar{A}$), en occultant complètement le point de vue du pôle dual $\bar{A}$ (respectivement A) pour définir cette classe de référence. Nous dénommerons *biais d'uni-polarisation* un tel type d'erreur de raisonnement. Les conditions de ce type de biais, en violation du principe d'indifférence dialectique, méritent toutefois d'être précisées. En effet, dans le présent contexte, on peut considérer qu'il existe certains cas, où la bi-polarisation par rapport à une dualité donnée $A/\bar{A}$ n'est pas requise. Tel est le cas lorsque les éléments du contexte ne présupposent pas des conditions d'objectivité et d'exhaustivité des points de vue. Ainsi, un avocat qui ne ferait valoir que les

[2]La présente construction s'applique également à des objets qui sont associés à plusieurs classes de référence. Nous nous limitons ici, dans un souci de simplification, à une seule classe de référence.

éléments à la décharge de son client, en ignorant complètement les éléments à charge, ne commettrait pas le type d'erreur de raisonnement précité. Dans une telle circonstance en effet, l'avocat ne commettrait pas un biais d'uni-polarisation dommageable, puisqu'il s'agit de la fonction qui lui est propre. Il en irait de même dans un procès pour le procureur qui, à l'inverse, mettrait uniquement l'accent sur les éléments à charge de la même personne, en ignorant complètement les éléments à décharge. Dans une telle situation également, le biais d'uni-polarisation en résultant ne serait pas inapproprié, car il résulte bien des éléments du contexte qu'il s'agit bien du rôle limité qui est assigné au procureur. En revanche, un juge qui ne prendrait en compte que les éléments à charge de l'accusé, ou bien qui commettrait l'erreur inverse, de ne considérer que les éléments à décharge de ce dernier, commettrait bien un biais d'uni-polarisation indésirable, car le rôle-même du juge implique qu'il prenne en considération les deux catégories d'éléments et que son jugement résulte de la synthèse qui en est effectuée.

En outre, ainsi que nous l'avons mentionné plus haut, la mention d'une classe de référence associée à l'objet o se révèle importante. En effet, ainsi que nous aurons l'occasion de le constater avec l'analyse des exemples qui suivent, sa définition-même se trouve associée à une dualité $A/\bar{A}$. Et la classe de référence peut être définie soit du point de vue de A, soit du point de vue de $\bar{A}$. Une telle particularité a pour conséquence que tous les objets ne sont pas susceptibles de donner lieu à un biais d'uni-polarisation. En particulier, les objets auxquels ne sont pas associés une classe de référence qui est elle-même susceptible d'être envisagée sous l'angle d'une dualité $A/\bar{A}$, ne donnent pas lieu à un tel biais d'uni-polarisation.

Avant d'illustrer la présente construction à l'aide de plusieurs exemples concrets, il apparaît utile à ce stade, de considérer le biais d'uni-polarisation qui vient d'être défini, et qui résulte de la définition-même du principe d'indifférence dialectique, à la lumière de plusieurs notions similaires. De manière préliminaire, nous pouvons observer qu'une description générale de ce type d'erreur de raisonnement avait déjà été formulée, en des termes voisins, par John Stuart Mill (On Liberty, II) :

> He who knows only his own side of the case, knows little of that. His reasons may be good, and no one may have been able to refute them. But if he is equally unable to refute the reasons on the opposite side; if he does not so much know what they are, he has no ground for preferring either opinion.

Dans la littérature récente, des notions très voisines ont également été décrites. Il s'agit en particulier du *biais dialectique* décrit notamment par Douglas Walton (1999). Walton (1999, pp. 76-77) se place ainsi dans le cadre la théorie dialectique des biais, qui oppose les argument uni-polarisés aux arguments bi-polarisés :

> The dialectical theory of bias is based on the idea [...] that an argument has two sides. [...] A *one-sided argument* continually engages in pro-argumentation for the position supported and continually rejects the arguments of the opposed side in a dialogue. A *two-sided* (*balanced*) argument considers all arguments on both sides of a dialogue. A balanced argument weights each argument against the arguments that have been opposed to it.

Walton décrit ainsi le biais dialectique (*dialectical bias*) comme un point de vue uni-polarisé qui survient au cours de l'argumentation. Walton souligne que le biais dialectique, qui est universellement répondu dans l'argumentation humaine, ne constitue pas nécessairement une erreur de raisonnement. Suivant en cela la distinction entre « bon » et « mauvais » biais due à Antony Blair (1988), Walton considère que le biais dialectique est incorrect seulement dans certaines conditions, et en particulier s'il survient dans un contexte qui est supposé être équilibré, c'est-à-dire où les deux facettes du raisonnement correspondant sont censées être mentionnées (p. 81) :

> Bad bias can be defined as "pure (one-sided) advocacy" in a situation where such unbalanced advocacy is normatively inappropriate in argumentation.

Une notion très voisine du biais d'uni-polarisation est également décrite par Peter Suber (1998). Suber décrit en effet une erreur de raisonnement qu'il dénomme *sophisme d'uni-polarisation* (*one-sidedness fallacy*). Il décrit ce dernier comme un raisonnement fallacieux qui consiste à ne présenter qu'un aspect des éléments qui justifient un jugement ou un point de vue donné, en occultant complètement l'autre aspect des éléments pertinents relatifs à ce même jugement :

> The fallacy consists in persuading readers, and perhaps ourselves, that we have said enough to tilt the scale of evidence and therefore enough to justify a judgment. If we have been one-sided, though,

> then we haven't yet said enough to justify a judgment. The arguments on the other side may be stronger than our own. We won't know until we examine them.

L'erreur de raisonnement consiste ainsi à ne pas prendre en compte qu'un point de vue concernant le jugement en question, alors même que l'autre point de vue pourrait se révéler décisif quant à la conclusion à en tirer. Suber entreprend également de donner une caractérisation du sophisme d'uni-polarisation et observe en particulier que le sophisme d'uni-polarisation constitue un argument valide. Car sa conclusion est vraie si ses prémisses en sont vraies. Plus encore, remarque Suber, il apparaît que l'argument est non seulement valide mais bien fondé (*sound*). Car lorsque les prémisses sont vraies, la conclusion de l'argument peut en être inférée valablement. En revanche, comme le fait remarquer Suber, l'argument pêche par le fait qu'un certain nombre de prémisses font défaut. Ce point est essentiel, car si ces prémisses manquantes sont replacées au sein de l'argument, la conclusion qui en résulte peut se révéler radicalement différente.

4. Instance du biais d'uni-polarisation

Afin d'illustrer les notions précédentes, il s'avère intéressant, à ce stade, de donner un exemple du biais d'uni-polarisation. À cette fin, considérons l'instance suivante, qui consiste en une forme de raisonnement, mentionnée par Philippe Boulanger (2000, p. 3)[3], qui l'attribue au mathématicien Stanislas Ulam. Le biais d'uni-polarisation s'y manifeste sous une forme déductive. Ulam estime ainsi que si une entreprise devait atteindre un niveau de main d'oeuvre suffisamment important, son niveau de performance serait paralysé par le grand nombre de conflits internes qui en résulteraient. Ulam estime ainsi que le nombre de conflits entre personnes augmenterait selon le carré du nombre n d'employés, alors que l'impact sur le travail qui en résulterait ne progresserait qu'en fonction de n. Ainsi, selon cet argument, il n'est pas souhaitable que le nombre d'employés au sein d'une entreprise devienne important. Cependant, il s'avère que le raisonnement d'Ulam est fallacieux, comme le souligne Boulanger, car il met exclusivement l'accent sur les relations conflictuelles entre employés. Or les n^2 relations parmi les employés de l'entreprise peuvent être de nature conflictuelle, mais peuvent consister

[3]Philippe Boulanger indique (correspondance personnelle) qu'il a entendu Stanislas Ulam développer ce point particulier lors d'une conférence à l'Université du Colorado.

aussi bien en relations de collaboration tout à fait bénéfiques pour l'entreprise. Et il n'y a donc pas de raison de privilégier les relations conflictuelles par rapport aux relations de collaboration. Et lorsque parmi les n^2 relations qui s'établissent entre les employés de l'entreprise, certaines sont d'authentiques relations de collaboration, cela a pour effet, au contraire, d'améliorer la performance de l'entreprise. Par conséquent, on ne peut pas conclure légitimement qu'il n'est pas souhaitable que l'effectif d'une entreprise atteigne une taille importante.

Dans un souci de clarté, il s'avère utile de formaliser quelque peu le raisonnement précédent. Il apparaît ainsi que le raisonnement d'Ulam peut être présenté de la manière suivante :

> ($D1_{\bar{A}}$) si <une entreprise présente un nombre important d'employés>
>
> ($D2_{\bar{A}}$) alors <il en résultera n^2 relations conflictuelles>
>
> ($D3_{\bar{A}}$) alors des effets négatifs en résulteront
>
> ($D4_{\bar{A}}$) $\therefore$ le fait qu' <une entreprise ait un nombre important d'employés> est mauvais

Ce type de raisonnement présente la structure d'un biais d'uni-polarisation, car il met uniquement l'accent sur les relations conflictuelles (pôle de *dissociation* dans la dualité *association/dissociation*), en passant sous silence un argument parallèle présentant la même structure qui pourrait être légitimement soulevé, mettant l'accent sur les relations de collaboration (pôle *d'association*), qui constituent l'autre aspect pertinent sur ce sujet particulier. Cet argument parallèle est le suivant :

> ($D1_A$) si <une entreprise présente un nombre important d'employés>
>
> ($D2_A$) alors <il en résultera n^2 relations de collaboration>
>
> ($D3_A$) alors des effets positifs en résulteront
>
> ($D4_A$) $\therefore$ le fait qu' <une entreprise ait un nombre important d'employés> est bon

Ceci met finalement en lumière comment les deux formulations de l'argument conduisent à des conclusions contradictoires, c'est-à-dire ($D4_{\bar{A}}$) et ($D4_A$). À ce stade, il est utile de souligner la structure-même de la conclusion du raisonnement ci-dessus, qui est la suivante :

($D5_{\bar{A}}$) la situation *s* est mauvaise du point de vue $\bar{A}$ (*dissociation*)

alors que la conclusion du raisonnement parallèle est la suivante :

($D5_A$) la situation *s* est bonne du point de vue A (*association*)

Mais si le raisonnement avait été complet, en prenant en compte les deux points de vue, une autre conclusion en aurait résulté :

($D5_{\bar{A}}$) la situation *s* est mauvaise du point de vue $\bar{A}$ (*dissociation*)

($D5_A$) la situation *s* est bonne du point de vue A (*association*)

($D6_{A/\bar{A}}$) la situation *s* est mauvaise du point de vue $\bar{A}$ (*dissociation*) et bonne du point de vue A (*association*)

($D7_{A/\bar{A}}$) la situation *s* est neutre du point de vue de la dualité $A/\bar{A}$ (*association/dissociation*)

Et une telle conclusion s'avère tout à fait différente de celle qui résulte de ($D5_{\bar{A}}$) et de ($D5_A$).

Finalement, nous sommes en mesure de replacer le biais d'uni-polarisation qui vient d'être décrit dans le cadre du présent modèle : l'objet *o* est le raisonnement précité, la classe de référence est celle des relations existant entre les employés d'une entreprise, et la dualité correspondante - permettant de définir la classe de référence - est la dualité *dissociation/association*.

5. Analyse dichotomique et méta-philosophie

Le principe d'indifférence dialectique précité et son corollaire - le biais d'uni-polarisation - est susceptible de trouver des applications dans plusieurs domaines[4]. Nous nous intéresserons, dans ce qui suit, à ses applications, à un

[4]Une application de la présente construction aux *distorsions cognitives*, introduites par Aaron Beck (1963, 1964) dans les éléments constitutifs de la thérapie cognitive, est donnée dans Franceschi (2007). Les distorsions cognitives sont classiquement définies comme des raisonnements fallacieux jouant un rôle déterminant dans l'émergence d'un certain nombre de troubles mentaux. La thérapie cognitive en particulier se fonde sur l'identification de ces distorsions cognitives dans le raisonnement usuel du patient, et leur remplacement par des raisonnements alternatifs. Classiquement, les distorsions cognitives sont décrites comme l'un des douze modes de raisonnement irrationnel suivants : 1. Raisonnement émotionnel 2. Hyper-généralisation 3. Inférence arbitraire 4. Raisonnement dichotomique 5. Obligations injustifiées (*Should statements*, (Ellis 1962)) 6. Divination ou lecture mentale 7. Abstraction sélective 8. Disqualification du positif 9. Maximisation et minimisation 10. Catastrophisme 11. Personnalisation 12. Étiquetage.

niveau méta-philosophique, à travers l'analyse de plusieurs paradoxes philosophiques contemporains. La méta-philosophie constitue cette branche de la philosophie dont l'objet est l'étude de la nature de la philosophie, de sa finalité et de ses méthodes propres. Dans ce contexte, un domaine spécifique au sein de la méta-philosophie est celui de la méthode à employer pour s'attacher à résoudre, ou à progresser vers la résolution des paradoxes ou des problèmes philosophiques. C'est dans ce domaine spécifique que s'inscrit la présente construction, en ce sens qu'elle propose l'*analyse dichotomique* comme un outil qui peut se révéler utile pour aider à la résolution de paradoxes ou de problèmes philosophiques.

L'analyse dichotomique, en tant que méthodologie pouvant être utilisée pour la recherche de solutions à certains paradoxes ou problèmes philosophiques, résulte directement de l'énoncé-même du principe d'indifférence dialectique. L'idée générale qui sous-tend la démarche dichotomique d'analyse des paradoxes, est que deux versions, correspondant à l'un et l'autre pôle d'une dualité donnée, peuvent se trouver mêlées dans un paradoxe philosophique. La démarche consiste alors à trouver une classe de référence associée au paradoxe en question et la dualité $A/\bar{A}$ correspondante, ainsi que les deux variations du paradoxe qui en résultent et qui s'appliquent à chacun des pôles de cette dualité. Cependant, toute dualité ne convient pas pour cela, car pour nombre de dualités, la version correspondante du paradoxe demeure inchangée, quel que soit le pôle que l'on envisage. Dans la méthode dichotomique, il s'agit de s'attacher à trouver une classe de référence et une dualité associée pertinente, telle que le point de vue de chacun de ses pôles conduise effectivement à deux versions *structurellement différentes* du paradoxe, ou bien à la disparition du paradoxe selon le point de vue de l'un des pôles. Ainsi, lorsque l'on envisage le paradoxe sous l'angle des deux pôles A et $\bar{A}$, et que cela n'a aucune incidence concernant le paradoxe lui-même, la dualité $A/\bar{A}$ correspondante ne se révèle donc pas, de ce point de vue, pertinente.

L'analyse dichotomique ne constitue pas un outil qui prétend résoudre tous les problèmes philosophiques, loin s'en faut, mais seulement une méthodologie qui est susceptible d'apporter un éclairage pour certains d'entre eux. Dans ce qui suit, nous nous attacherons à illustrer, à travers plusieurs travaux de l'auteur, comment l'analyse dichotomique peut s'appliquer pour progresser vers la résolution de trois paradoxes philosophiques contemporains : le paradoxe de Hempel, le paradoxe de l'examen-surprise et l'argument de l'Apocalypse.

De manière préliminaire, on peut observer ici que dans la littérature, on trouve également un exemple d'analyse dichotomique de paradoxe chez Da-

vid Chalmers (2002). Chalmers s'attache ainsi à montrer comment le *paradoxe des deux enveloppes* comporte deux versions fondamentalement distinctes, dont l'une correspond à une version *finie* du paradoxe et l'autre à une version *infinie*. Une telle analyse, bien que conçue indépendamment de la présente construction, peut ainsi être caractérisée comme une analyse dichotomique fondée sur la dualité *fini/infini*.

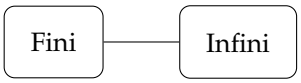


Figure 34.3: Les pôles duaux dans l'analyse de David Chalmers du paradoxe des deux enveloppes

6. Application à l'analyse des paradoxes philosophiques

À ce stade, il convient d'appliquer ce qui précède à l'analyse de problèmes concrets. Nous nous efforcerons ainsi d'illustrer cela à travers l'analyse de plusieurs paradoxes philosophiques contemporains : le paradoxe de Hempel, le paradoxe de l'examen-surprise et l'argument de l'Apocalypse. Nous nous attacherons à montrer comment un problème de biais d'uni-polarisation associé à un problème de définition d'une classe de référence se rencontre dans l'analyse des paradoxes philosophiques précités. En outre, nous montrerons comment la définition-même de la classe de référence associée à chaque paradoxe est susceptible d'être qualifiée à l'aide des pôles duaux A et $\bar{A}$ d'une dualité $A/\bar{A}$ tels qu'ils viennent d'être définis.

Application à l'analyse du paradoxe de Hempel

Le paradoxe de Hempel est basé sur le fait que les deux assertions suivantes :

(H) Tous les corbeaux sont noirs

(H*) Tout ce qui est non-noir est un non-corbeau

sont logiquement équivalentes. Par sa structure, (H*) se présente en effet comme la forme contraposée de (H). Il en résulte que la découverte d'un corbeau noir confirme (H) et également (H*), mais aussi que la découverte d'une chose non-noire qui n'est pas un corbeau telle qu'un flamand rose ou même un parapluie

gris, confirme (H*) et donc (H). Cependant, cette dernière conclusion apparaît comme paradoxale.

Nous nous attacherons maintenant à détailler l'analyse dichotomique sur laquelle se trouve basée la solution proposée dans Franceschi (1999). La démarche se trouve fondée sur la recherche d'une classe de référence associée à l'énoncé du paradoxe, qui est susceptible d'être définie à l'aide d'une dualité $A/\bar{A}$. Si l'on examine ainsi avec soin les concepts et les catégories qui sous-tendent les propositions (H) et (H*), on remarque tout d'abord qu'il en existe quatre : les corbeaux, les objets noirs, les objets non-noirs et les non-corbeaux. Un *corbeau* tout d'abord se trouve défini de manière précise dans la taxinomie au sein de laquelle il s'insère. Une catégorie comme celle des corbeaux peut être considérée comme bien définie, car elle est basée sur un ensemble de critères précis définissant l'espèce *corvus corax* et permettant l'identification de ses instances. De même, la classe des objets *noirs* peut être décrite avec précision, à partir d'une taxinomie des couleurs établie par rapport aux longueurs d'onde de la lumière. Enfin, on peut constater que la classe des objets *non-noirs* peut également faire l'objet d'une définition qui ne souffre pas d'ambiguïté, à partir notamment de la taxinomie précise des couleurs qui vient d'être mentionnée.

En revanche, qu'en est-il de la classe des *non-corbeaux* ? Qu'est-ce qui constitue donc une instance d'un non-corbeau ? Intuitivement, un merle bleu, un flamand rose, un parapluie gris, voire même un entier naturel, constituent des non-corbeaux. Mais doit-on envisager une classe de référence qui aille jusqu'à inclure les objets abstraits ? Faut-il ainsi considérer une notion de *non-corbeau* qui englobe des entités abstraites tels que les entiers naturels et les nombres complexes ? Ou bien convient-il se limiter à une classe de référence qui n'embrasse que les animaux ? Ou doit-on considérer une classe de référence qui englobe tous les êtres vivants, ou bien encore toutes les choses concrètes, incluant cette fois également les artefacts ? Finalement, il en résulte que la proposition (H*) initiale est susceptible de donner lieu à plusieurs variations, qui sont les suivantes :

(H_1*) Tout ce qui est non-noir parmi les *corvidés* est un non-corbeau

(H_2*) Tout ce qui est non-noir parmi les *oiseaux* est un non-corbeau

(H_3*) Tout ce qui est non-noir parmi les *animaux* est un non-corbeau

(H_4*) Tout ce qui est non-noir parmi les *êtres vivants* est un non-corbeau

(H_5*) Tout ce qui est non-noir parmi les *choses concrètes* est un non-corbeau

(H_6*) Tout ce qui est non-noir parmi les *objets concrets et abstraits* est un non-corbeau

Ainsi, il apparaît que l'énoncé du paradoxe de Hempel et en particulier la proposition (H*) se trouve associée à une *classe de référence*, qui permet de définir les *non-corbeaux*. Une telle classe de référence peut s'assimiler aux corvidés, aux oiseaux, aux animaux, aux êtres vivants, aux choses concrètes, ou encore aux choses concrètes et abstraites, etc. Cependant, dans l'énoncé du paradoxe de Hempel, on ne dispose pas de critère objectif permettant d'effectuer un tel choix. À ce stade, il apparaît que l'on peut choisir une telle classe de référence de manière *restrictive*, par exemple en l'assimilant aux corvidés. Mais de manière aussi légitime, on peut choisir une classe de référence de manière plus *extensive*, par exemple en l'identifiant à l'ensemble des choses concrètes, incluant alors notamment les parapluies. Alors pourquoi choisir telle classe de référence définie de manière restrictive plutôt que telle autre définie de façon extensive ? On ne possède pas en réalité de critère pour légitimer le choix, selon que l'on procède par *restriction* ou par *extension*, de la classe de référence. Dès lors, il apparaît que celle-ci ne peut être définie que de manière *arbitraire*. Or le choix d'une telle classe de référence se révèle déterminant, car selon que l'on choisira telle ou telle classe de référence, un objet donné tel qu'un parapluie gris confirmera ou non (H*) et donc (H). Ainsi, si nous choisissons la classe de référence par extension, incluant ainsi l'ensemble des objets concrets, un parapluie gris confirmera (H). Cependant, si nous choisissons une telle classe de référence par restriction, en l'assimilant seulement aux corvidés, un parapluie gris ne confirmera pas (H). Une telle différence se révèle essentielle. En effet, si l'on choisit une définition extensive de la classe de référence, on a bien l'effet paradoxal inhérent au paradoxe de Hempel. Mais dans le cas contraire, si l'on opte pour une classe de référence définie de manière restrictive, on perd alors l'effet paradoxal.

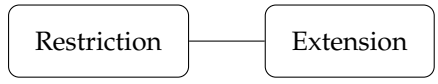


Figure 34.4: Pôles duaux au sein de la classe de référence des non-corbeaux dans le paradoxe de Hempel

Ce qui précède permet de décrire avec précision les éléments de l'analyse qui précède du paradoxe de Hempel, en termes de biais d'uni-polarisation ainsi qu'il a été défini plus haut : au paradoxe et en particulier à la proposition (H*) se trouve associée la classe de référence des *non-corbeaux*, qui est elle-même susceptible d'être définie par rapport à la dualité *extension/restriction*. Or, pour un objet donné tel qu'un parapluie gris, la définition de la classe de référence par extension donne lieu à un effet paradoxal, alors-même que le choix de cette dernière par restriction ne conduit pas à un tel effet.

Application à l'analyse du paradoxe de l'examen-surprise

La version classique du paradoxe de l'examen-surprise (Quine 1953, Sorensen 1988) est la suivante : un professeur annonce à ses étudiants qu'un examen aura lieu la semaine prochaine, mais qu'ils ne pourront pas connaître à l'avance le jour précis où l'examen se déroulera. L'examen aura donc lieu par surprise. Les étudiants raisonnent ainsi. L'examen ne peut avoir lieu le samedi, pensent-ils, car sinon ils sauraient à l'avance que l'examen aurait lieu le samedi et donc il ne pourrait survenir par surprise. Aussi le samedi se trouve-t-il éliminé. De plus, l'examen ne peut avoir lieu le vendredi, car sinon les étudiants sauraient à l'avance que l'examen aurait lieu le vendredi et donc il ne pourrait survenir par surprise. Aussi le vendredi se trouve-t-il également éliminé. Par un raisonnement analogue, les étudiants éliminent successivement le jeudi, le mercredi, le mardi et le lundi. Finalement, ce sont tous les jours de la semaine qui sont ainsi éliminés. Toutefois, cela n'empêche pas l'examen de survenir finalement par surprise, le mercredi. Ainsi, le raisonnement des étudiants s'est avéré fallacieux. Pourtant, un tel raisonnement paraît intuitivement valide. Le paradoxe réside ici dans le fait que le raisonnement des étudiants est semble-t-il valide, alors qu'il se révèle finalement en contradiction avec les faits, à savoir que l'examen peut véritablement survenir par surprise, conformément à l'annonce faite par le professeur.

Afin de présenter l'analyse dichotomique (Franceschi 2005) qui peut être effectuée par rapport au paradoxe de l'examen-surprise, il convient de considérer tout d'abord deux variations qui apparaissent structurellement différentes du paradoxe. Une première variation est associée à la solution au paradoxe proposée par Quine (1953). Quine considère ainsi la conclusion finale de l'étudiant selon laquelle l'examen ne peut avoir lieu par surprise aucun jour de la semaine. Selon Quine, l'erreur de l'étudiant réside dans le fait de n'avoir pas envisagé dès le début l'hypothèse selon laquelle l'examen pourrait ne pas avoir lieu le dernier jour. Car le fait de considérer précisément que l'examen

n'aura pas lieu le dernier jour permet finalement à l'examen de survenir par surprise, le dernier jour. Si l'étudiant avait également pris en compte cette possibilité dès le début, il ne serait pas parvenu à la conclusion fallacieuse que l'examen ne peut pas survenir par surprise.

La seconde variation du paradoxe qui se révèle intéressante dans le présent contexte, est celle qui est associée à la remarque, effectuée par plusieurs auteurs (Hall 1999, p. 661, Williamson 2000), selon laquelle le paradoxe émerge nettement, lorsque le nombre n d'unités est grand. Un tel nombre est habituellement associé à un nombre n de jours, mais on peut aussi bien utiliser des heures, des minutes, des secondes, etc. Une caractéristique intéressante du paradoxe est en effet que celui-ci émerge intuitivement de manière plus nette lorsque de grandes valeurs de n sont prises en compte. Une illustration frappante de ce phénomène nous est ainsi fournie par la variation du paradoxe qui correspond à la situation suivante, décrite par Timothy Williamson (2000, p. 139) :

> Advance knowledge that there will be a test, fire drill, or the like of which one will not know the time in advance is an everyday fact of social life, but one denied by a surprising proportion of early work on the Surprise Examination. Who has not waited for the telephone to ring, knowing that it will do so within a week and that one will not know a second before it rings that it will ring a second later ?

La variation décrite par Williamson correspond à l'annonce faite à quelqu'un qu'il recevra un coup de téléphone dans la semaine, sans pouvoir toutefois déterminer à l'avance à quelle seconde précise ce dernier événement surviendra. Cette variation souligne comment la surprise peut se manifester, de manière tout à fait plausible, lorsque la valeur de n est élevée. L'unité de temps considérée par Williamson est ici la seconde, rapportée à une période qui correspond à une semaine. La valeur correspondante de n est ici très élevée et égale à 604800 (60 x 60 x 24 x 7) secondes. Cependant, il n'est pas indispensable de prendre en compte une valeur aussi grande de n, et une valeur de n égale par exemple à 365 convient également très bien.

Le fait que deux versions qui semblent a priori assez différentes du paradoxe coexistent, suggère que deux versions structurellement différentes du paradoxe pourraient se trouver inextricablement mêlées dans le paradoxe de l'examen-surprise. De fait, si l'on analyse la version du paradoxe qui donne lieu à la solution de Quine, on s'aperçoit qu'elle présente une particularité :

elle est susceptible de se manifester pour une valeur de n égale à 1. La version correspondante de l'annonce du professeur est alors la suivante : « Un examen aura lieu demain, mais vous ne pourrez savoir à l'avance que cet examen aura lieu et par conséquent, il surviendra par surprise. » L'analyse de Quine s'applique directement à cette version du paradoxe pour laquelle $n = 1$. Dans ce cas, l'erreur de l'étudiant réside, selon Quine, dans le fait de n'avoir considéré que la seule hypothèse suivante : (a) « l'examen aura lieu demain et je prévoirai qu'il aura lieu ». En fait, l'étudiant aurait dû considérer également trois autres cas : (b) « l'examen n'aura pas lieu demain et je prévoirai qu'il aura lieu » ; (c) « l'examen n'aura pas lieu demain et je ne prévoirai pas qu'il aura lieu » ; (d) « l'examen aura lieu demain et je ne prévoirai pas qu'il aura lieu ». Et le fait de considérer l'hypothèse (a) mais également l'hypothèse (d) qui est compatible avec l'annonce du professeur aurait empêché l'étudiant de conclure que l'examen n'aurait finalement pas lieu. Par conséquent, souligne Quine, c'est le fait de n'avoir pris en considération que l'hypothèse (a) qui peut être identifié comme la cause du raisonnement fallacieux.

On le voit, la structure-même de la version du paradoxe sur laquelle est fondée la solution de Quine présente les particularités suivantes : d'une part, la non-surprise peut effectivement survenir le dernier jour, et d'autre part, l'examen peut également survenir par surprise le dernier jour. Il en va de même pour la version du paradoxe où $n = 1$: la non-surprise ainsi que la surprise peuvent survenir le jour n. Ceci permet de représenter une telle structure du paradoxe sous forme de la matrice S[k, s] suivante (où k dénote le jour où l'examen a lieu et S[k, s] dénote si le cas correspondant de non-surprise ($s = 0$) ou de surprise ($s = 1$) est rendu possible (dans ce cas, S[k, s] = 1) ou non (dans ce cas, S[k, s] = 0)) :

jour	non-surprise	surprise
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1

Structure matricielle de la version du paradoxe correspondant à la solution de Quine pour n = 7 (une semaine)

jour	non-surprise	surprise
1	1	1

Structure matricielle de la version du paradoxe correspondant à la solution de Quine pour n = 1 (un jour)

Compte tenu de la structure correspondante de la matrice qui admet des valeurs égales à 1 à la fois au niveau des cas de non-surprise et de surprise, pour un jour donné, nous dénommerons *conjointe* une telle structure de matrice.

Si l'on étudie la variation du paradoxe énoncée par Williamson et mentionnée plus haut, elle présente la particularité, à l'inverse de la variation précédente, d'émerger de manière nette lorsque n est grand. Dans ce contexte, l'annonce du professeur correspondante par exemple à une valeur de n égale à 365, est la suivante : « Un examen aura lieu dans l'année à venir mais la date de l'examen constituera une surprise ». Si l'on analyse une telle variation en termes de matrice des cas de non-surprise et de surprise, il apparaît qu'une telle version du paradoxe présente les propriétés suivantes : la non-surprise ne peut survenir le 1er jour alors que la surprise est possible ce même 1er jour ; en revanche, le dernier jour, la non-surprise est possible alors que la surprise n'est pas possible.

jour	non-surprise	surprise
1	0	1
...	...	...
365	1	0

Structure matricielle de la version du paradoxe correspondant à la variation de Williamson pour n = 365 (un an)

Ce qui précède permet maintenant d'identifier avec précision ce qui pêche dans le raisonnement de l'étudiant, lorsqu'il s'applique à cette version particulière du paradoxe. Dans ces circonstances, l'étudiant aurait alors dû raisonner de la manière suivante. La surprise ne peut se manifester le dernier jour mais peut survenir le 1er jour ; la non-surprise peut se manifester le dernier jour, mais ne peut survenir le 1er jour. Il s'agit ici d'instances propres de

non-surprise et de surprise, qui se révèlent disjointes. Cependant, la notion de surprise n'est pas capturée de manière exhaustive par l'extension et l'anti-extension de la surprise. Or une telle définition est conforme à la définition d'un prédicat vague, qui se caractérise par une extension et une anti-extension mutuellement exclusives et non-exhaustives. Ainsi, la conception de la surprise associée une structure disjointe est-elle celle d'une notion *vague*. Aussi l'erreur à l'origine du raisonnement fallacieux de l'étudiant réside-t-elle dans l'absence de prise en compte du fait que la surprise correspond dans le cas d'une structure disjointe, à une notion vague, et comporte donc la présence d'une zone de pénombre correspondant à des cas-limites (*borderline*) entre la non-surprise et la surprise. Car la seule prise en compte du fait que la notion de surprise est ici une notion vague aurait interdit à l'étudiant de conclure que $S[k, 1] = 0$, pour toutes les valeurs de k, c'est-à-dire que l'examen ne peut survenir par surprise aucun jour de la période considérée.

Finalement, il apparaît ainsi que l'analyse conduit à distinguer au niveau du paradoxe de l'examen-surprise deux variations indépendantes. La définition matricielle des cas de non-surprise et de surprise conduit à distinguer deux variations du paradoxe, en fonction de la dualité *conjoint/disjoint*. Dans un premier cas, le paradoxe est basé sur une définition *conjointe* des cas de non-surprise et de surprise. Dans un second cas, le paradoxe se trouve fondé sur une définition *disjointe*. Chacune de ces deux variations conduit à une variation structurellement différente du paradoxe et à une solution indépendante. Lorsque la variation du paradoxe est basée sur une définition *conjointe*, la solution développée par Quine s'applique alors. En revanche, lorsque la variation, du paradoxe est fondée sur une définition *disjointe*, la solution retenue est fondée sur la reconnaissance préalable de la nature vague de la notion de surprise associée à cette variation du paradoxe.

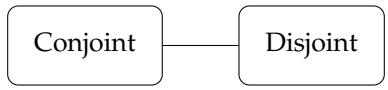


Figure 34.5: Pôles duaux au sein de la classe des matrices associées au paradoxe de l'examen-surprise

On le voit finalement, l'analyse dichotomique du paradoxe de l'examen-surprise conduit à envisager la classe des matrices associées à la définition-même du paradoxe et à distinguer selon que leur structure est conjointe ou

bien disjointe. Dès lors, il en résulte une solution indépendante pour chacune des deux versions structurellement différentes du paradoxe qui en résultent.

Application à l'analyse de l'Argument de l'Apocalypse

L'*argument de l'Apocalypse*, attribué à Brandon Carter, a été décrit par John Leslie (1993, 1996). Il convient d'en rappeler préalablement l'énoncé. Considérons la proposition (A) suivante :

(A) L'espèce humaine disparaîtra avant la fin du XXIème siècle

On peut estimer, pour fixer les idées, à une chance sur 100 la probabilité que cette disparition survienne : P(A) = 0,01. Soit également la proposition suivante :

(Ā) L'espèce humaine ne disparaîtra pas à la fin du XXIème siècle

Soit encore E l'événement : je vis durant les années 2010. On peut par ailleurs estimer aujourd'hui à 60 milliards le nombre d'humains ayant existé depuis la naissance de l'humanité. De même, la population actuelle peut être évaluée à 6 milliards. On calcule ainsi qu'un humain sur dix, si l'événement A survient, aura connu les années 2010. On évalue alors la probabilité que l'humanité soit éteinte avant la fin du XXIème siècle, si j'ai connu les années 2010 : $P(E, A) = 6x10^9/6x10^{10} = 0,1$. Par contre, si l'humanité passe le cap du XXIème siècle, on peut penser qu'elle sera appelée à une expansion beaucoup plus importante, et que le nombre des humains pourra s'élever par exemple à $6x10^{12}$. Dans ce cas, la probabilité que l'humanité ne soit pas éteinte à la fin du XXIème siècle, si j'ai connu les années 2010 s'évalue ainsi : $P(E, \bar{A}) = 6x10^9/6x10^{12} = 0,001$. À ce stade, nous pouvons assimiler à deux urnes distinctes - l'une contenant 60 milliards de boules et l'autre en comportant 6000 milliards - les populations humaines totales qui en résultent. Ceci conduit à calculer la probabilité a posteriori de l'extinction de l'espèce humaine avant la fin du XXIème siècle, à l'aide de la formule de Bayes : $P'(A) = [P(A) \times P(E, A)] / [P(A) \times P(E, A) + P(\bar{A}) \times P(E, \bar{A})] = (0,01 \times 0,1) / (0,01 \times 0,1 + 0,99 \times 0,001) = 0,5025$. Ainsi, la prise en compte du fait que je vis actuellement fait passer la probabilité de l'extinction de l'espèce humaine avant 2150 de 1 % à 50,25 %. Une telle conclusion apparaît comme contraire à l'intuition et en ce sens, paradoxale.

Il convient maintenant de s'attacher comment une analyse dichotomique (Franceschi 1999, 2009) peut s'appliquer à l'argument de l'Apocalypse. En premier lieu, nous nous attacherons à montrer comment l'argument de l'Apoca-

lypse comporte un problème de définition de *classe de référence*[5] liée à une dualité A/Ā. Considérons en effet l'assertion suivante :

(A) L'espèce humaine disparaîtra avant la fin du XXIème siècle

Une telle proposition présente une connotation dramatique, apocalyptique et tragique, liée à la disparition très prochaine de l'espèce humaine. Il s'agit là d'une prédiction de nature tout à fait catastrophique et alarmante. Cependant, si on analyse une telle proposition avec soin, on est conduit à remarquer qu'elle comporte une imprécision. Si la référence temporelle elle-même - la fin du XXIème siècle - se révèle tout à fait précise, le terme d' « espèce humaine » proprement dit apparaît comme ambigu. En effet, il s'avère qu'il existe plusieurs façons de définir cette dernière. La notion la plus précise permettant de définir l' « espèce humaine » est notre présente taxinomie scientifique, basée sur les notions de genre, d'espèce, de sous-espèce, etc. En adaptant cette dernière taxinomie à l'assertion (A), il s'ensuit que la notion ambiguë d' « espèce humaine » est susceptible d'être définie par rapport au genre, à l'espèce, à la sous-espèce, etc. et en particulier par rapport au genre *homo*, à l'espèce *homo sapiens*, à la sous-espèce *homo sapiens sapiens*, etc. Finalement, il s'ensuit que l'assertion (A) est susceptible de revêtir les formes suivantes :

(A_h) Le genre *homo* disparaîtra avant la fin du XXIème siècle

(A_{hs}) L'espèce *homo sapiens* disparaîtra avant la fin du XXIème siècle

(A_{hss}) La sous-espèce *homo sapiens sapiens* disparaîtra avant la fin du XXIème siècle

À ce stade, la lecture de ces différentes propositions conduit à un impact différent, eu égard à la proposition initiale (A). Car si (A_h) présente bien à l'instar de (A) une connotation tout à fait dramatique et tragique, il n'en va pas de même pour (A_{hss}). En effet, une telle proposition qui prévoit l'extinction de notre sous-espèce actuelle *homo sapiens sapiens* avant la fin du XXIème siècle, pourrait s'accompagner du remplacement de notre actuelle race humaine par

[5] L'analyse de l'argument de l'Apocalypse du point de vue du problème de la classe de référence est effectuée de manière détaillée par Leslie (1996). Mais l'analyse de Leslie vise à montrer que le choix de la classe de référence, par extension ou par restriction, n'a pas d'incidence sur la conclusion de l'argument lui-même.

une nouvelle sous-espèce plus évoluée, que l'on pourrait dénommer *homo sapiens supersapiens*. Dans ce cas, la proposition (A_{hss}) ne comporterait pas de connotation tragique, mais serait associée à une connotation positive, car le remplacement d'une race ancienne par une espèce plus évoluée constitue un processus naturel de l'évolution. Plus encore, en choisissant une classe de référence encore plus restreinte telle que celle des humains n'ayant pas connu l'ordinateur (*homo sapiens sapiens antecomputeris*), on obtient la proposition suivante :

> (A_{hsss}) L'infra-sous-espèce *homo sapiens sapiens antecomputeris* disparaîtra avant la fin du XXIème siècle

qui ne présente plus du tout la connotation dramatique inhérente à (A) et qui se révèle même tout à fait normale et rassurante, et qui ne présente plus aucun caractère paradoxal ni contraire à l'intuition. Dans ce cas en effet, la disparition de l'infra-sous-espèce *homo sapiens sapiens antecomputeris* s'accompagne de la survie de l'*infra-sous-espèce* plus évoluée *homo sapiens sapiens postcomputeris*. Il s'avère ainsi qu'un classe de référence restreinte coïncidant avec une infra-sous-espèce est définitivement éteinte, mais qu'une classe plus étendue correspondant à une sous-espèce (*homo sapiens sapiens*) survit. Dans ce cas, on observe bien le décalage bayesien décrit par Leslie, mais l'effet de ce décalage se révèle cette fois tout à fait inoffensif.

Ainsi, le choix de la classe de référence pour la proposition (A) se révèle-t-il déterminant pour la nature paradoxale de la conclusion associée à l'argument de l'Apocalypse. Si l'on choisit ainsi une classe de référence étendue pour la définition-même des humains, en l'associant par exemple au genre *homo*, on conserve le caractère dramatique et inquiétant associé à la proposition (A). Mais si on choisit une telle classe de référence de manière restrictive, en l'associant par exemple à l'infra-sous-espèce *homo sapiens sapiens antecomputeris*, une nature rassurante et normale se trouve désormais associée à la proposition (A) qui sous-tend l'argument de l'Apocalypse.

Finalement, nous sommes en mesure de replacer l'analyse qui précède dans le présent contexte. La définition-même de la classe de référence des « humains » associée à la proposition (A) inhérente à l'argument de l'Apocalypse est susceptible d'être définie selon les pôles de la dualité *extension / restriction*. Une analyse fondée sur un point de vue bi-polarisé conduit à constater que le choix par extension entraîne un effet paradoxal, alors-même que le choix par restriction de la classe de référence fait disparaître ce même effet paradoxal.

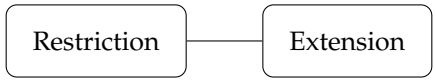


Figure 34.6: Pôles duaux au sein de la classe de référence des « humains » dans l'Argument de l'Apocalypse

L'analyse dichotomique, toutefois, en ce qui concerne l'argument de l'Apocalypse, ne se limite pas à cela. En effet, si on étudie l'argument avec soin, il apparaît qu'il recèle une autre classe de référence associée à une autre dualité. Ceci peut être mis en évidence en analysant l'argument opposé par William Eckhardt (1993, 1997) à l'argument de l'Apocalypse. Selon Eckhardt, la situation humaine correspondant à DA n'est pas analogue au modèle des deux urnes décrit par Leslie, mais plutôt à un modèle alternatif, qui peut être appelé le distributeur d'objets consécutifs (*consecutive token dispenser*). Le distributeur d'objets consécutifs est un dispositif qui éjecte à intervalles réguliers des boules numérotées consécutivement : « (...) suppose on each trial the consecutive token dispenser expels either 50 (early doom) or 100 (late doom) consecutively numbered tokens at the rate of one per minute ». S'appuyant sur ce modèle, Eckhardt (1997, p. 256) souligne le fait qu'il est impossible d'effectuer une sélection aléatoire, dès lorsqu'il existe de nombreux individus qui ne sont pas encore nés au sein de la classe de référence correspondante : « How is it possible in the selection of a random rank to give the appropriate weight to unborn members of the population ? ». L'idée forte d'Eckhardt qui sous-tend cette objection diachronique est qu'il est impossible d'effectuer une sélection aléatoire lorsqu'il existe de nombreux membres au sein de la classe de référence qui ne sont pas encore nés. Dans une telle situation, il serait tout à fait erroné de conclure à un décalage bayesien en faveur de l'hypothèse (A). En revanche, ce que l'on peut inférer de manière rationnelle dans un tel cas, c'est que la probabilité initiale demeure inchangée.

À ce stade, il apparaît que deux modèles alternatifs pour modéliser l'analogie avec la situation humaine correspondant à l'argument de l'Apocalypse se trouvent en concurrence : d'une part le modèle à caractère synchronique (où toutes les boules sont présentes dans l'urne au moment où s'effectue le tirage) préconisé par Leslie et d'autre part, le modèle diachronique d'Eckhardt, où des boules peuvent être ajoutées dans l'urne après le tirage. La question qui se pose est la suivante : la situation humaine correspondant à l'argument de l'Apocalypse est-elle en analogie avec (a) le modèle de l'urne synchronique, ou bien avec (b) le modèle de l'urne diachronique ? Afin d'y répondre, la ques-

tion suivante s'ensuit : existe-t-il un critère objectif qui permette de choisir, de manière préférentielle, entre les deux modèles concurrents ? Il apparaît que non. En effet, ni Leslie ni Eckhardt ne présentent une motivation objective qui permette de justifier le choix du modèle qu'ils préconisent, et d'écarter le modèle alternatif. Dans ces circonstances, le choix de l'un ou l'autre des deux modèles - synchronique ou diachronique - apparaît comme arbitraire. Par conséquent, il s'avère que le choix au sein de la classe des modèles associée à l'argument de l'Apocalypse est susceptible d'être défini selon les pôles de la dualité *synchronique/diachronique*. Et une analyse fondée sur un point de vue bi-polarisé conduit à constater que le choix du modèle synchronique conduit à un effet paradoxal, alors-même que le choix du modèle diachronique fait disparaître ce dernier effet paradoxal.

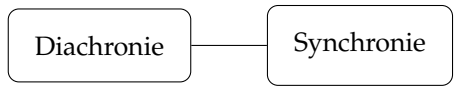


Figure 34.7: Pôles duaux au sein de la classe des modèles de l'Argument de l'Apocalypse

Finalement, compte tenu du fait que le problème précité concernant la classe de référence des *humains* et le choix dans la dualité *extension/restriction* qui lui est associé, ne concerne que le modèle synchronique, la structure de l'analyse dichotomique à un double niveau concernant l'argument de l'Apocalypse, peut être représentée de la manière suivante :

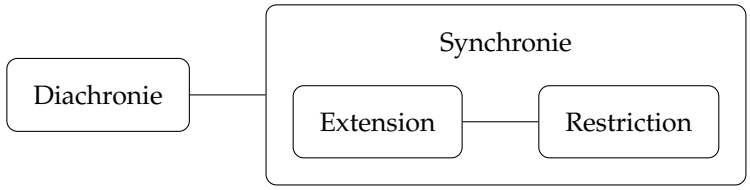


Figure 34.8: Structure de pôles duaux imbriqués Diachronie/Synchronie et Extension/Restriction pour l'Argument de l'Apocalypse

On le voit, les développements qui précèdent mettent en oeuvre la forme de contextualisme dialectique qui a été décrite plus haut, en l'appliquant à

l'analyse de trois paradoxes philosophiques contemporains. Dans le paradoxe de Hempel, à la proposition (H*) se trouve associée la classe de référence des non-corbeaux, qui est elle-même susceptible d'être définie par rapport à la dualité *extension/restriction*. Or, pour un objet *x* donné tel qu'un parapluie gris, la définition de la classe de référence par extension donne lieu à un effet paradoxal, alors-même que le choix de cette dernière par restriction élimine un tel effet. En second lieu, les structures matricielles associées au paradoxe de l'examen-surprise sont analysées sous l'angle de la dualité *conjoint/disjoint*, mettant ainsi en évidence deux versions structurellement distinctes du paradoxes, qui admettent elles-mêmes deux résolutions indépendantes. Enfin, au niveau de l'argument de l'Apocalypse, une analyse dichotomique double met en évidence que la classe des humains est liée à la dualité *extension/restriction*, et que l'effet paradoxal qui est manifeste lorsque la classe de référence est définie par extension, se dissout dès lors que cette dernière est définie par restriction. En second lieu, il s'avère que la classe des modèles peut faire l'objet d'une définition selon la dualité *synchronique/diachronique* ; au point de vue synchronique se trouve associé un effet paradoxal, alors que ce même effet disparaît si l'on se place du point de vue diachronique.

7. Références

Beck, AT. (1963) Thinking and depression : Idiosyncratic content and cognitive distortions, *Archives of General Psychiatry*, 9, 324-333.

Beck ,AT. (1964) Thinking and depression : Theory and therapy, *Archives of General Psychiatry*, 10, 561-571.

Blair, J. Anthony (1988) What Is Bias ? in *Selected Issues in Logic and Communication*, ed. Trudy Govier, Belmont, CA : Wadsworth, 1988, 101–102).

Boulanger, P. (2000) Culture et nature, *Pour la Science*, 273, 3.

Chalmers, D. (2002) The St. Petersburg two-envelope paradox, *Analysis*, 62 : 155-157.

Eckhardt, W. (1993) Probability Theory and the Doomsday Argument, *Mind*, 102, 483-488.

Eckhardt, W. (1997) A Shooting-Room view of Doomsday, *Journal of Philosophy*, 94, 244-259.

Ellis, A. (1962) *Reason and Emotion in Psychotherapy*, Lyle Stuart, New York.

Franceschi, P. (1999). Comment l'urne de Carter et Leslie se déverse dans celle de Carter, *Canadian Journal of Philosophy*, 29, 139-156.

Franceschi, P. (2002) Une classe de concepts, *Semiotica*, 139 (1-4), 211-226.

Franceschi, P. (2005) Une analyse dichotomique du paradoxe de l'examen surprise, *Philosophiques*, 32-2, 399-421.

Franceschi, P. (2007) Compléments pour une théorie des distorsions cognitives, *Journal de Thérapie Comportementale et Cognitive*, *17-2*, 84-88. Preprint in English : www.cogprints.org/5261/

Franceschi, P. (2009) A Third Route to the Doomsday Argument, *Journal of Philosophical Research*, 34, 263-278.

Hall, N. (1999) How to Set a Surprise Exam, *Mind*, 108, 647-703.

Leslie, J. (1993) Doom and Probabilities, *Mind*, 102, 489-491.

Leslie, J. (1996) *The End of the World : the science and ethics of human extinction*, London : Routledge

Quine, W. (1953) On a So-called Paradox, *Mind*, 62, 65-66.

Sorensen, R. A. (1988) *Blindspots*, Oxford : Clarendon Press.

Stuart Mill, J. (1985) *On Liberty*, London : Penguin Classics, original publication in 1859.

Suber, E. (1998). *The One-Sidedness Fallacy.* Manuscript, http ://www.earlham.edu/˜peters/courses/inflogic/onesided.htm. Retrieved 11/25/2012

Walton, D. (1999) *One-Sided Arguments : A Dialectical Analysis of Bias*, Albany : State University of New York Press.

Williamson, T. (2000) *Knowledge and its Limits*, London & New York : Routledge.

35

Intentionality as a Genuine Relation (All You Need is Love)

FRANÇOIS CLEMENTZ

Intentionality is commonly defined either as the relational " property " that most mental states have to refer to, or to be about, something external to themselves, or simply as this " aboutness " relation as such. A seemingly equivalent idea, which is part and parcel of Brentano's heritage although it could be in fact traced back to such late Medieval philosophers as, *e.g.*,Thierry de Freiberg, is that of an intentional state as being " directed " at its target-object.

As a French philosopher who, while a student nearly half-a-century ago, was first exposed to Sartre's and other such phenomelogico-existentialist subjectivist metaphors about consciousness " aiming at " its intentional object, or about intentionality itself as some kind of unlikely *ex-stasis*, I must confess that, for many years, I have remained somewhat suspicious towards the very idea of an intrinsic " direction ", or " sense ", of mental acts. More recently, however, I came to realize that such misleading metaphors should be peeled off from the kernel of truth which they tend to conceal and which lied, in part, at the heart of the Medieval account of intentional relations as " unilateral " (or " non-mutual ").

Is intentionality, really, a full-blooded *relation* ? It is the first and main contention of this paper that some, though presumably not all, mental states are, indeed, genuinely relational. I shall then further argue – in contradistinction, particularly, to Ingvar Johansson whith whom I am, nevertheless, in full agreement for the remaining of this matter– that the relation invoved is endowed with an *intrinsic*, non derivative, asymmetry and direction.

1. The problem of intentionality in a nutshell

It is clearly not the aim of this short note about intentional *relations* to provide a comprehensive account of intentionality as such. Were I to offer such an overall account, I suppose that I would have to draw at least a rough sketch of the complicated genealogy of this concept across centuries, beginning with Aristotle's *De Anima* and then proceeding, say, from Ibn Senna's *mana*, through the Medieval's theory of " intentions ", towards Brentano's modern rediscovery (as well as re-interpretation) of the Aristotelo-Scholastic tradition - with its famous and ambiguous focus on the intentional " inexistence " of objects of thought – and further on, *via* Chislhom and many others, until the late XX th century's debates about the so-called " naturalisation " of intentionality.

Although he never used the term himself (the paternity of which, it seems, should be attributed to Husserl), it was clearly Franz Brentano who famously gave birth to the modern concept of intentionality. Of course, in the much-quoted passage of his *Psychology from an Empirical Standpoint* which is usually invoked in this context, Brentano explicitly refers to the Scholastic tradition, as well as to Aristotle himself, although there also many reasons to consider that his own approach actually follows a quite different path. Suffice it to say that whereas most contemporary philosophers would regard the " problem of intentionality " as belonging primarily to the philosophy of mind – with a few of them taking into account, however, its metaphysical import -, there are some good reasons to think that for the Medievals, or at any rate from Aquinas' and his immediate followers' " realist " standpoint, the issue was basically a concern for epistemology.

Brentano's own initial characterization of intentionality has been so much commented and discussed that one has got somehow wary of quoting it once again. Yet, this is just what I am to do. According to Brentano,

> " Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to content, direction towards an object (which is not to be understood here as meaning a thing), or immanent objectivity . Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgment something is affirmed or denied, in love loved, in hate hated, in desire

desired and so on ".[1]

I shall limit myself to a few more or less cursory remarks about this over-quoted passage. To begin with, a distinction should certainly be drawn between Brentano's qualification of intentionality, on the one hand, and what has come to be known as " Brentano's thesis " according to which intentionality, thus understood, is both a necessary and sufficient condition for, as well as a principled hall-mark of, the mental as such, on the other hand.

Brentano's thesis is that all, and only, mental states are " intentional " in his sense - or, anyway, appear to have a relational structure. *Prima facie*, it should count as a major rationale in favour of this claim that it provides a twofold uniform account of mental " acts ". First, it is supposed to subsume altogether such various mental states as perceptual or emotional experience, on the one hand, and knowledge, belief, desire and various similar propositional attitudes on the other hand. Second, it is also meant to apply whether the " object " of the intentional so-called " relation " actually exists or not, and to be uniform across both cases. Let's call the cunjunction of the two claims the *principle of uniformity*.

Uniformity is surely a nice thing by itself. But, in the present case, does it amount to a real advantage ? One traditional line of attack relies upon the question whether intentionality, as contrued in genuinely relational terms, is both a necessary and a sufficient condition for mentality as such : what, in particular, about so-called purely " qualitative " states such as raw sensations and would-be *qualia* ? And, on the other hand, *quid* about what Searle defines as " secondary " intentionality with a view, especially, on the linguistic " expression " of our mental states ? These are widely discussed issues that I do not intend, however, to examine in this paper.

More relevant to the present discusion is whether Brentano's thesis really implies, as it would seem, that every mental state is genuinely " about " *O*, *whether O exists or not*.

This leads us back to Brentano's initial description of intentionality, which is, famously enough, at least twice ambiguous. A first well-known source for ambiguity has to do with Brentano's equivocation as between the " object " and the " content " of the intentional " act ". Another concerns even more directly the very notion of an " intentional " state. Does Brentano neo-scholastic idiom mean that the " object " of every *bona fide* intentional state is an imma-

[1] Franz Brentano, *Psychologie vom empirische Standpunk,* Leipzig, 1874) ; *Psychology from an Empirical Standpoint*, p. 88

nent (" in-existent ") " intentional object ", *per forc*e distinct from its purported " real " object ?

The above-quoted few lines could seem to encourage such a hasty conclusion. Yet, after he had been criticized on this score by some of his best students, Brentano came to deny that he ever conceived of the intentional object of whichever kind of mental act as some " immanent " entity to be distinguished from its (putative) real object - a distinction that Husserl himself famously and rightly rejected. But then, of course, he had to cope with the issue of " empty " terms and cognitive states.

Regarding this problem (*i.e.* that of would-be referring expressions and/or intentional mental items without an actual " object "), there would seem to be just two answers only, once rejected the intentional/real object spurious divide One is Twardowski's (and, for a part Meinong's) more distinction between " object " and " content ". Another is the otherwise Meinongian overgenerous attribution of *som*e ontological status to every purported " object of thought ", whether actually existing or not.

Actually, Brentano rejects both ways out. No wonder, then, that he afterwards kept wavering on this issue, hesitating, as it seems, between a neo-Thomist and a neo-Scotist approach to the very idea of a mental " representation " [2]. No wonder either that he eventually came to regard intentionality as just " quasi-relational " (*Relativische*)[3].

Ever since the Scholastics (at least) it has widely assumed that a *genuine* relation can only hold of relata that are really existing, themselves, and are really distinct from each other. Actually, this dictum might be disputed, in view not so much of the controversial " relation " of identity than of such relations as self-love or self-destruction (*e.g*,.suicide). However, for the present purpose, I shall leave this complication aside. Suppose, thus, that the dictum is taken to hold generally. What, then, of the intentional " relation " when the purported " object " of such or such mental act does not actually exist (and provided we don't turn toward some kind of neo-meinongian solution) ? Should we regard this, along with Reinhardt Grossmann[4], as showing that the Scholastic criterion does not apply to *intentional* relations ? Or, following Keith Campbell[5] should we interpret this, rather, as a clear indication that

[2] on this issue, see Dermot Moran, " Brentano Thesis ", *Supplementary Volume of the Aristotelian Society*, 1996, pp. 1-27

[3] F.Brentano, *op. cit., p. 272*

[4] R. Grossmann,, *The Categorical Structure of the World*, Indiana University Press, 1983, pp. 1977 ff,

[5] K. Campbell, *Abstract Particulars*, Blackwell, 1990, p. 178

" intentional relations " are not, as a matter of fact, genuine relations ? At the fist blush, it would seem that a famous remark by the " last " Russell might reinforce the latter view :

> " The doctrine of internal relations held that every relation between two terms expresses, primarily, intrinsic properties of the two terms and, in ultimate analysis, a property of the whole which the two compose. With some relations, this view is plausible. Take, for example, love or hate. If *A* loves *B*, this relation exemplifies itself and may be said to consist in certain states of mind of *A*. Even an atheist must admit that a man can love God. It follows that love of God is a state of the man who feels it, and not properly a relational fact "[6]

At first sight, Russell seems to concede that *some* relations - *v.i.z.* some psychological relations, such as love or hate - can be indeed analysed away, or might be reducible to monadic states or properties. However, as aptly remarked by Vincent Descombes[7], this admission is, in fact, merely apparent, as Russell hurries to stress that " relations " of this kind are not genuine *relations* in the end.

Well, maybe so. However, this passage raises at leat two different, though complementary, questions. Firstly, is Russell right to assert that even an atheist should allow that a believer can love God ? Secondly, does this particular kind of case suffice to licence the conclusion that love and hate in general are not genuinely *relational* states ? Let us begin by considering this second issue. It will be easily granted, I guess, that John love's for Mary requires the existence in John of some internal states (be they conscious or not), or, to simplify, the exemplification of a number of monadic (in most cases, actually, dispositional) properties. Yet – unless we find ourselves in the extreme and most unusual circonstance where John has got mad about a wholly imaginary Mary, or has fallen in " love " with a really existing Mary he has been told about but actually never met –, I cannot conceive of any reason why we should *a priori* decree that John's love towards Mary consists (or, at any rate, consists *exclusively*) in the co-occurrence of such more or less " intrinsic " properties in John. Thus, I cannot think of any good reason either why we should reject the idea that it is Mary *herself* who, in the most favourable and (hopefully) most frequent case, is the object of John's love for Mary, nor of any serious ground,

[6] B. Russell, *My Philosophical Development*, Unwin & Allen, 1959, p. 42
[7] V. Descombes, *Les institutions du sens*, Minuit, 1996, p. 191

therefore, to deny that the latter is (or, at the very least involves) a genuine relation. But, then, how are we to account, all the same, for the the former case – that is, for the scenario by the lines of which Mary, for instance, does not really exist (and never, in fact, actually existed), so that John's " love " would seem to involve, at best, what Medieval philosophers used to call a " relation of reason " and what Brentano, in turn, dubbed a " quasi-relation " ? Once one has renounced, as I think we should do, the temptation to think that in both cases - *i.e.* whether Mary, say, exists or not -, John is (at least immediately) in a loving-relation with some representation of Mary (or with some unlikely " immanent " Mary, construed as a mere " intentional object "), we seem to be left with just two options. One of them is to allow that the verb " to love " is, as it stands, open to two different interpretations : *loving* could be read, as it were, either *de re* or *de dicto*. Now, it is a well-known fact that there exists, in common parlance, some sense in which John may well be said to be in " love " with some merely imaginary Mary – just as there is a sense in which Lady McBeth can be said to " see " her hands covered with blood, or the average serial killer to " hear " extra-terrestrial incitements to further slaughters. However, one might as well decide that these are just non-literal, and more or less parasitic, uses of both kind of verbs. If so, John cannot any more literally " love " Mary than he can, literally again, " see " a pink elephant in front of him or " hear " any voices in his head. And, thus, an atheist should obviously *not* allow that a man can really love God : as Mark Sainsbury rightly observed[8], if there is no God, even a sincere monotheist can at the very best believe or imagine that he loves God. The latter type of solution amounts to what one might call, *cum grano salis,* a " disjunctive " account of the loving experience - by analogy, of course, with the so-called " disjunctive " theory of perceptual experience. After all, the very principle of a " disjunctive " theory in the philosophy of perception (J.M. Hinton, P. Snowdon, J. McDowell) has already been extended to the epistemology of other kinds of cognitive states – and, primarily, to the interpretation of *knowledge* (McDowell himself and, above all, T. Williamson). And, clearly, *seeing that* is no less a " factive " mental state than *kowing that*, which just means that, if *S* knows/sees that *p*, then (*ex hypothesis*), *p* is the case. Not all perceptual experiences, however, are of the " epistemic " or " doxastic " kind – far from it -, so that it is of the highest importance that we should make, at the vey least, a principled distinction between seing *O* (or even, for that matter, seing *O as* (an) F in a non-conceptual manner) and seeing *that O* is *F*. Yet, it remains that, even though it does not

[8] R.M. Sainsbury, *Russell*, Routledge, 1979, p. 230

always imply the existence of the state of affairs that *p* – that, say, *O* is *F* -, the statement " *S* sees *O* " as (an) *F* ", or just " *S* sees *O* ", does in any case entail, at least, the existence of *O* itself. In short, whether or not " factive " *stricto sensu*, perceptual states are clearly " object-dependent "

However that may be, I shall contend that *seeing* and *loving* are, for that matter, on the same boat. Once again, whether John can be said to genuinely *love* this ideal, and therefore non-existent, woman whom he secretly calls " Mary ", might look as a merely verbal issue. To speak the truth, I am inclined to take this would be an easy way-out. But never minds : what really matters is that in both cases (that is, wether we embrace some kind of " disjunctive theory ", or, rather, make ourselves content with the more traditional *de re*/*de dicto* distinction), we shall have to renounce what I have called above the *principle of uniformity*.

Clearly enough, what holds both of those such " factive " states as propositional knowledge or epistemic seeing, and of such typically object-dependent states as (genuine) love or " simple seing ", is not generally true of many other cognitive states as ordinary beliefs or the mere fact that you are currently thinking o*f X*. There is no need to say that the sheer *belief* that *O* is *F* does not, *per se*, no more implies that *O* is *F* than it entails, at the very least, the actual existence of *O* itself. Still, it remains that, for a long and rich intellectual tradition which bloomed towards the former century's latest decades (although its origin goes back to Russell's seminal intuitions about this issue), an important distinction has to be drawn, amongst beliefs and more or less similar propositional attitudes, between those wich essentialy, or constitutively, depend on the identity and very existence of their purported object (genuinely singular thoughts) and those which don't (descriptive thoughts). What I have in mind, of course, is the by now well-established philosophical tradition (Kripke, Donnellan, Putnam, Kaplan, Mc Dowell & *alii*) which kept putting an emphasis – and, as I see it, quite rightly so – on the " object-dependence " of both genuinely singular phrases and thoughts. This is plainly not the circumstance for revisiting the formerly widely discussed issue wether such linguistic or mental items should be regarded as " directly " referential or whether their semantic rôle (as I keep, in fact, inclined to believe as for me) could not be more accurately accounted for within a more or less strict Fregean framework. In my view, it is quite possible to sucribe to the overall principle of the so-called " causal theory of reference " - or, at least, to aknowledge that various categories of terms, as well as mental representations or acts, crucially depend upon the existence of a causal relation with it source on the side of the relevant " referent " or " intententional " object.-, and yet not

succumb to what Gareth Evans famously and ironically dubbed " the photographic model ", with its dubious implication that this causal relation, just by itself, might suffice to determine the meaning of the linguistic or mental items in question. Even so, another vivid issue in the eighties was between those, among the so-called " neo-Fregean ", who claimed to remain wholly faithful, in their own way, to Frege's principle that the sense of any term or phrase strictly determines its reference (G. Evans, J.McDowell) and those who, themselves faced with Putnam's Twin-Earth thought-experiment and various similar externalist arguments and challenges, favoured what was called, in those days, the " twofold " theory or mental content (*e.g.* K. Bach, C. McGinn, A. Woodfield and so forth), pleading in favour of some minimal mutual independence between " sense " (most commonly construed in terms of cognitive " or functional terms) and " reference " (presumably accountable upon a mere causal, informational and, at any rate, " external " basis). Although my own inclination was, in those days, and currently remains in favour of the former option, I have no intention whatever, in this note, to revisit these most complicated issues. But could, please, my reader keep them in mind, all the same, when giving a look to the the two next sections ?

2. What is a " real " relation ?

For a while, suppose, in any case, that intentional " mental acts " - or at least *some* of them, including genuine love -involve a " real " relation towards their object. But what, then, is a real a genuine or " real " relation ? According to the Aristotelian and Scholastic tradition, which thought of relations in terms of relational properties *(relative accidents*), a dyadic " real " relation, as opposed to a mere " relation of reason ", is such as (i) it holds of two really existing terms, (ii) its terms are, themselves, really distinct ; (iii) the relation has a (monadic) foundation wihin boh its relata. Clause (iii) is particularly important, as it means that, for the Scholastics, a real (categorical) relation is above all a *grounded* relation. Let's name this third requirement the *foundation criterion*. A major issue among Scholastic discussions concerning these subjects, however, has famously to do with the question whether " grounded " relations should be allowed some kind of first-class and distinctive being, over and above that of their monadic foundations. This is surely not the place to revisit the many intricacies of the sophisticated and most fascinating debate which took place on this score, more particularly, some time between the late 13 th and the early 14 th centuries (and was *mutatis mutandis* revived re-

cently, about comparative and other supposedly " supervenient " relations). Suffice it to mention another ontological requirement which, however variously interpreted, seems to have then played a major rôle, as well, in this context : that according to which a real relation is one such that its occurrence makes a *genuine difference* to its relata (let's call this the *genuine change criterion*).

So far as good old comparative " mutual " *(i.e.* multilateral or just, say, bilateral) relations are concerned, how to reconcile the foundation criterion with the genuine change requirement was, and remains, a most tricky issue, especially in view of Aristotle's famous " indication " (as Peirce put it) that there is " no change " (*i.e.* no *real* change) within the category of relation, since " it may happen that when one correlative changes, the other can truly be said not to change at all, so that in these case the motion is accidental " (Aristotle, *Physics*, V, c. 2 225 b, 11-13). To provide just an example, if both *A* and *B* are white, it would seem that, according to the " foundation " criterion, they are really similar, since the relation they have to each other – that is, in fact, *A*'s property of being similar (in colour) to *B* and *B*'s property of being similar to *A* – is grounded on both terms. But, on the other hand, does not Aristotle's observation show that *A* might become similar to *B* merely in virtue of the fact that *B* has just been painted white, or *vice versa* ? Here, clearly, we seem to record some tension between the foundation and the genuine change criteria. How could *A*, for instance, become " really " related to *B* without any change among its intrinsic properties ? One remembers the answer put forward by Scotus : necessarily, if the relation is real, *A* undergoes a genuine change, but a real change involves the acquisition of some real proprety and, as there is no actual change in *A*'s absolute acidents, the new acquired property has to be distinct from anyone of them. Thus reformulated, Scotus' argument clearly invites Ockham's reply : how could I, just by repainting a wall in Rome, really change the (colour of) a wall in Oxford or in London ?

Be that as it may, things look quite different when we turn to unilateral, or " non-mutual ", relations – that is, to a two-term relation with a foundation in just one of is relata, like God's relation to His creatures or like intentional relations within Aristotle's (sub) category of the measure and the measured. This time, on the contrary, the foundation criterion and the genuine change requirement would happen to converge. Take Aristotle's own example of the knower and the known : if *A* knows *B*, *A*'s relation to *B* would seem to be have its foundation in *A* alone, so that *A* cannot acquire or loose it without undergoing some intrinsic change - while *B*'s property of being known by *A* does not make or imply any such change in *B*. Hence Aquinas's well-known view, endorsed by many Scholastic philosophers, that *A*'s relation to (relational prop-

erty directed at) *B* is, indeed, a " real relation ", whereas *B*'s relation to *A* is but a " relation of reason ".

Now, most obviously, I don't wish to suggest that we should return to the Medieval view of relations. Philosophers, nowadays, do not conceive of relations, generally speaking, in tems of " relative accidents ". Far from reducing relations to relational properties, most them would rather regard the former as being (at best) both logically and ontologically prior to the latter – and I fully adhere to this post-Russellian wiew. Nevertheless, I would like to hint to what I take to be the main insight behind the Medieval account as far as intentional relations are concerned.

3. Inherent or extrinsinc direction ?

However, before I endeavour to do so, let me first point at just two of the many difficulties met by the mainstream Scholastic tradition on this score. One of them is that, however construed (and insofar as the " foundation " requirement can be interpreted in the light of the more recent notion of " supervenience "), it is quite doubtful, to say the very least, that intentional relations " supervene ", *stricto sensu*, upon their unique subject-sided foundation. Clearly enough, that John *really* loves Mary seems to imply some form of acquaintance, or causal relationship, with dear Mary herself, As an aside, this comes rather as a piece of good news, considering the popular, albeit much controversial view, opinion that supervenient entities have no reality of their own over and above that of the underlying substances or properties (Armstrong's famous " free ontological lunch ").

The second, and presumably the main, difficulty has to do with what Medieval philosophers used to call the *esse-ad* (as opposed to the *in-esse*) of relations and directly flows from the very fact that we regard relations, today, either as genuinely *polyadic* properties or as some kind of connective entities standing somehow " between " their relata. Although modern logic has it that every relation has a converse and that, exception being made for symmetric relations, a relation and its converse are, from a purely formal point of view, distinct from each other, many philosophers within the analytic tradition are inclined to think that every relation – wether symmetric or non-symmetric - is actually, metaphysically speaking, identical with its converse. In other words, since it is the case that Paris is north of Maseilles, it is ipso facto the case that Marseilles is south of Paris, and clearly this amounts to the very same

state of affairs. This is on this ground that Kit Fine[9] (following Russell 1913)[10] recently objected to what he regards as the " standard " view of relations, as previously and famously put forth by Russell himself (1903[11]), according to which non-symmetric, or at any rate asymmetric, relations involve some form of intrinsic " sense ", or direction, and relate their terms in a given order. Fine's argument is based on the consideration of " an important class of metaphysical and linguistic contexts which call for an alternative conception of relation ", in that they seem to involve the existence of relations " for which there is no meaningful notion of converse". Since I have recently examined and discussed at full-length Fine's own account of " neutral relations " thus understood, I shall not repeat, here, the detail of my objections[12]. To put it in a nutshell, and putting aside some further more or less technical difficulties, I seems to me that what Fine's has in mind under the name of the " standard view " of (non-symmetric) relations is actually the conjunction of two distinct theses which - just like Russell himself did - he takes to be so closely associated that they may well be regarded as forming just a single philosophical conception of relations in the end. A first thesis is that every non-symmetric two-term relation has a " sense " and is, to that extent inherently directional. A second thesis is the claim that every non-symmetric two-term relation has a converse which is, not just logically or conceptually, but also and above all ontologically speaking, distinct of itself. Now, as I see it, no only are those two claims quite distinct. I shall furthermore contend that they are independent from each other – or, at at rate, that the first thesis does not imply the second one, so that the falsity of the latter does not entail that of the former. As Erwin Tegtmeier and Ingvar Johansson have remarked, a important step towards dispelling any risk of conflation, here, is to realize how ambiguous the very notion of " sense " (of a relation) itself turns to be[13].

Indeed, it is one thing for a relation to hold of its terms in some specific order, and it is quite another to enjoy some kind of inherent directionality.

[9] K. Fine, " Neutral Relations " *The Philosophical Review*, vol. 109, n° 1, 2000

[10] B. Russell, *Theory of Knowledge* (1913), in *Collected Papers*,vol. VII, George Allen & Unwin, 1983, part. II ; chap.1

[11] B. Russell, The Principle of Mathematics, Cambridge University Press, 1903, pp. 140-141

[12] F. Clementz, " Asymétrie, ordre et direction : la notion de " sens " d'une relation ", in A. Gay (éd.), *Autour des* ***Principia Mathematica*** *de Russell et Whitehead*, Editions Universitaires de Dijon, 2012

[13] E. Tegtmeier, " The Ontological Problem of Order ", in K. Mulligan & H. Hochberg (eds), *Relations and Predicates*, Ontos Verlag, 2004 ; I. Johansson, " Order, Direction, Logical Priority and Ontological Categories " in J, Cumpa & E. Tegtmeier (eds.), Ontological Categories, Ontos Verlag, 2011

As it happens, some relations – like, say, *(temporally) precedes* – do enjoy both properties. But many ordering relations, as *being greater than* for instance, do not display any kind of inherent direction. Most obviously, if $a > b$ and $b > c$, then $a > c$ - something we might as well express as " if $c < b$ and $b < a$, then $c < a$ " - no matters which way we are to read this ordered series of dyadic relations : clearly, there is no objective and inherent direction going from to a to b, for instance, rather than the other way round. (My own view, actually, is that such relations nevertheless involve some kind of " dissymmetry " (no to be confused with asymmetry), or functionnal non-interchangeability – some kind of *proto-order* - between their terms. But this is another story[14]) On the other hand, and to take another example among those put forth by Russell in *Theory of Knowledge*, consider the *loves* relation. If A loves B, nothing forbids, but unfortunately nothing ensures either, that B loves A. Such a relation is non-symmetric (which means, according to me, that it also implies some form of intrinsic dissymetry), but, thanks God, not asymmetric (which means that, not being transitive either, it can in no way serve as foundation for any kind of relation of order). But, by contrast with the *greater than* relation, it would seem to harbour some kind of essential " sense " or direction. Like it or not, if A loves B (and, therefore, if B is loved by A), it is not just accidental, or due simply to some obscure linguistic convention, that we usually describe this (unique) state of affairs using the former formulation rather than the latter alternative.

Intentional relations, such as loving, are clearly directional. But are they *intrinsically* so ? As for me, I am inclined to give an affirmative answer, and this is where I would now depart from, say, Ingvar Johansson's full view as expressed lately[15]. As a matter of fact, Johansson's main contention is that there are actually *three* sub-categories of relations with a " sense " : *order*, *priority* and *direction*, and that, in each case, sense actually comes " from the outside ". For sake of brevity, I shall only consider *direction*.. According to Johansson, in the (unique) state of affairs expressed both by " A loves B " and by " B is loved by A ", there is no inherent direction going from A to B : what " smacks of sense, " in this case, has its source within A alone. Johansson's wiew, to begin with, is that we should distinguish between the actual relation of loving as such *(R-love*) and the corresponding intentional mental state of being in love with B (*I-love*). His main argument is that if A " loves " B, while B, unbeknownst to him, is in fact departed, and since there can be no real rela-

[14] see Clementz 2012

[15] I. Johansonn, *op.cit*., pp. 100-101

tions except among actually (spatio-temporally), existing items, *A* cannot be said to " *R*-love " *B*, although, in one sense, John presumably remains in the same intentionnal state (*I*-Love). From this argument, he infers, first, that the " direction " of this particular instance of *R*-love , far from being intrinsic to it, comes from the underlying psychological state (*I*-love), whose intentional character is a " un-reducible phenomenon ", and, second, that *R*-love is an *internal* relation, in that its exemplification by *A* and *B*, in our example, supervenes upon the existence of both *B* and *A*'s corresponding intentional state which Johansson takes to be " logically independent " from the existence of *B*. Indeed, according to Johansson -whose analysis presents some acknowledged similarities with Searle's two-components view of intentionality, with internal as well as external conditions of satisfaction -, " all intentional act and states have so to speak a from-to structure ", in that they are directed towards a topole (an intentional object), *which may or may not exist* " (p. 100 ; my emphasis). *Prima facie*, this looks as a " conjunctive " view of the *loves* relation, as opposed to the " disjunctive " account which I suggested above. Which view should we favour ?

Consider, again, Johansson main argument :

> " When *b*, unbeknownst to *a*, dies, the *I*-love remains, but the corresponding *R*-love disappears, since it requires the existence of both the relata *(. . .) It is as simple as that* " (p. 100 ; my emphasis).

Well, is it really " as simple as that " ? One might wonder whether such a view does not rely upon an excessively narrow construal of the Scholatic dictum that a real relation requires the actual existence of each relatum. However, I shall no dwell on this complicated issue, which I leave open to discussion. Indeed, my main worry concerns Johansson's account of " *I*-love " and, in particular, his claim that its instantiation by John, considered by itself, is " logically independent " from the existence of Mary. For, how should we describe, in the first place, John's intentional state (which, of course, must not be confused with the pseudo " relational property " of just being in love with Mary ? Presumably, *I*-love is a quite complex psychological state, which is in part comprised of a number of more or less general dispositions, such as John's well-known fascination toward Irish girls with both green eyes and a solid sense of humour, or his propension to associate lasting relationship with mutual intellectual esteem. But there is also every reason to believe that – except, of course, in the rather pathological (and unusual) case where John gets suddendly enamoured of a wholly imaginary Mary -, John's *I*-love psychological state also (and mainly) consists of the semi-actualization of these

" primay " dispositional states in the form of more specific " secondary " dispositions, more directly *en rapport* with Mary herself and, as it were, " Mary-dependent ". If so, the " internal " foundation of the intentional relation under consideration consists of a complex of various psychological states (desires, feelings, emotions, beliefs, etc.) whose genuinely intentional nature in turn depends, if not on this very relation itself, at least of a whole serie of subvenient intentional relations such as their own directionality cannot be, on pain of infinite regression, explained away in terms of so-called merely " internal " states. The only alternative, in my view, would be to explicitly renounce the view that love is, generally speaking, a *de re* intentional state and to allow that its " intentional object ", whether its " real " counterpart actually exists or not, enjoys some form of immanent (in)existence in its own right. However, I presume this is *not* the sort of conclusion that Johansson would welcome (or, at any rate, that we should endorse).

Arguably, the same is true, *mutatis mutandis*, of quite a number of psychological attitudes as well. My own view, indeed, is that most intentional relations, or many of them anyway, are both " real " and *inherently* directional. John's *R*-love for Mary, in particular, is (in part[16]) an " internal " relation indeed, although not because it has a "foundation " in John's intentional state (*I*-love), but, rather, because it is partly *constitutive* of this intentional state itself. I told you : all you need is love.

[16] but in part only, since it is " external " to Mary.

36

Fregean Inferences *

OLAV GJELSVIK

The main aim of this paper is to argue that a Fregean conception of inference is fruitful and well equipped to help us sort out some intriguing philosophical questions. This is interesting because a Fregean conception of inference is quite different from today's standard conception of inference. It is not much attended to, but lends itself to virtue epistemological considerations, a central concern of Pascal.

I focus on two problem areas: The grand aim of making progress in our understanding, or perhaps explanation, of how we can extend our knowledge by inferring, and the equally important aim of understanding practical inference. Progress on either front would in itself be a very significant result. I shall argue that these challenges are related, and that the Fregean approach helps us see that. Seeing them as related is, furthermore, something of a novelty in today's discussion. A subsidiary aim of this paper is to make some progress on how to think about mental acts like judging and inferring, not least with regard to issues which arise when we think of judging and inferring as mental acts, and also think of doing something intentionally as a special way of being related to a propositional content.

The paper will mainly limit itself to the act of inferring deductively. This limitation will not do any harm to my purposes. This kind of act (the act of inferring deductively) might be seen as a challenge for an approach to doing something intentionally that conceives of the latter as the conclusion of

*I am very grateful to an audience at CSMN in Oslo for comments, and especially Dagfinn Føllesdal, Øystein Linnebo and Jon Litland. I am equally grateful for a comments from an audience in Mülheim, Germany, and especially Miguel Hoeltje, Jennifer Hornsby and David Velleman.

an inference (a practical inference). I shall explain what this challenge is and respond to it, and in so doing try to show how fruitful such an approach to doing something intentionally is: It can contribute to our understanding of inferring, the distinction between theoretical and practical inference, and perhaps also, more indirectly, to important questions in our conception of the epistemology of logic. When it comes to the latter, however, it would be necessary to extend my focus beyond what I can deal with properly in this article. I shall only touch on issues around validity and logical consequence.

1. Setting: Inference and inferring.

Here is an old question recently sharply posed by Dag Prawitz (Prawitz 2013): Why do some inferences confer evidence on their conclusions when applied to premises for which one already has evidence? What is it that gives inferences this epistemic power? This is a fundamental problem, Prawitz claims, and in the literature it has received no obvious solution. One basic problem is that appealing to the relation of logical consequence holding between the premises and the conclusion seems unable to do the job: this relation of logical consequence can hold between premises and conclusion without the subject making the inference displaying the right or even any epistemic sensitivity towards this very fact. Frequently, the answering of this question has been the motivating force behind the development of various types of epistemic account of the meaning of the logical constants (as developed both by Dummett and Prawitz respectively in somewhat different ways), and it has also inspired what we might call inferentialist approaches to meaning (Paul Boghossian (2003) might serve as an example). These positions achieve their aims by building the relevant epistemic sensitivities into the meaning of the logical constants. But not without criticism (see, for instance, T. Williamson 2003 and 2007) and there are, generally speaking, reasons not to accept outright epistemic or inferentialist accounts of both meaning and truth. That being the case, the problem of epistemic transfer in inference presents an interesting challenge.

Frege saw logic as the study of inference, and thought of inference as an act. In contrast, today's standard view sees logic as the study of logical consequence understood as the study of the relationship between propositional contents. Here is how Frege saw inference according to Dag Prawitz:

> An inference in the course of an argument or proof is not an assertion or judgment to the effect that a certain conclusion *B* "fol-

> lows" from a number of premisses $A1, A2, \ldots, A_n$ but is first of all a transition from some assertions (or judgments) to another one. In other words, it contains the $n + 1$ assertions $A1, A2, \ldots, A_n$, and B, and in addition, the claim that the assertion B is supported by the assertions $A1, A2, \ldots, A_n$, a claim commonly indicated by words like "then", "hence", or "therefore". ...
>
> This is how Frege saw an inference, as a transition between assertions or judgments. To make an assertion is to use a declarative sentence A with assertive force, which we may indicate by writing $\vdash A$, using the Fregean assertion sign. We may also say with Frege that a sentence A expresses a thought or proposition p, while $\vdash A$, the assertion of A, is an act in which p is judged to be true. (Prawitz 2013)

What seems very clear is the focus on inference as an act. The premises, the judgings, are also acts, as is the conclusion. Nicholas Smith (Smith 2009) has stressed that for Frege, inference is really a relation between actions. Putting things like that emphasises very strongly the act aspect in Frege. It would, however, be a mistake to think of Frege's inferential transitions as processes in time occurring between entities in time with different temporal extensions and locations. (Ian Rumfitt (2011) defends the view that inferences are not transitions. I cannot here go into his interesting views).

There is also something in the passage by Prawitz quoted above that in my view needs clarification or, probably, reassessment, namely the point that the inference in addition to being a transition "*contains the claim that the assertion B is supported by the assertions* $A1, A2, \ldots, A_n$*, a claim commonly indicated by words like "then", "hence", or "therefore""* (my italics). This, I think, is a somewhat questionable statement. (Of course, Prawitz is here thinking of inferences of the sort that he calls reflective inferences. [See below for a discussion of this concept.] Still, there remain problems here.)

My view (and, I think, any developed Fregean view), is that an inference might be seen as an act exhibiting a commitment to a correctness claim, but should not be seen *as containing such the claim*. We should take the "because" to alert us to the commitment made, but not as indicating that a claim is made. This point is a delicate one, and I shall return to it below. It relates to, but does not coincide with, Wittgenstein's point that in order to follow a rule we need no rule for how to follow the rule, we only need the rule – and to grasp that following the rule is 'this'. It relates by the same token to points Boghossian makes when he speaks of 'blind reasoning'. The point concerns the relation-

ship between an act of following the rule and reflective knowledge or belief about how to follow the rule that we might have. In the end it concerns the heart of the present debate about knowing how (Stanley 2011).

This point aside, we see that an inference here is seen as a transition from judgment(s) to a judgment; the latter being the conclusion. Both premises and the conclusion are judgments, but the inference itself is precisely not a judgment (for instance to the effect that the conclusion follows from the premises), but the transition from the premises to the conclusion, a transition that in some sense aims to answer to correctness norms. Judgments are, on this picture, constituents of inferences; inferences are not judgments. (There might of course also be judgments about inferences, and about their correctness.)

Frege, or the Fregean, thus sees judgments (represented by the corresponding judgment stroke) as a primitive or special kind of mental act, and also sees the transition in inferences as a primitive or special kind mental act governed by the (normative) laws of thought. For both kinds of mental act there is a question about the relationship to acts characterized as doing something intentionally. That question in turn raises the issue of how we are to think about doing something intentionally, and how our conception of that relates to our conceptions of inference and judgment. I turn now to the subject of doing something intentionally.

2. Doing something intentionally.

According to Anscombe (1957), doing something intentionally is at the heart of intentional action, and makes up the starting point for understanding what it is to do something with an intention, or to intend something. It is indeed natural to think of the Anscombian approach to doing something intentionally as a way of being related to a propositional content. She definitely thinks that doing something intentionally can be thought of as a conclusion (of practical inference). This view is also Aristotle's, and Davidson was also willing to entertain it.[1]

[1] Here is Davidson: "In the case of intentional action, at least when the action is of brief duration, nothing seems to stand in the way of an Aristotelian identification of the action with a judgement of a certain kind — an all-out, unconditional judgement that the action is desirable (or has some other positive characteristic). The identification of the action with the conclusion of a piece of practical reasoning is not essential to the view I am endorsing, but the fact that it can be made explains why, in our original account of intentional action, what was needed to relate it to pure intending remained hidden. — In the case of pure intending, I now suggest that the intention simply is an all-out judgement" (Davidson 1980 p. 99).

Now, if our conception of inference is Fregean, a transition between propositional contents to which we relate in the required ways, then we can very simply extend this account of inference into the practical realm as well: if an intentional action is a conclusion of an inference, and also a way of relating to a propositional content, then practical inferences would be inferences with such conclusions, i.e. inferences with conclusions that exhibit this practical way of being related to a propositional content. Frege himself limited his account of inference to inferences in demonstrative science, i.e. theoretical inferences, but it is a very small step, and a step already taken by Aristotle, to think that there are indeed two basic ways of relating to propositional contents in general and thus to the propositional content of a conclusion, namely a practical way and a theoretical way.[2] This practical way of relating to p is exhibited in doing p intentionally. I shall later give more substance to this view.

This move definitely raises the issue, which must be faced, of whether judging and inferring within the context of a Fregean approach to inference should be thought of *as things we do intentionally*. It is a challenging question, and I cannot settle all or even many aspects of it here. This is, however, what I think we should say: inferences are clearly intentional phenomena: they are personal level phenomena (and not sub-personal phenomena), they are clearly things we (i.e. inferrers) do for reasons, but they are not things we typically need to be aware of doing when we do them (as Anscombe argued was the case for all the things we do intentionally). Inferences are unlike intentional actions (things we do intentionally) in their relation to what we might be tempted to think of as the will, as they do not seem to be subject to the same sort of control we normally exercise over things we do intentionally. They are typically responses to many things we do intentionally, like reasoning, deliberating, considering, gathering evidence etc. But, and here is the catch, while we can decide what to do, we cannot decide what to believe of what to infer.[3]

[2] A very natural way of interpreting Anscombe's shopping example is that it shows these two different ways of relating to the same propositional content (the shopping list). The identification of this distinction with a distinction between belief and desire is a big mistake. It probably started with Searle's work on directions of fit.

[3] Of course there are people who argue that forming a belief or making an inference is just as much a utility-maximising choice as any other. I cannot deal with such views here. A prominent example is George Ainslie (Ainslie 1992), but the view is clearly wrong. There is, at this point, a new and interesting exchange between Pascal Engel (2013) and Ernest Sosa (2013) about whether forming a belief can be an action or not. Engel provides arguments against the view that forming a belief is a (possibly intentional) action, an appeals, among other things, to the different kinds of reasons that are relevant in the two cases. I see the crux of the matter addressed in the main text above, and as being about the range of what we can do intentionally. Answering this question

Judging and inferring are not extended in time (as intentional actions are). Many of the aspects pointed out above – like the lack of the standard type of control and also the lack of awareness of being engaged in doing the thing – relate to this point of no temporal extension. (Note that we are aware of deliberating, engaging in reasoning, checking proofs, etc., and that such activities have temporal extensions. We need not, however, be aware of making an inference exactly when we make it, even in cases in which we quickly become aware of having made it.) Judgings and inferrings are typically (i.e. generically) rational responses to the many things we do intentionally when we gather evidence and consider its merit, and deliberate and reason about what we ought to believe or ought to do. (This generic fact about judging and inferring should not, of course, stand in the way of recognizing that we sometimes respond irrationally to evidence, sometimes fall short of judging and inferring the way we should in the light of the evidence we have and the deliberations we have conducted.)

These points together seem to establish that judging and inferring are not necessarily or typically things we do intentionally, even if they are things we do, and they are personal-level intentional phenomena that display reason-sensitivity of as high a degree as any other thing we do. The necessary awareness criterion, argued by Anscombe for intentional action, i.e. our being non-observationally aware of doing what we do intentionally when we engage in doing it, is important for seeing things this way. If we needed to be aware of all the inferences we make in making them, that would be an enormous cognitive burden, and the same goes for judging that things are so and so. The point about control by the will, and that control in these cases of judging and inferring is very different from the normal type of control we have over what we do when doing something intentionally, is related and equally important. Both lack of awareness and lack of control relate to the lack of temporal extension.[4]

negatively does not prevent a notion of teleology to apply to belief and belief formation, at least not when we see belief's point in relation to all these other things we do intentionally. I probably see things differently from Sosa, and, maybe, more the way Engel sees them, but I also recognize the need to go much deeper into this matter. The point about reasons concerns what it is proper to look into, how to conduct an inquiry, and thus about the proper employment of our capacities in forming beliefs.

[4] It might be argued that the way to understand Hume's stance on skepticism, is to understand that we do not with our higher cognitive and reflective capacities control the making of all the judgments we actually make. When a friend knocks on the door and offers a game of backgammon, our reflective skeptical reasoning loses its hold on us. This could not be so if reflective control extended to judgment and inference. Nature has done us an enormous service

3. Virtue epistemology and inferring.

The intimate connections between things we do intentionally in terms of considering evidence, reasoning and deliberating, and things like judging and inferring, are sufficient to uphold the possibility of using the type of virtue-theoretic considerations Ernest Sosa (see Sosa 2012) has insisted on in connection with belief-formation (which in this essay is another way of speaking of judging). I find Sosa's approach very useful for thinking generally about epistemic normativity, and very fruitful when it comes to these basic epistemic acts. The standing as acts is very important.

Bluntly put, Sosa's idea is that we should identify three levels of knowledge in ordinary cases of successful belief formation. The basic level we can think of as animal knowledge. Successful belief formation of this sort is *apt*, Sosa says, and that means that it shows real competence in arriving at a true belief in judging. Sosa thinks of possessing this competence as typically exhibiting good reliability in arriving at true belief, but not as reducible to reliability. Here we could also, perhaps, think of competence as providing safety, as Williamson does, and safety as not reducible to reliability. In any case, the next level of belief formation is when belief formation is meta-apt. In that case, the agent or the judger takes into account that the first level competence is intact in the circumstances, that the conditions are appropriate for exercising it, and also assesses the likelihood that the epistemic action from the competence will succeed in the circumstances. Here we have apt reflective intentional activity about first-level belief formation. The third level is that of fully apt, which is when the action on the first-order level is apt *because* it is meta-apt. In that case the agent manifests his/her meta-competence in when and how to exercise their first-order competence when exercising their first-level competence, the competence that is aptly deployed in delivering true belief or judgment. This last third level then exhibits *knowing full well*.

How do these levels identified by Sosa come out when seen in the context of something like Prawitz's concept of a reflective inference? Here is Prawitz on the concept:

> Reflective inferences must be understood as aiming at getting support for the conclusion. This may be articulated in different ways. We may say that the primary aim is to get a good *reason* for the as-

in not letting us (in one sense of us, the reflectively informed will) control all this activity. The distinction in question also relates to Kahneman's distinction between system one and system two.

> sertion that occurs as conclusion. Since the term reason also stands for cause or motive, another and better way to express the same point is to say that the aim is to get adequate *grounds* for assertions or sufficient *evidence* for the truth of asserted sentences. Since assertions are evaluated among other things with respect to the grounds or evidence the speakers have for making them, we may also say that the aim of reflective inferences is to make assertions *justified* or *warranted*. (Prawitz 2013, page 6)

It follows that the point of reflective inference is to arrive at the conclusion with warrant, and that this warrant seems to be provided by the grounds for it, and those are presumably seen at a meta-level where one reflects on the correctness of the inference in question. In that case, the concept of reflective inference will tend to have the same extension as Sosa's knowing full well case, where the first-level aptness is seen as resulting from the meta-aptness. Note that on Sosa's view, knowledge on the first-order level, and the aptness found there, need not result from meta-aptness (i.e. the case of animal knowledge). Such (animal) knowledge should therefore not be accounted for by meta-aptness or reflective knowledge about the correctness of the inference. No such thing needs to be ascribed to a knowing inferrer. I shall employ Sosa's way of thinking about epistemic normativity, noting the connections to Prawitz's as I have just done.

It also seems quite obvious that the three-levelled structure identified by Sosa can be applied just as easily to inferring as to judging. (Sosa himself focuses on judging.) A conclusion is aptly reached when that reaching exhibits or manifests logical competence, a competence to correctly reach such conclusions in a relevant range of inferences. The inference is meta-apt when the inferring person exhibits meta-competence about when to infer and when not to infer, i.e. in the ability to stop inferring in the cases where the competence one has will not succeed. (Again there are connections to Kahneman.) Lastly, we have the cases where the inference is apt because it is meta-apt, where the meta-competence is actively employed and thus plays a real role in the production of an inference that is also apt. The meta-aptness consists both in active reflective knowledge about what good inferences are, accurate knowledge about one's own first-order competence, and active use of this knowledge in influencing the first-order activity.

It is important to distinguish an act such as asserting the correctness of the inference from the act of actually making the inference. It is very easy to think of an inference not as comprising a transition between judgments/assertions,

a transition subject to correctness or incorrectness in its execution, but as the judgment that the connection of logical consequence is in place between the premises and the conclusion. But if we think like that, we lose sight of something fundamentally important regarding the activity of inferring. We might also become subject to a version of the Lewis Carroll regress point; we would never get to the concluding, only to asserting the correctness of the inference, then to asserting the way we established that correctness, and the correctness claim regarding that way of establishing the establishing of the correctness etc. Making an inference thus cannot be making a claim to be asserted, even a correct claim about an implication, but must be to make the transition from premises to conclusion. Making a claim about the correctness of the inference is definitely not part of making the inference aptly, the animal knowledge inference. And there seems to be no good reason to let inferring *contain* the making of such a claim in the reflective case, even if one is prepared to make such a claim, and that knowledge is, so to speak, active. Also, for reflective inference and knowing full well, a sound commitment to correctness seems to be enough for the making of the inference.

Modern logic approaches inference in different ways, but not in this Fregean way where acts of judgment are parts, and inferring is an act. As said above, inference is typically approached as an abstract relation of logical consequence between a set of premises and a conclusion. Both premises and conclusion are then thought of as abstract propositional objects. We might think of the issue here as something like a producing/product ambiguity. One view (the standard view) focuses on the product rather than the producing, while the Fregean focuses on the latter. In any case, the view of inference as a mental act is then lost from the standard view, and the possibility of applying the Sosa apparatus is also lost. And, when this is so, epistemic transfer in inference becomes quite puzzling, something we really need to explain. One great advantage of the Fregean approach to inference, which sees inference as a mental act, is therefore to preserve the applicability of the Sosa virtue-epistemological apparatus, and the possibility of using that apparatus for explanatory purposes, as in the explanation of epistemic transfer. This is part of my motivation for staying fully Fregean.

4. Interlude: Some representations of Fregean inferences.

I shall introduce some very simple examples of the Fregean picture in order to communicate more clearly the way I am thinking. I shall use the judgment

stroke and explain the work it is doing in the inference, and index the stroke as to whether it is a normal theoretical judgment we are speaking about, or whether we are speaking about the practical way of being related to a propositional content. In the first case I index with a '*J*' for judgment, in the latter I use a '*P*' for practical. The first is a case of simple modus ponens.

1a. $\vdash_J$ (I am driving to Stockholm)
2a. $\vdash_J$ (If I am driving to Stockholm, I turn left here)
———————-
3a. $\vdash_J$ (I turn left here)

What is extra here, compared with the standard way of thinking about inference, is the presence of the indexed judgment stroke. This added level brings with it the possibility that each premise is apt, meta-apt or fully apt, and the same distinction can be applied to the inferring, it can be apt, meta-apt and also fully apt. The first level, that of being apt, carries with it straight knowledge in the case of judging, and the natural way of thinking about the inferring is that when an act of inferring is apt, then if the premises are held aptly, then the conclusion is also aptly held when it is the result of apt employment of some inferential capacity or competence. It is also natural to think that even if the employment of inferential capacity also is meta-apt or fully apt, the conclusion cannot be more than aptly held if the premises are only aptly held but not fully known. On the other hand, it is also natural to think that if the premises are fully apt, then the conclusion can be fully apt, but only as long as the employment of the inferential capacity also is fully apt.

We can therefore see the possibility of thinking about inferences on at least two levels, a ground level which is apt when the premises are aptly held and the inferential capacity is aptly employed, and the case where the meta-capacity is also active and one knows full well in Sosa's sense. We could also introduce the intermediate level if there were a point of doing so. I shall not do so for now.

This example above is deliberatively chosen because we can also use it in a case of practical reasoning. The only difference is in the way we relate to the propositional contents or thoughts, not in the thoughts or the way they relate semantically. Here is the example

1b. $\vdash_P$ (I am driving to Stockholm)
2b. $\vdash_J$ (If I am driving to Stockholm, I turn left here)
———————

3b. $\vdash_P$ (I turn left here)

Most of the things said about apt and meta-apt in the cases of the theoretical inference above carry over to this inference as well. The main difference is that we here have a practical way of relating to one premise, and also to the conclusion. I shall maintain that we need a practical way of relating to a premise in order to get a practical way into the conclusion, and I argue in some detail for this in Gjelsvik 2013 ('Understanding Enkratic Reasoning')[5]. The richness of the Fregean approach becomes even more striking when one considers a practical inference on the model of the theoretical.

Let me make some remarks about aptness in the practical case. First, the practical case as exhibited here is very close to Anscombe's late account of practical inference (in 'Practical Inference', first written for the von Wright volume of Library of Living Philosophers). In fact, or so I shall claim, the '$\vdash_P$' symbol stands exactly for what Anscombe already in "Intention" called *practical knowledge*, something she claimed philosophy had forgotten all about, i.e. a (legitimate and factive) way of being practically related to a content. When you are thus related to a proposition, you are engaged in doing intentionally the propositional content to which you relate. (This is phrased awkwardly, but that should not deter us.) In our example there are two such propositional contents, 'I am driving to Stockholm', and 'I turn left here'. In both cases — the practical premise and the practical conclusion — we take the whole premise to represent an intentional action. I say more about this practical stroke in other connections. I also agree with Anscombe that being so related to a propositional content, i.e. 'I am driving to Stockholm', implies awareness of me being engaged in driving to Stockholm. This awareness is a way of knowing that I am driving to Stockholm. Doing something intentionally carries non-observational knowledge of what you are engaged in doing with it. This knowledge is propositional.

There are further issues here concerning the point that an intentional action exemplifies knowledge how to do the thing in question. Such knowledge must be employed with success for the intentional action to be there. We get further layers when we consider whether the action is apt, whether it is also meta-apt, and apt because it is meta-apt. I shall not here take a stand on how to think about knowing how to x, and I want at this point to remain neutral on the contested and controversial questions about the relationship between knowledge how to something and knowledge that.[6]

[5] Jay Wallace made a good case for this in Wallace 2001.

[6] Jason Stanley (2011) is an important new contribution I shall not engage with here.

Intellectualists, philosophers who advocate the reduction of knowledge how to do something to knowledge that, operate with practical ways of being related to constituents of propositions or thoughts. That is, however, not the same as the practical way of being related to a whole proposition about which I am speaking. As I said, I will not be going into issues about knowledge how to do something, just stress the need to operate with a practical way of being related to whole propositions when thinking about doing something intentionally. This point shows some of the complexities in the relationship between Anscombe's use of practical knowledge and the discussion about Ryle's distinction between knowing how to do something and knowing that. (As discussed in Stanley 2011.)

Let me close with some further examples of a theoretical and a parallel practical inference, before going on to show how enkratic inference can be dealt with on the present approach. The first example is interesting because it shows how to extend the practical way of relating to a proposition to the case of intentions. There are no conditional actions, but there are conditional intentions. The central case, the case of doing something intentionally, thus needs to be extended to intentions and conditional intentions. A great deal of practical reasoning, as Michael Bratman has shown, concerns plans within plans, and relations between intentions. Without discussing all of that, I shall just provide the example with intentions.

4a. $\vdash_J$ (If I ought to take a break, then I shall take a break)

5a. $\vdash_J$ (I ought to take a break now)

———————————

6a. $\vdash_J$ (I shall take a break now)

The practical analogue to this must be reasoning between two intentions, which is shown by the way 'shall' enters the actual content. We still represent the reasoning in the same way as that of action:

4b. $\vdash_P$ (If I ought to take a break, then I shall take a break)

5b. $\vdash_J$ (I ought to take a break now)

———————————

6b. $\vdash_P$ (I shall take a break now)

The enkratic case is Broome's case in which you move by inference from the recognition that you ought to take a break to an intention to take a break (see

Broome 2013.) '*B*' stands for belief by this rendering of Broome's approach, and has a parallel, but not a full parallel in the judgement stroke on my approach. The '*I*' stands for intention on Broome's approach, and has a parallel but not a full parallel in the practical stroke on my approach. I use the letters in this way for comparative purposes; I don't think it causes any problems. (I use the strokes to indicate legitimate ways of being related to propositions.)

This is the practical inference according to Broome:

2. *B*(I ought to take a break)

_____

3. *I*(I shall take a break)

(To be precise: On Broome's view this reasoning is enthymematic: I also need to believe that it is up to me whether I take a break or not. The full and correct representation of the inference is something like this:

2. *B*(I ought to take a break)

2*. *B*(It is up to me whether or not I take a break.)

_____

3. *I*(I shall take a break))

On the view I am pursuing, this is not correct reasoning, and Broome is wrong. There is a logical step from the modal verb 'ought' to the modal verb 'shall' which is not correct inference – satisfying the one modal predicate does not entail satisfying the other.

This is the correct practical inference on my view:

1. $\vdash\!\!\!-_P$ (If I ought to take a break then I shall take a break)
2. $\vdash\!\!\!-_J$ (I ought to take a break)
3. $\vdash\!\!\!-_P$ (I shall take a break)

Without the first premise being true of you, you will not reach the conclusion. Note, it is not enough for you to judge the first premise to be true, you have to have adopted the practical way of relating to the first premise to be able to infer this conclusion. If you only judge the propositional content of first premise to be true, but do not relate to it practically, then you might exhibit akrasia or weak will in this case. This shows that weak will is not typically a failure of reasoning (as it would be on Broome's account), but also why

we need a practical way of relating to a premise in order to get a practical conclusion.

This demonstrates the ability of the present neo-Fregean approach to practical inference to handle some of the most contested issues in today's discussion of practical inference. And there is more here: dominant in the discussion has been the role of rationality requirements in practical inference, their form, whether they are wide in scope or narrow, and so forth.[7] I submit from the perspective of the present approach to inference that we have all the resources we need in the ways of relating to propositional content, and there is no point in going into the issue of rationality requirements. Or rather, all the work that can be done by rationality requirements will be done by the resources we already have at our disposition, by what goes into the legitimate ways of relating to propositional content, practical and theoretical.

Work is also done by the interaction between judging and inferring, and the recognition that inferential connections may force us to reconsider some of the judgings to which we are committed: if a conclusion of a valid inference must be rejected, we must reject at least one premise. That goes for both theoretical and practical inference. The great virtue of the present approach is the way we get a full parallel between practical and theoretical inference in this matter, and a full parallel in the way entailment relations matter. This was indeed Anscombe's Aristotelian aim.[8] If we were to extend the present approach to hypothetical thinking, then we might as an additional benefit be able to see the structure of reduction arguments as fully parallel and as arising out of some hypothetical premises leading to unacceptable conclusions. Of course, we engage in such reasoning all the time. Any full approach to inference needs to deal with that. I return to this kind of extension in the concluding overview, but just let me say that I want to remain neutral on *how* to extend to hypothetical judgments, and that from the present perspective we start from the categorical in both the theoretical and the practical case. The extension to the hypothetical may take quite different forms in the two cases.

This concludes the discussion of enkratic inference. Let me end by appending one further point. Keeping the practical case in view makes it easy to see how difficult it is to think of inferring as something we do intentionally.

[7] There is now a huge literature on this topic. Important early contributors are Broome and Kolodny.

[8] If we were to extend the present approach to hypothetical thinking since we might benefit further by being able to see the structure of reduction arguments as fully parallel and as arising out of some hypothetical premises leading to unacceptable conclusions. Of course, we engage in such reasoning all the time. Any full approach to inference needs to deal with that.

If we do, it will result in a vicious regress in the practical case: If we think of inferring as a way of being practically related to a propositional content, as we would be if the action was intentional, then that intentional action of inferring should also be able to be a conclusion of another practical inference, and so forth. We get an analogue to the Lewis Carroll problem. The problem is not solved as long as you think of the transition as something we do intentionally, and think of doing something intentionally along the present lines.

5. Returning to the theoretical case: Discussion of Prawitz's explanation.

Prawitz formulates the explanation we are seeking of the transfer of epistemic value as an explanation of how some inferences come to be legitimate, conceived in this case as inferences that confer evidence on a conclusion when there is evidence for the premises. Prawitz supplies this explanation at what he calls the level of generic inference. The notion of legitimate inference goes then like this:

> In sum, it is required of a proof that all its inferences are successful. To characterize a proof as a chain of inferences, as we usually do, we thus need this notion of successful inference. It is convenient to have a term for this, that is, for inferences that can be used legitimately in a proof, and I have called them legitimate inferences (Prawitz 2011). Accordingly, a generic inference is said to be legitimate, if a subject who makes the inference and has evidence for its premisses thereby gets evidence for the conclusion; or more precisely, it should follow that she has evidence for the conclusion from the assumptions that she performs the inference and has evidence for the premisses. We can now say that a deductive proof is a chain of legitimate inferences.

We see that the starting point here is that of successful inference in relation to producing a proof. On the Frege–Sosa approach, we might think of the notion of success such that a successful inference extends our knowledge. This is because we start from inferences where we judge/assert the premises, successful judgment is knowledge, and the rule of assertion can be seen as 'assert only what you know'. Proof in the case of an inference will then typically play a role in knowing full well that knowledge has been extended in making the inference.

There is a big difference between this way of simply speaking about knowledge and the extension of knowledge, and speaking of the conferring of evidence on the conclusion. These differences may have some of their background in whether one's thought is structured by the tripartite analysis of knowledge, or whether knowledge is not to be analysed into three parts. It also has a background in the extension of the role of asserting in Prawitz's approach. Prawitz extends asserting from judging something to be so, or assert correctly, to simply assuming a propositional content to be correct. In Prawitz's revised terminology, an assumed premise is also an assertion. With that move, he departs in a quite specific way from the basic Fregean picture of inference. That departure has its clear motivation, and that motivation needs to be addressed; for now, however, let it just be duly noted. The result is that while the old fashioned Fregean considers inferences that basically extend our empirical or mathematical knowledge, and I in this paper extend that old Fregean picture of inference to commonsense knowledge and also to practical knowledge (in Anscombe's sense), Prawitz generalizes in a different direction by thinking of assumptions as assertions, moving swiftly in fact from categorical judgments to hypothetical ones. Inference can then take place between assumptions, not only between legitimate judgments, and this may well suit logic. I am not extending in this way to assumptions and to hypothetical judgments. I hold that we should deal with those extensions as a special or derivative case to be properly considered in due course in both the practical and theoretical case: our basic business is inferential competence.

With this noted, here is some of what Prawitz says to explain the transfer of evidence in legitimate inferences:

> If we pay attention to the agent who performs an inference and the occasion at which it is performed, we are considering an *individual* inference act. By abstracting from the agent and the occasion, as we usually do in logic, we get a *generic* inference act.

He formulates the required explanation at the level of generic inference thus identified. *How* we abstract will, however, matter as well. The issue can be formulated as being about what we do when we abstract 'from the agent and the occasion'. Should we abstract in such a way that we can still keep all the Sosa distinctions properly in place or not? To a virtue epistemologist it is necessary to keep the various competences of the inferring agents properly in view when we abstract. I therefore answer my own question affirmatively, but cannot see that Prawitz abstracts in such way as would be required by such conditions

on the abstraction. That is, it seems to me, related to his abstracting not from inferences where we go from actual legitimate judgments to a conclusion, but from inferences where we start from assumptions (assertions/judgments) in the extended sense. I suggest we should abstain from moving to assumptions (assertions in the extended sense), and only generalize on well-functioning epistemic agents performing old-fashioned Fregean inferential acts successfully, inferential acts on actual judgments or actual premises.

To show that this might matter, let me introduce Prawitz's explanation of how an inference comes to be legitimate:

> We can spell out what it is show that a condition *C* on generic inferences is sufficient for legitimacy as the task of establishing for any generic inference *I* and subject *S* that from the three facts
> (1) the inference *I* satisfies the condition *C*,
> (2) the subject *S* has evidence for the premisses of *I*,
> (3) *S* performs *I*,
> it can be derived that (4) *S* gets evidence for the conclusion of *I*.
>
> To find such a derivation is the business of the philosopher who seeks an explanation. This may be described as taking place on a meta-level, where the subject's activities on the object-level is explained. The subject is not to do anything except performing the inference *I;* the point is that thereby, without doing anything else, she gets evidence for the conclusion.

Here we have a subject S who performs an inference, and not much more is said about the relevant dimensions in epistemic normativity that are exemplified by the subject S in performing that inferential act. This is where we seem to go different ways. From my perspective the great advantage of the Fregean way of thinking about inference is precisely the way we can theorize about competences and abilities of the person doing the inferring when approaching issues like that of transfer of knowledge. This applies both to the competences of the subject in question in making each premise legitimate, and the competences of the subject in making the legitimate inferential transition. I contend, therefore, that all abstractions we make when explaining legitimate inferences should respect and acknowledge the normative epistemic distinctions Sosa identifies. We should bring in the competences of the inferring subject, the aptness and meta-aptness displayed by the inferring person when doing the inferential transition, and use such resources in explaining the transfer of epistemic properties in inference. And note: the successful employment of

these competences seems to match the extension of the *explanandum*, i.e. when knowledge *is* extended in inferring, and the problem that the explanation provided explains epistemic transfer when there is no transfer disappears. There is, possibly, therefore, disagreement about how to conceive of the shape of explanation of what I think of as getting knowledge extended by inference.

There are also, of course, different ways of thinking about the explanation provided. Timothy Williamson (Williamson 2009) has recently attempted to explain the extension of knowledge by logical competence when we move from premises to conclusion. That explanandum is close to the explanandum I focus on, but not the same. This is because I want only to explain that we extend knowledge when employing logical competence in successful/legitimate inferences. Williamson wants to explain that such inferences *are* successful, even if the probability of the truth of the conclusion is lower than the truth of each premise separately. My explanandum presupposes my Fregean framework, and starts from legitimate premises, as identified above, and aims to explain *how* epistemic standing of the premises is being transferred to the conclusion *when* it is transferred (and I say by logical competence). I am not explaining *that it is* transferred. Williamson has a stronger aim than me: he wants to explain *that* epistemic properties *are* transferred by logical competence. There is, of course, an issue of transfer of legitimate ways of relating to a propositional content or a thought that I am not discussing, and an issue of the relationship between taking oneself to know and actually knowing. Some of these issues have been discussed under the heading of rationality requirements. These issues need addressing, but not here, and there will be different challenges in the practical and the theoretical ways of relating to contents.[9]

It also seems to me that what goes into Prawitz's condition C above will vary a lot on whether we are explaining transfer of knowing full well, or just transfer of (animal) knowledge. For instance, knowing that the conclusion follows from the premises by knowing how to see the last step of the inference as accounted for by introduction rules and reduction to canonical form, might

[9] My assumption is that all premises are legitimate. Of course in real life the probability that all premises in a deduction are legitimate will be lower than the probability that a particular premise is legitimate, and lower than the probability that the premise with the lowest probability of being legitimate is legitimate. Williamson's project is to show that safety does not work the way probability does, that the conclusion might still be safe, and that safety is what matters for knowledge. That is an extremely interesting project. But I am simply assuming that the premises are legitimate, and that we know each premise. I believe this example shows advantages of the Fregean approach in how to conceptualize the issues, but cannot go further into that.

plausibly be a way of knowing the conclusion full well, but it does not seem required for explaining anything on the part of the subject S if the transfer of knowledge when inferring is not meta-apt. In that case, the ascription of simple logical competence is enough to explain transfer of knowledge.

There is a further point concerning the notion of evidence, and the extent to which one's thinking is coloured or structured by the way one thinks of the relationship between knowledge and evidence. If one accepts a traditional tripartite account of knowledge, and sees knowledge as justified (or evidence-based) true belief, then things looks quite different from when one does not accept that analysis, and thinks of knowledge as too central a concept to be subjected to analysis, and, moreover, thinks that our evidence is simply what we know. On this way of thinking, the Fregean conception of inference is a natural ally, and we can think of successful inference simply as extending our knowledge. We will no longer, it seems to me, have to pose the question of the transfer of evidence in quite the same way as Prawitz.

Let me add one further consideration here. Prawitz goes on to provide an extremely elegant proof. He shows how we can introduce a language of grounds such that he can prove that a subject who performs a valid inference also then obtains evidence for the conclusion, in an externalist sense of evidence (in the sense of being related to a truth-maker). Here is Prawitz again:

> The proposed explication according to which a subject who makes a valid inference gets to know a truth-maker of the sentence asserted by the conclusion, *although she may not know that it is a truth-maker of the sentence*, seems therefore optimal with respect to what a subject can become aware of and get to know by just performing an inference. (my italics)

The difficulty — or rather the limitation — in this as I see it, is that we do not from this account seem able to explain how the correct inference makes the right room for the subject's making the judgment that the conclusion represents. Without any awareness of the fact that the truth-maker is a truth-maker for the conclusion, the subject might have evidence for the conclusion without having any awareness of having evidence for the conclusion. If the explanation were a somewhat simpler one (simpler on the present conception), namely that of explaining that the conclusion is known, then the explanation would only need to appeal to competence at extending our knowledge, and one could say quite simply that the (known) premises make up the evidence for the conclusion. If the explanation would have to account for a *separate*

transfer of evidence seen as necessary for knowledge, the task is different, and possibly much harder. What we wanted explained at first was, I think, the transfer of the epistemic value of the premises such that if we knew the premises we knew the conclusion. The issue now is about achieving that aim through a transfer of evidence from the premises to the conclusion. If we think of inference ultimately as the transfer of knowledge, and we here seem only to be able to transfer grounds for knowledge, (externalistically conceived grounds), without transfer of any awareness that the grounds are grounds, then we seem to fail in our ambition to explain that the inferrer knows the conclusion.

On the other hand, I think Prawitz's reasoning is completely correct at every step given his starting points. But what I am insisting on is that we start in a completely different place; the explanation we seek should focus, I suggest, on the competence exhibited by the inferring person in the individual inference acts, not on the properties of the generic inference as conceived by Prawitz.

This is a point where it might be useful to bring in the practical case once more. In practical inference there is also a transfer of something, but can it be evidence? That seems to be altogether wrong in the practical case. If we have to bring in the practical case, and we do it seems to me, then we're in trouble. It cannot be accommodated by Prawitz's way of thinking. Still, the practical case seems to exhibit exactly the same logical competence as the theoretical. Further, if we think of these skills or competences just as skills we use to extend our knowledge, there is space for the possibility that the knowledge extended can be either type of knowledge, theoretical or practical. Bringing in the practical therefore broadens our view of inference, something I consider a great advantage. It is an advantage of the Fregean view that it seems ideally suited for such a broadening.

This broadening move, someone might think, might not in the end be entirely defensible. Of course, that is up for further discussion. Thinking as I suggest does, however, change the ground on which we stand when we look at how the inferring agent employs their inferential skills and competences. It avoids, at least initially, any commitment to a specific type of meaning theory in explaining the epistemic transfer. On the suggested view, all successful inference extends knowledge, and the same competence/skill is employed in extending both practical and theoretical knowledge.

6. Concluding overview

By bringing in both the practical and the theoretical, we are able to identify a general way of looking at legitimate inferences that is of great theoretical interest, and which brings unity to the basic types of inference: they are both knowledge-extending when they are legitimate. They extend practical and theoretical knowledge respectively.

Such inferences can fail in two ways: we can fail in the inference being correct, and we can fail in the way we relate to a premise, as our relating to that premise might not be legitimate. Legitimacy therefore crops up twice on this Fregean view, and provides a unique way of thinking about inferences. In the last type of case, the case where our relating to a premise is not legitimate, the factivity of the premise will break down, and we, in the theoretical version, will entertain something false or something that is true by luck, and in the practical case something we take ourselves to be doing something we will fail to do. This is so even if the inference would have been legitimate had the premises been legitimate.

In the practical case, then, we fall back on intention in doing it, and we say that that is what I intended to do. In the theoretical case we fall back on belief. Just as belief is then seen as failed (theoretical) knowledge, intention (in this sense of intention, i.e. intention in doing something when we fail to do it) is seen as failed (practical) knowledge. There is luck in the theoretical case, where our belief happens to be true by some fluke. Not so in the case of doing something intentionally. The practical case is, however, different in another way. There is also prior intention, intention prior to the act when, for instance, a temporal gap obtains between the forming of an intention and acting, and there is the deviant case where you happen to do the thing because of the intention you have but you do not do it intentionally.

The practical case highlights the need to ponder whether we should think of inferring as an intentional action. I highlight the point because the conclusion is itself an intentional action in the central practical case. I submit that we should not think of inferring as an intentional action, and that going wrong here shows us one of the ways we might be led to posit a Lewis Carroll type conclusion. This point does not stand in the way of recognizing many intentional actions in the neighbourhood, like engaging in reasoning, deliberating, considering evidence, considering whether something follows etc.

The paper has not attempted to explain what logical competence is, nor to explain what logical correctness is, nor what it is to know logical truths, nor to know that some inference is valid and the conclusion a logical consequence of

the premises. It is not committed to a specific view about how to extend from categorical judgments to hypothetical ones, only to the view that we should think of the basic logical competence as displayed in categorical judgments and intentional actions. Of course, we can do the abstraction Prawitz does and think of the premises simply as assumptions, not as Fregean judgments in the theoretical case. In that scenario, we may be well placed for focusing on whether something *is* a logical consequence of something else, and we might reach a judgment about that. Reaching such judgments is something logic perceived as a discipline does. We reach that judgment by employing logical competence, and we can employ the result to extend and improve our logical competence. We do something similar in reductio arguments, where the focus is different from doing logic; it is to explore the tenability of some assumption.

There is, as pointed out by Dummett, no real justification of deduction, but a sort of explanation. Prawitz's proof that there will be truth-makers for the conclusion in the case of a correct inference from legitimate premises can be put to use in this explanation. I do not want to commit to truth-makers in any metaphysical sense, but bringing in truth-makers (rather than evidence) can be extended to cover both practical and theoretical inference. We only need to think of truth-makers in the right (metaphysically innocent) way.

This paper must remain largely neutral on how we think in more detail about logical knowledge and logical competence and the relationship between them. But the paper does impose some constraints on the theoretical work. It must provide a view on what validity is, a view of logical competence such that we can on the whole see people as inferring correctly and competently in an interesting range of cases. If Prawitz's proof or a Bolzano-Tarski approach can be seen as doing the first part of this, then we can think of the ways of breaking down complex logical steps into simple ones as exhibiting the possibility of there being logical competence among finite beings. Those things together will help explain how there could be such a thing as logical competence. In actual inference, then, for there to be transfer of epistemic value there has to be an employment of this sort of competence, a competence that is generally accounted for on the theoretical level. The important point is that there are different explanations at different levels. That makes the tasks achievable, and also makes them semi-independent, which in its turn changes the dialectical situation

It is hard to see how the work that is needed on this theoretical or meta-level can be further constrained than by making the connections between the explanatory levels possible. When that work is done, we are in a position to understand whether knowledge can be extended by inference in the range of

cases so depicted. Of course, logical competence can vary between individuals, from dogs to logicians, and some individuals are much better at extending their knowledge than others. Also, there is no need to be able to account for an extension of knowledge when the inference is too complex for the skills the inferrer in question possesses. The competence of an actual inferrer can be determined by identifying the kinds of inference the inferrer in question knows how to perform.

From this perspective, a division of labour obtains between explaining on the one side how an actual inferrer can extend knowledge by inference, and, on the other, how inference in general can extend knowledge. The explanations are quite different. The latter question is dealt with by work in logic and logical theory while the former is dealt with by looking at the competence of an individual inferrer. The one question focuses on inference generally conceived and in the abstract, the other on actual inferrings, and can employ virtue epistemology. As I see things, the Fregean conception of inference is exemplified in the latter, and it helps us separate the two, and by separating the explanatory tasks, progress on very difficult questions is made.

7. References

Ainslie, George. *Picoeconomics*. Cambridge, CUP. 1992.

Anscombe, G. E. M. *Intention*, Blackwell's, Oxford, 1957.

—, "Practical Inference". In Hursthouse, Lawrence and Quinn, (eds), *Virtues and Reasons*, Oxford, OUP. 1995, pp. 1-34.

Boghossian, Paul. Bind Reasoning, *Proceedings of the Aristotelian Society*, Supplementary Volume, 2003, 77 pp. 225–248.

Broome, John. 1999: "Normative Requirements", *Ratio*, 12. pp. 398-419.

—, 2001 "Normative Practical Reasoning", *Proceedings of the Aristotelian Society*.

—, 2003 "Practical Reasoning", in Bermudez and Millar (eds), *Reason and Nature. Essays in the Theory of Rationality*. Oxford, OUP, pp. 85-112.

—, "The Unity of Reasoning", in *Spheres of Reason*, edited by Simon Robertson, John Skorupski and Jens Timmerman.

—, "How to be Rational", circulated book manuscript. Published as *Rationality through Reasoning*, Blackwell, Wiley 2013.

—, "Rationality", in *A Companion to the Philosophy of Action*, edited by Timothy O'Connor and Constantine Sandis, Blackwell (2010), pp. 285-92.

Davidson, Donald. *Essays on Actions and Events*. Oxford, OUP, 1980.

Dummett. Michael. 'The Justification of Deduction', in *Truth and Other Enigmas*, Duckworth, London 1978. Pp- 290-319.

Engel, Pascal, Sosa on the Normativity of Belief. *Philosophical Studies*, 166, 617-624, 2013.

Gjelsvik, Olav. 'Understanding Enkratic Reasoning'. In *Organon F*, vol XX, 2013, pp. 464-484..

Kolodny, Niko, 'Why be rational?', *Mind*, 114 (2005), pp. 509-63.

Prawitz, Dag. (2011): Proofs and Perfect Syllogisms, in *Logic and Knowledge*, C. Cellucci et al (red), Cambridge Scholars Publishing, Newcastle on Tyne, pp 385-402.

–, . "Explaining Deductive Inference". To appear in *Dag Prawitz on Proofs and Meaning*, ed. H. Wansing, forthcoming 2013.

Rumfitt, Ian. (2011) "Inference, deduction, logic". In John Bengson and Marc A. Moffett, eds., *Knowing How: Essays on Knowledge, Mind and Action* (New York: Oxford University Press, 2011), pp.333-359.

Sosa, Ernest. *Knowing Full Well*, Princeton, Princeton University Press, 2011.

—, "Responses to four critics", *Philosophical Studies*, (2013) 166, pp 625–636

Smith, Nicholas "Frege's Judgement Stroke and the Conception of Logic as the Study of Inference not Consequence", *Philosophy Compass* 4, 2009, pp. 639-665.

Stanley, Jason. *Knowing How*. Oxford, OUP, 2011.

Wallace, R. Jay. "Normativity, Commitment and Instrumental Reason.". *Philosophers' Imprint*, Vol 1. No 3, 2001, pp. 1-26.

Williamson, Timothy. 'Understanding and Inference', (Symposium on Blind Reasoning), *Aristotelian Society* Supplementary Volume 77, 2003, pp. 249-93.

—, *The Philosophy of Philosophy*, Oxford, Blackwell, 2007

—, "Probability and Danger", *Amherst Lecture in Philosophy*, 4. 1-35.

37

Remarques sur un placard : Descartes contre Regius

ALAIN DE LIBERA

« L'une des grandes différences entre la démarche analytique et la démarche "historienne" en histoire de la philosophie étant, selon lui, que la première privilégie la remontée d'aval en amont d'une problématique contemporaine à ses antécédents passés, alors que la seconde privilégie l'étude de l'émergence de problématiques d'amont en aval, Pascal Engel demande s'il n'y aurait pas « un point où elles puissent se rejoindre, un peu comme deux équipes qui creusent un tunnel de chaque côté d'une montagne » [1]. Il répond positivement, donne quelques exemples – la meilleure démonstration de sa thèse étant au demeurant son propre travail, tant en philosophie de la psychologie qu'en philosophie de la logique. C'est en ce point de rencontre que je voudrais me situer aujourd'hui, en hommage au philosophe, au professeur et à l'ami, en un point où se rencontrent aussi philosophie médiévale et philosophie moderne, et, on ne s'en étonnera pas, plusieurs formes de lectures philosophiques, de la *Geistesgeschichte* à la reconstruction historique et, on l'espère ici, rationnelle : l'invention du « sujet cartésien ».

Dans le §52 des *Principes de la philosophie*, sur lesquels Heidegger a, dans *Sein un Zeit*, fondé son interprétation de la *Substanzialität* du Je ou du Moi cartésien en termes de *Vorhandenheit* (liée à une *non élucidation* du « sens de

[1]. Cf. P. Engel, « Retour aval », *Les Études philosophiques*, n°4/1999, p. 453-463.

l'être » « enveloppé » ou « renfermé dans l'idée de substantialité »)[2], Descartes explique en quoi le nom de « substance peut être *attribué à l'âme et au corps en même sens* ; *et comment on connaît la substance* ». Pour bien comprendre l'argument, il faut citer le texte en entier :

> 52. *Qu'il peut être attribué à l'âme et au corps en même sens, et comment on connaît la substance*. [IXb,47]. Et la notion que nous avons ainsi de la substance créée, se rapporte en même façon à toutes, c'est-à-dire à celles qui sont immatérielles comme à celles qui sont matérielles ou corporelles ; car il faut seulement, pour entendre que ce sont des substances, que nous apercevions qu'elles peuvent exister sans l'aide d'aucune chose créée. Mais lorsqu'il est question de savoir si quelqu'une de ces substances existe véritablement, c'est-à-dire si elle est à présent dans le monde, ce n'est pas assez qu'elle existe en cette façon pour faire que nous l'apercevions ; car cela seul ne nous découvre rien qui excite quelque connaissance particulière en notre pensée. Il faut, outre cela, qu'elle ait quelques attributs que nous puissions remarquer ; et il n'y en a aucun qui ne suffise pour cet effet, à cause que l'une de nos notions communes est que *le néant ne peut avoir aucun attribut, ni propriété ou qualité* : c'est pourquoi, *lorsqu'on en rencontre quelqu'un, on a raison de conclure qu'il est l'attribut de quelque substance, et que cette substance existe* [3].

---

[2] . Cf. *Sein und Zeit*, §20, éd. F.-W. Von Hermann, Francfort, Vittorio Klostermann, [17]1993, GA 2, p. 125 ; [trad. E. Martineau, en ligne]. Sur ce thème, cf. V. Carraud, « Qui est le moi ? », *Les Études philosophiques*, n° 1/2009, p. 61-81, spéc. p. 65-66. Sur Heidegger et Descartes, cf. J.-F. Courtine, « Les méditations cartésiennes de Martin Heidegger », *ibid*., p. 101-113.

[3]. Cette dernière phrase est la version cartésienne de la proposition de fond de l'attributivisme/attributivisme* : « *Qui dit attribut (propriété) dit sujet* ». Angelelli la rapproche de l'argument ontologique. Cf. I. Angelelli, *Studies in Gottlob Frege and the traditional philosophy*, Dordrecht, D. Reidel, 1967, 1.42 (trad. J.-F. Courtine *et al*., Paris, Vrin, 2006) : « [Dans] *Les Principes de la philosophie*, §52 *in fine*, [Descartes] affirme que quand nous rencontrons une propriété nous pouvons conclure qu'*il y a* une substance qui a cette propriété. Ce n'est pas l'argument ontologique. . .mais probablement un argument évident au sujet des propriétés entendues comme accidents individuels. »

L'argument du néant est l'un des arguments favoris de Descartes [4]. Les *Principes* lui font jouer un rôle dans la démonstration métaphysique de la substantialité de l'*âme*, en arguant du fait *qu'on ne peut rien attribuer au néant*, celui-ci (ou pour mieux dire : ce qui n'est rien) n'ayant aucune propriété ou qualité. Le §52, plaide dans « le langage de l'École » pour l'*univocité* du terme « substance » rapporté au corps et à l'âme. Il fait suite au §51, où Descartes a établi son *équivocité* rapporté à Dieu et aux créatures :

> 51. *Ce que c'est que la substance, et que c'est un nom qu'on ne peut attribuer à Dieu et aux créatures en même sens*. [IXb,46]. Pour ce qui est des choses que nous considérons comme ayant [IXb,47] quelque existence, il est besoin que nous les examinions ici l'une après l'autre, afin de distinguer ce qui est obscur d'avec ce qui est évident en la notion que nous avons de chacune. Lorsque nous concevons la substance, nous concevons seulement une chose qui existe en telle façon, qu'elle n'a besoin que de soi-même pour exister. En quoi il peut y avoir de l'obscurité touchant l'explication de ce mot : n'avoir besoin que de soi-même ; car, à proprement parler, il n'y a que Dieu qui soit tel, et il n'y a aucune chose créée qui puisse exister un seul moment sans être soutenue et conservée par sa puissance. *C'est pourquoi on a raison dans l'École de dire que le nom*

[4] . Sur ce point, voir A. de Libera, *Archéologie du sujet*, I, *Naissance du sujet*, Paris, Vrin (Bibliothèque d'histoire de la philosophie), 2007, p. 171, n. 3 ; *Archéologie du sujet*, II, *La Quête de l'identité*, Paris, Vrin, 2008, p. 98 ; *Archéologie du sujet. III/1. L'acte de penser. La double révolution*, Paris, Vrin, 2014, p. 37-39. L'axiome qui le fonde : *nihili nulla sunt attributa* (le néant n'a pas d'attributs) en fait l'argument attributiviste* par excellence. C'est également en s'appuyant sur le *nihil* que, dans les *Troisièmes objections*, Hobbes réfute la thèse de Descartes affirmant que « je suis une chose qui pense, *c'est-à-dire un esprit, une âme, un entendement, une raison* », à savoir, en l'occurrence, en s'appuyant sur le principe que *ce qui pense n'est pas un rien* (« ex eo quod sum cogitans, sequitur, Ego sum, *quia id quod cogitat non est nihil* »). C'est également dans la discussion et la réfutation de ladite thèse que, comme on y reviendra plus bas, Hobbes *introduit la notion de sujet dans le cartésianisme*, créant du même coup *à terme* la possibilité d'un « sujet cartésien » (évidemment différent du sien). Son objectif est clair : il s'agit de montrer que la *chose* qui pense n'est pas *nécessairement* incorporelle ou, ce qui revient au même, qu'elle *peut* être corporelle. C'est dans cet esprit que les *Troisièmes objections* recourent au « sujet » : Hobbes veut montrer que, si tout les monde s'accorde à dire qu'il ne peut y avoir de pensée sans sujet, l'équation *res cogitans* = sujet est non seulement *impossible* chez Descartes, mais nécessairement *exclue*, car l'admettre serait, pour lui ouvrir la porte au matérialisme, autrement dit à la thèse faisant du corps le sujet de la pensée, thèse qu'il refuse *a priori* ou, si l'on préfère, qu'il refuse sans rien prouver, dans la mesure où pour établir sa conclusion (*i.e.* que *la chose qui pense est nécessairement incorporelle*), il prend pour *fondement*, par une grossière pétition de principe, que *la chose qui pense ne peut être sujet de l'esprit*. Sur le débat Hobbes-Descartes, cf. E. Curley, « Hobbes contre Descartes », *in* J.-M. Beyssade, J.-L. Marion (éd.), *Descartes. Objecter, répondre*, Paris, PUF, 1994, p. 149-162.

> *de substance n'est pas univoque au regard de Dieu et des créatures, c'est-à-dire qu'il n'y a aucune signification de ce mot que nous concevions distinctement, laquelle convienne à lui et à elles* ; mais parce qu'entre les choses créées quelques-unes sont de telle nature qu'elles ne peuvent exister sans quelques autres, nous les distinguons d'avec celles qui n'ont besoin que du concours ordinaire de Dieu, en nommant celles-ci des substances, et celles-là des qualités ou des attributs de ces substances.

Dans aucun de ces textes Descartes ne parle du moi ou de l'*ego*, mais seulement de l'âme. *Il ne parle pas non plus de sujet* [5]. Il se demande comment l'on aperçoit clairement et distinctement que l'âme est une substance. Il répond qu'on le voit à ce qu'on lui peut attribuer quelque attribut, propriété ou qualité qui a « besoin » d'elle pour exister. On peut donc dire que, dans les *Principes*, c'est dans le cadre de l'attributivisme* que s'effectue la reconnaissance du caractère substantiel de l'âme : l'âme est substance *ssi* elle est un sujet d'attribution possible pour des attributs. Je rappelle que par attributivisme*, j'entends :

> Attributivisme*$_{\text{déf.}}$ Toute doctrine de l'âme, de la pensée, de l'intellect ou de l'esprit, fondée sur (ou présupposant ou impliquant) une assimilation explicite des états ou des actes psychiques, noétiques ou mentaux à des attributs ou des prédicats d'un *sujet* défini comme *ego*.

L'attributivisme* se distingue de l'attributivisme ou *Attribute-theory* :

> Attributivisme$_{\text{déf.}}$ Toute doctrine faisant de l'âme, de l'esprit, voire de l'intellect une *propriété ou disposition du corps*.

[5] . Sur le « sujet » chez Descartes, voir J.-L. Marion, « Descartes hors sujet », *Les Études philosophiques*, n°1/2009, p. 52-53 : « À notre connaissance, jamais, dans les *Meditationes* ou les *Principia,* Descartes ne nomme l'*ego,* la *mens,* ou la *res cogitans* un *subjectum,* ni *sujet* leurs équivalents français dans le *Discours de la méthode.* Au contraire, les variations de *sujet/subjectum* renvoient le plus souvent à ce qui se trouve *soumis* [...] éventuellement soumis à la pensée elle-même, au titre de ce que nous nommerions facilement aujourd'hui des *objets.* » Selon nous, la remarque vaut, contrairement à ce que dit J.-L. Marion lui-même, pour l'expression « subjectum meae cogitationis » de *Meditationes,* A.–T, VII, 37, 9, qui, loin d'être le « *hapax* d'un subjectum de la pensée » (« Descartes hors sujet », p. 51, n. 3) désigne bel et bien un *objet* de pensée, comme l'avaient compris John Veitch et Julius Heinrich von Kirchmann, qui, au XIX^{e} siècle, traduisaitent respectivement « semper quidem aliquam rem ut *subjectum* meae cogitationis apprehendo », par « I always, indeed, apprehend something as the *object* of my thought » et « Ich erfasse da zwar immer einen Gegenstand als *Unterlage* meines Gedankens. »

L'attributivisme* est le fondement épistémique du substantialisme, défini :

> Substantialisme$_{\text{déf.}}$ Toute doctrine définissant l'âme, l'esprit ou l'intellect comme une substance ou une chose [6].

Dans les *Principes*, le sujet n'a qu'une présence tacite, pour ne pas dire virtuelle. Le mot n'est pas prononcé : il est seulement impliqué par l'utilisation du mot « attribut » .

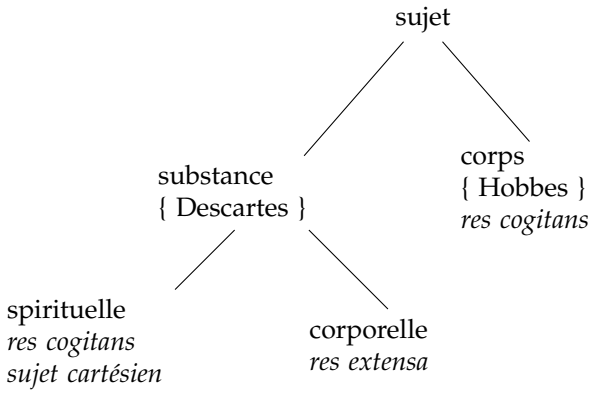
Le §53 pose qu'il y a deux sortes d'attributs : l'attribut principal, qui « constitue la nature ou essence » de la substance et les autres attributs qui « suffisent à faire connaître » qu'elle existe, mais dépendent pour exister eux-mêmes de cette « nature ou essence ». L'attribut principal de l'âme substance(-sujet) est la pensée. L'âme substance(-sujet) est appelée « chose qui pense », l'attribut principal, « pensée », et les divers attributs secondaires « façons différentes de penser » – à savoir dans l'énumération du §53 (dont rien ne dit qu'elle soit exhaustive) : « l'imagination, le sentiment et la volonté » [7]. L'analyse attributiviste est clairement menée dans un cadre « dualiste » : les §52-53 ne mentionnent que deux sortes de « substances » (ou « choses ») qui n'ont besoin pour exister que du « concours ordinaire de Dieu » : l'âme et le corps, la première, immatérielle, ayant pour attribut principal la pensée (d'où la formule : « la chose qui pense », la seconde, matérielle, l'étendue (d'où la formule : « la chose étendue »).

---

[6] . Sur le substantialisme, les dualismes et la « Théorie de l'attribut », cf. A. de Libera, *Naissance…*, p. 154-158 et 162-163 ; cf. en outre D.M. Armstrong, *A Materialist Theory of the Mind*, Routledge, 1993, p. 5-14. Le philosophe définit en ces termes le dualisme dit « cartésien » : **Cartesian dualism**$_{\text{déf.}}$: Any view that holds that a person's mind is a single, continuing, non-material substance in some way related to the body. Pour une discussion de l'*Attribute-Theory* chez Aristote, voir (*pro*) J. Barnes, « Aristotle's Concept of Mind », *Proceedings of the Aristotelian Society*, 1971-2, p. 101-114, et (*contra*), H. Granger, *Aristotle's Idea of the Soul*, Dordrecht, Kluwer, 1996.

[7] . *Principes de la philosophie*, §53, A.-T. IX-2, p. 48 : « Que chaque substance a un attribut principal, et que celui de l'âme est la pensée, comme l'extension est celui du corps. Mais, encore que chaque attribut soit suffisant pour faire connaître la substance, il y en a toutefois un en chacune, qui constitue sa nature et son essence, et de qui tous les autres dépendent. A savoir l'étendue en longueur, largeur et profondeur, constitue la nature de la substance corporelle ; et la pensée constitue la nature de la substance qui pense. Car tout ce que d'ailleurs on peut attribuer au corps, présuppose de l'étendue, et n'est qu'une dépendance de ce qui est étendu ; de même, toutes les propriétés que nous trouvons en la chose qui pense, ne sont que des façons différentes de penser. Ainsi nous ne saurions concevoir, par exemple, de figure, si ce n'est en une chose étendue, ni de mouvement, qu'en un espace qui est étendu ; ainsi l'imagination, le sentiment et la volonté dépendent tellement d'une chose qui pense, que nous ne les pouvons concevoir sans elle. Mais, au contraire, nous pouvons concevoir l'étendue sans figure ou sans mouvement, et la chose qui pense sans imagination ou sans sentiment, et ainsi du reste. »

Dualisme, substantialisme et attributivisme : voilà donc ce qui caractérise la détermination cartésienne de la substantialité(-subjectité) de l'âme.* Point ici de « déduction métaphysique de l'ego » – il faudrait pour cela qu'on entreprît de montrer (ou de rappeler) l'existence d'une équation comme *je* = (une) *chose qui pense*. Dans les *Principes*, Descartes ne parle pas d'*ego*, mais il formule clairement la thèse fondamentale de l'attributivisme* substantialiste dualiste, position philosophique opposée à l'attributivisme substantialiste non-dualiste (fonctionnaliste, naturaliste, matérialiste), où la substance-sujet de la pensée est le corps (l'âme n'étant pas distincte du corps ou n'étant qu'une disposition ou une qualité du corps) – une position proche de celle de Hobbes – les deux positions se laissant ainsi schématiser :



Hobbes est le premier à parler de « sujet » à Descartes. C'est le pivot d'une des critiques qu'il lui adresse dans les *Troisièmes objections* [8]. Résumons : le refus de *sub-jecter* l'esprit, l'âme, l'entendement ou la raison dans la *chose qui pense* (pour Hobbes, évidemment, le corps) conduit Descartes à identifier la *chose qui pense* à l'esprit (ou à l'âme, ou à l'entendement ou à la raison) en l'identifiant pour ce faire aux actes de l'esprit et aux facultés mentales correspondantes. Descartes élimine toute référence à un sujet, et pose :

$$\text{esprit} = \text{actes} = \text{facultés}$$

[8]. Cf. *Œuvres de Descartes*, éd Ch. Adam & P. Tannery (= A.-T.), Paris, Vrin, 1964–1974, VII, p. 172-173 (*Meditationes de prima philosophia*) et IX (*Méditations*), p. 134

C'est, pour Hobbes, une erreur majeure [9]. Pour l'historien attentif à l'usage des termes, c'est la preuve que le cartésianisme n'a originairement que faire du supposé *sujet cartésien*. Au contraire, sans *élimination du sujet* (*i.e.* du corps sujet, la thèse non cartésienne par excellence), on ne saurait avoir ni « substantialisme » ni « dualisme » *cartésiens* (*i.e.* les deux thèses authentiques de Descartes). Telle est, pour *nous*, l'ombre portée de la lecture hobbésienne du *cogito*. Une ombre *singulière*, qui peut en cacher une autre et, de fait, la cache pour la majorité des interprètes : le déficit historiographique des lectures du cartésianisme axées sur la solidarité, indéniable *a parte post*, mais fourvoyante *a parte ante*, du sujet cartésien et du sujet transcendantal. Si l'on veut échaper au « mouvement rétrograde du vrai » en histoire, il faut, s'agissant de Descartes et du sujet, lire Descartes à partir de Hobbes plutôt qu'à partir du binôme formé par Schelling et Heidegger [10]. Cela veut dire saisir le « sujet » là où il *entre par effraction* dans l'univers de discours cartésien. Cela veut dire le regarder apparaître *de l'extérieur* dans le débat de Descartes avec Hobbes, plutôt que s'esquisser « subjectivement », *de l'intérieur*, dans une anticipation tâtonnante du *je pense* comme « sujet insigne », conduisant par une sorte de nécessité immanente au « je suis libre » de Kant, puis d'une volonté l'autre, au je *(me)* veux de la Volonté de volonté nietzschéenne entendue comme Volonté de puissance. Le rejet de l'attributivisme est fondamental dans le cartésianisme. Il commande celui de la sub-jectité. Le corps n'est pas le *sujet* de l'âme. Partant, il n'y a pas de sujet de ma pensée : ni ce corps que « je » ne suis pas ; ni ce « je », ni ce « moi » qui devraient *composer* avec ce *corps-sujet* pour accéder à une fonction, la *fonction-sujet*, qui ne ferait, m'égalant à mon corps, que me soustraire à la dignité de chose pensante et, ainsi, m'ôter à moi-même. C'est pourquoi nous disons qu'il n'y a pas *originairement* chez Descartes de *sujet pensant*. Le sujet cartésien est d'abord un *article d'importation anglaise* ; on verra bientôt qu'il est aussi batave. Angleterre, Pays-Bas : *pauvre France*.

Si Descartes n'est pas attributiviste, il est en revanche attributiviste*. Parler d'attributivisme* à propos de Descartes n'a philosophiquement rien d'anachronique. Le langage de l'attribut est fondateur. Posé qu'il y a deux sortes

[9]. La position de Hobbes est fondée sur deux équivalences : *sujet* = *matière* (matière sujette) et *actes* = *propriétés* (attributs ou dispositions), qui permettent de mettre en implication réciproque attributivisme et attributivisme*. La position hobbésienne se définit donc : attributivisme « attributivisme*. L'esprit est une disposition du corps ; la pensée, un acte du corps ainsi disposé (une activité, donc, mais aussi un acte au sens où, selon Hobbes, lecteur de Descartes, les qualités ou propriétés de la cire sont les actes d'une même chose ou matière sujette).

[10]. Sur ce thème, cf. A. de Libera, « Sujet insigne et *Ich-Satz*. Deux lectures heideggériennes de Descartes », *Les Études philosophiques*, n° 1/2009, p. 83-99.

de substances, que chaque substance a un attribut, que la pensée est l'attribut principal de l'esprit et que l'extension est l'attribut principal du corps, le problème de Descartes est de montrer que chacune, esprit et corps, a *un seul attribut* principal. De fait, pour surmonter les objections de Hobbes, il faut exclure qu'*une* même substance, l'esprit ou le corps, puisse avoir *deux* attributs principaux, la pensée et l'extension, autrement dit se doter sur ce point d'une prémisse additionnelle, que M. Rozemond appelle « *The Attribute Premise* », prémisse qui, toutefois, selon elle, « is generally not at all explicit when Descartes argues for the real distinction » [11]. Le *fond* du problème cartésien est celui de *l'unité* d'un homme *composé de deux substances hétérogènes* : l'âme ou esprit et le corps.

Distinguant *opposita* et *diversa*, Martial Gueroult commente : « s'il s'agissait là seulement, comme chez Aristote, de deux *opposita*, extrêmes d'un même genre (matière et forme, puissance et acte), leur unité serait celle de leur genre commun et ne poserait aucun problème. Mais comme il s'agit, en l'espèce, de *diversa*, c'est-à-dire de réalités uniques en leur genre et par conséquent incommensurables, il est contradictoire que », en l'homme, « elles ne fassent qu'un » [12]. Le problème en un sens est comparable à celui de l'union des deux natures dans le Christ : celui que la tradition théologique a justement appelé « union *hypostatique* ». Qu'il y ait une seule nature en l'homme, bien qu'il soit constitué de deux substances – hypostases ou *sujets* – voilà ce qu'il faut comprendre : c'est la réciproque du problème christologique, qui est de comprendre l'union de deux natures dans une seule hypostase [13]. La thèse hobbésienne est une sorte d'hérésie anthropologique, qui professe l'existence

[11]. Cf. M. Rozemond, « Descartes's case for Dualism », in *Descartes's Dualism*, Cambridge (Mass.), Harvard University Press, 1998, chap. I, p. 1-37 (reprise de l'article paru en 1995, dans le *Journal of the History of Philosophy*, 33, p. 29-63).

[12] . Cf. M. Gueroult, *Spinoza, I : Dieu (Ethique, I)*, Paris, Aubier-Montaigne, 1968, p. 229-230.

[13] . On notera que le problème christologique comporte une variante contrefactuelle également transposable au problème de l'union de l'âme et du corps : cette variante, liée à la théorie de la toute-puissance divine est ce que l'on appelle depuis Oberman *« asinus-Christology »*, à savoir que Dieu aurait pu assumer la nature d'âne, au lieu de la nature d'homme, s'il l'avait voulu ; on peut, de fait, dans le même sens demander si Dieu aurait pu unir l'âme d'un homme à un autre corps que le nôtre. Sur la « christologie de l'âne », cf. H.A. Oberman, *The Harvest of Medieval Theology : Gabriel Biel and Late medieval nominalism*, Cambridge (Mass.), Harvard UP, p. 255, §3 : « The intention of Occam's *asinus*-Christology : rejection of the Charge of Nestorianism ». Sur ce thème, cf. J. Haga, *Was there a Lutheran Metaphysics ? The interpretation of communicatio idiomatum in Early Modern Lutheranism*, [202A ?]Göttingen, Vandenhoeck & Ruprecht, 2012[202C ?], p. 78 et O. D. Crisp, *Divinity and Humanity. The Incarnation Reconsidered*, Cambridge (UK), CUP, p. 84-85. Sur le nestorianisme d'Occam, cf. Marilyn McCord Adams, « Relations, Inherence and Subsistence : or, Was Ockham a Nestorian in Christology ? », *Noûs*, Vol. 16, No. 1, 1982, p. 62-75. Une des premières discusions de l'hypothèse de l'âne figure chez Bonaventure, *Sent*. III, d.2, q.1.

de deux attributs contraires dans une même et unique substance. Mais l'orthodoxie anthropologique cartésienne, qui professe l'opposé : l'existence d'une seule nature, la nature humaine, *constituée* par l'union de deux substances incommunicables, tient du mystère. Pourtant, pour citer encore Gueroult, « c'est un fait qu'elles constituent dans l'homme une seule et même nature », et que, « de ce fait absurde, un sentiment invincible nous atteste la réalité ». Insoluble pour les idées claires et distinctes, le problème de Descartes se résout par une sorte *d'extension de garantie*, prenant effet de la véracité divine.

> La raison doit seulement reconnaître que Dieu, étant vérace, n'a pas pu vouloir nous tromper en mettant en nous ce sentiment sans rien nous donner pour démentir son enseignement ; en conséquence, elle garantit que ces deux natures, bien que conservant toujours chacune son essence irréductible, sont effectivement unies en nous de façon incompréhensible [14].

Mystère ou pas, c'est en mettant Descartes face au problème de l'attribut unique, que ses adversaires lui ont imposé, une deuxième fois après Hobbes, mais plus efficacement, de *parler sujet*. Bizarrement, le coup est parti d'un ancien cartésien, qui l'avait entraîné dans ce qu'on a appelé la « Querelle d'Utrecht », et auquel il avait longuement prodigué ses conseils, face aux aristotéliciens calvinistes de l'université, menés par le recteur Gijsbert Voet (Gisbertus Voetius, 1589–1676). C'est à cette seconde polémique que l'on va à présent s'intéresser. Place donc à Henricus Regius (Hendrik De Roy, 1598-1679), alias Henry Le Roy, et à sa petite affiche : l'*Explicatio Mentis humanæ, sive Animæ rationalis, ubi explícatur quid sit, et quid esse possit* [15].

On notera que certains distinguent union hypostatique et union personnelle. Cf., sur ce point, M. McCord Adams, *What sort of human nature ? : Medieval philosophy and the systematics of Christology (The Aquinas Lecture 1999)*, Marquette University Press, 1999, p. 29, qui distingue l'union hypostatique *potentiellement ouverte* et l'union personnelle *réclamant l'humanité* : « Natures that lack life, sense, or reason are mere vestiges that are incapable of personal (as opposed to mere hypostatic) union. But rational natures are made in God's image and likeness and are capable of making deiformity explicit. Thus, hypostatic union with a donkey nature would not show forth God's power, goodness, and wisdom, in the way personal union with a rational nature would. » La question mériterait d'être reprise au double niveau christologique et anthropologique. Elle est cruciale pour le cartésianisme, et communique avec les problèmes d'identité contrefactuelle.

[14]. Cf. M. Gueroult, *Spinoza…*, *loc. cit*., p. 230.

[15]. Un étudiant de Regius, Petrus van Wassenaer a joué un rôle dans la confection du « programme ». Sur ce point, cf. Th. Verbeek, « Le contexte historique des *Notae in programma quoddam* », *in* Th. Verbeek (éd.), *Descartes et Regius. Autour de l'Explication de l'esprit humain*, Amsterdam-Atlanta, Rodopi (Studies in the History of Ideas in the Low Countries 2), 1993, p. 1 : « Avant

*

Le « sujet cartésien » est sorti du placard en janvier 1648, avec les *Notae in Programma* publiées en réponse au libelle de Regius, et mises à l'*Index* dès 1663. Traduites par Claude Clerselier dans son édition des *Lettres de Mr. Descartes* (1657), les *Notae* et l'*Explicatio* de « Monsieur Le Roy », ont circulé en français sour le titre de *Remarques de René Descartes sur un certain placard imprimé aux Pays-Bas vers la fin de l'année 1647, qui portait ce titre : Explication de l'esprit humain, ou de l'âme raisonnable : où il est montré ce qu'elle est, et ce qu'elle peut être* [16]. Brentano est, à ma connaissance, le premier (et quasiment le dernier) philosophe à avoir partiellement fondé son interprétation de Descartes sur ce texte négligé par l'historiographie [17]. Le ton de Descartes y est des plus

d'être divulguées comme placard, les ving-et-une thèses qui composent *L'Explication de l'esprit humain* de Henricus Regius (1598-1679) étaient attachées, comme des corollaires, à une disputation médicale, proposée pour être publiquement discutée dans l'université d'Utrecht, le 2 octobre 1647. Il s'agit d'une "*disputatio sub praeside*", exercice académique dont le texte avait été écrit par un professeur ou "*praeses*", en l'occurrence Regius, mais qui devait être défendu par un étudiant, le "répondant", qui, dans ce cas-ci, était l'étudiant de médecine Petrus Wassenaer. Enfin, les corollaires ne faisaient pas partie de l'argument. Proposés le plus souvent par le répondant, ils roulaient sur des questions d'actualité, dont le lien avec le sujet de la thèse était assez lâche. D'une façon générale, le "*praeses*" devait les trouver "défendables" mais sa responsabilité était moins stricte que pour les "thèses". En l'occurrence, ils étaient présentés comme des *corollaria respondentis*, ce qui impliquerait qu'ils avaient été écrits par Wassenaer. » Le titre de l'édition latine de 1657 peut laisser entendre cette « collaboration ». Cf. *Brevis explicatio mentis humanae, sive, Animae rationalis antea publico examini proposita* [= par Petrus Wassenaer]*, et deinde operà Henrici Regii ... nonnihil dilucidata, & à Notis Cartesii vindicata. Editio postrema prioribus auctior et emendatior, ad calumniarum quarundam rejectionem, nunc evulgata, Trajecti ad Rhenum, Typis Theodori ab Ackersdijck & Gisberti à Zijll, anno 1657*. Sur Wassenaer, cf. le *Biographical Lexicon*, publié en appendice de sa thèse par J.J.F.M. Bos, *The Correspondence between Descartes and Henricus Regius De briefwisseling tussen Descartes en Henricus Regius*, Dissertation (Proefschrift Universiteit Utrecht), 2002, publiée en ligne dans la série *Quaestiones Infinitae* (Publications of the Department of Philosophy Utrecht University, volume XXXVII, p. 255, s.v. « WASSENAER, Petrus († Utrecht 1680) ».

[16]. Cf. *Lettres de Mr Descartes*, Claude Clerselier (éd.), 1657-1667, t. I, p. 434-462, repris dans *Oeuvres philosophiques*, textes établis, présentés et annotés par Ferdinand Alquié, t. III (1643-1650), Paris, Classiques Garnier, 1973, p. 787-820. Sur le dossier et le texte, cf. Descartes, *Lettres à Regius* et *Remarques sur l'explication de l'esprit humain*. Texte latin, trad., introd. et notes, par G. Rodis-Lewis, Paris, Vrin, 1959. Cf., également, René Descartes. *Die Prinzipien der Philosophie.* Mit Anh. Bemerkungen René Descartes über ein gewisses in den Niederlanden gegen Ende 1647 gedrucktes Programm, Übers. u. erl. von Artur Buchenau (7. Aufl., Reprogr. d. 4. Aufl. 1922), Hambourg, Meiner (Philosophische Bibliothek, 28), 1965. Pour la correspondance ave Regius, on peut également se reporter désormais à la belle édition de J.-R. Armogathe, René Descartes, *Correspondance*, 2, in *Œuvres complètes*, Paris, Gallimard (Tel), 2013, p. 729-786.

[17] . Cf. F. Brentano, *Vom Ursprung sittlicher Erkenntnis*, Leipzig, Dunker & Humblot, 1889 (2e éd. par Oskar Kraus, Hamburg, Meiner 1921 ; 3e éd. Meiner, 1934) ; trad. fr. *L'Origine de la connaissance*

vifs. Pris sous le feu croisé de la *Consideratio theologica* de Revius [18] et de l'affiche de Regius, son ancien disciple, il adresse à l'un son mépris et réplique à l'autre en ces termes :

morale [suivi de *La Doctrine du jugement correct*], trad. M. de Launay & J.-C. Gens, Préface de J.-C. Gens, Paris, Gallimard, 2003. Brentano mobilise les *Notae* pour prouver que, contrairement aux idées reçues, Descartes soutient la division tripartite des « phénomènes psychiques » en trois « classes fondamentales » réintroduite dans la *Psychologie vom emprisichen Standpunkt* de 1874, en représentations, jugements et mouvements affectifs (émotions).

[18]. Jakob Reefsen (1586-1658). Pour une édition moderne de l'oeuvre du théologien réformé, cf. Jacobus Revius, *A Theological Examination of Cartesian Philosophy. Early Criticisms (1647)*, éd. A. Goudriaan, Leyden, E.J. Brill (Kerkhistorische Bijdragen, 19), 2002.

Alius autem libellus magis me movet : quamvis enim nihil in eo aperte de me habeatur, prodeatque sine nomine Authoris et Typographi, quia tamen continet opiniones quas judico perniciosas et falsas, editusque est forma Programmatis, quod vel templorum valvis affigi, et quibuslibet legendum obtrudi possit, dicitur autem jam antea typis mandatus fuisse sub alia forma cum adjuncto nomine cujusdam, tanquam Authoris, quem multi putant non alias quam meas opiniones docere : cogor detegere ejus errores, ne mihi forte imputentur ab illis, qui casu incident in obvias istas chartas, et mea scripta non legerunt.

Pour l'autre [livret, sc. celui de Regius], je m'en mets davantage en peine ; car bien qu'il ne comprenne rien qui s'adresse ouvertement à moi, et qu'il paraisse sans aucun nom, ni de l'auteur ni de l'imprimeur, toutefois, parce qu'il contient des opinions que je juge être très pernicieuses et très fausses, et qu'il a été imprimé en forme de placard, afin qu'il pût être commodément affiché aux portes des temples, et ainsi qu'il fût exposé à la vue de tout le monde, et aussi parce que j'ai appris qu'il a déjà été une autre fois imprimé en une autre forme, sous le nom d'un certain personnage qui s'en dit l'auteur, que la plupart estiment n'enseigner point d'autres opinions que les miennes, je me trouve obligé d'en découvrir les erreurs, de peur qu'elles ne me soient imputées par ceux qui, n'ayant pas lu mes écrits, pourront par hasard jeter les yeux sur de telles affiches.

Le *Placard* s'inscrit dans un contexte particulier. Un contentieux existe en effet entre les deux hommes depuis que, dans la *Préface* des *Principes de philosophie*, Descartes a pris officiellement ses distances avec les *Fundamenta Physices* publiés – malgré son *veto* – par Regius en 1646. C'est de ces lignes meutrières que le philosophe néerlandais entend tirer raison :

> Je sais bien qu'il y a des esprits qui se hâtent tant, et usent de si peu de circonspection en ce qu'ils font, que, même ayant des fondements bien solides, ils ne sauraient rien bâtir d'assuré ; et parce que ce sont d'ordinaire ceux-là qui sont les plus prompts à faire

> des livres, ils pourraient en peu de temps gâter tout ce que j'ai fait, et introduire l'incertitude et le doute en ma façon de philosopher, d'où j'ai soigneusement tâché de les bannir, si on recevait leurs écrits comme miens, ou comme remplis de mes opinions. J'en ai vu depuis peu l'expérience en l'un de ceux qu'on a le plus cru me vouloir suivre, et même duquel j'avais écrit, en quelque endroit, « que je m'assurais tant sur son esprit, que je ne croyais pas qu'il eût aucune opinion que je ne voulusse bien avouer pour mienne » : car il publia l'an passé un livre, intitulé *Fudamenta Physicæ*, où, encore qu'il semble n'avoir rien mis, touchant la physique et la médecine, qu'il n'ait tiré de mes écrits, tant de ceux que j'ai publié que d'un autre encore imparfait touchant la nature des animaux, qui lui est tombé entre les mains, toutefois, à cause qu'il a mal transcrit, et changé l'ordre, et nié quelques vérités de métaphysique, sur qui toute la physique doit être appuyée, je suis obligé de le désavouer entièrement, et de prier ici les lecteurs qu'ils ne m'attribuent jamais aucune opinion, s'ils ne la trouvent expressément en mes écrits, et qu'ils n'en reçoivent aucune pour vraie, ni dans mes écrits ni ailleurs, s'ils ne la voient très clairement être déduites des vrais principes.

Il le fait en faisant imprimer et placarder *in Belgio* vingt-et-une thèses susceptibles de compromettre Descartes – qui durant l'hiver 1647 réside encore au Pays-Bas –, dans la mesure où, bien que *foncièrement opposées* à celles du philosophe, certaines *pourraient sembler* rendre un son cartésien ou se pouvoir tirer de certains de ses principes. Il avait, il est vrai, de bonnes raisons de redouter le style d'application du principe de charité interprétative pratiqué dans les années 40 par ses adversaires bataves, après l'épisode de l'*Admiranda Methodus Philosophiae Renati des Cartes* (1643), et ses démêlés avec le recteur Voetius et son homme de main Martin Schoock (le bien nommé) [19]. On peut prendre *doublement* au sérieux le désir cartésien de se « démarquer » de Regius, quand on relit le *best-off* des épithètes dont l'avait couvert le tandem d'Utrecht : après

[19] . A prendre les choses au pied de la lettre Martin Schoock est le premier inventeur du « sujet cartésien » : c'est lui qui, en effet, reproche au philosophe des « globes éthériens » de se prendre pour le *sujet unique de la Raison*, en s'arrogeant son monopole, « telle qu'elle existe en lui *subjectivement* ». Voir sur ce point, A. de Libera, *La Double révolution*, p. 29-31 et 615-616, d'après M. Schoock, *L'Admirable Méthode* de René Descartes, Section II, chapitre vii, *in* Th. Verbeek, *La Querelle d'Utrecht. René Descartes et Martin Schoock*, préface de J.-L. Marion, Paris-Bruxelles, Les Impressions Nouvelles (Bâtons rompus), 1988, p. 239-240.

avoir été successivement traité, à cause de lui, de « bipède », de « bouche menteuse » vomissant « des calomnies », de « bâtard du christianisme », et de « girouette de toutes les heures » ; après avoir été accusé, « tout en voulant avoir l'air de combattre les athées par ses achilles », d'« injecte[r] finement et secrètement le venin de l'athéisme à ceux qui, par la faiblesse de leur entendement, ne sont pas à même de surprendre partout le serpent caché sous l'herbe », Descartes n'avait sans doute plus qu'un désir : *lui présenter l'adddition*. C'est ce qu'il fait dans ses *Remarques sur le susdit placard*.

Mais c'est précisément en répondant à l'*Explicatio* « pour dissiper toute confusion [20] », que l'auteur des *Méditations* est attiré sur le terrain du sujet. Un terrain où nul ne saurait dire exactement *qui* parle *avant* d'avoir lu le *démenti* de Descartes. Un sujet *ventriloque* que le tour d'écriture pervers de Regius installe anonymement au centre d'un dispositif où, jusque là, on l'a dit, il n'avait rien à faire.

La sub-jectivation, entendons : l'application de ce que j'appelle, suivant Heidegger, le schème de la « sub-jectité » (*Subiectität*)[21] à la *mens* / à l'*esprit* (Clerselier), intervient dès le deuxième des 21 « articles » du placard.

[20] . Cf. F. Alquié, *in* Descartes, *Œuvres philosophiques*, p. 788, n. 1. Pour la controverse d'Utrecht et la *Lettre apologétique* écrite par Descartes en réponse à ses accusateurs, voir la somme de Th. Verbeek citée *supra*. Pour un panorama d'ensemble, cf. du même, *Descartes and the Dutch. Early Reactions to Cartesian Philosophy, 1637-1650*, Carbondale-Edwardsville, Southern Illinois University Press, 1992.

[21] . J'emprunte le schème à Heidegger, qui par '*Subiectität*' désigne le fait d'être sujet, au sens originaire du terme grec ὑποκείμενον ou du latin *subiectum* : être le support, le substrat de qualités ou d'accidents, être une substance, un constituant, existant par lui-même, stable et permanent, de la réalité physique. « Subjectivité », *Subjektivität*, terme introduit par Kant, désigne l'application de ce schème à l'esprit. Cf. M. Heidegger, *Die Metaphysik als Geschichte des Seins* in *Nietzsche*, t. II, Pfullingen, Neske, 1961, p. 399-458 [GA 6.2] = « La métaphysique comme histoire de l'être », in *Nietzsche*, t. II, trad. fr. P. Klossowski, Paris, Gallimard, 1971, p. 319-365. La thèse centrale de Heidegger est que la « subjectivité » (*Subjektivität*) de la métaphysique moderne est un « mode de la subjectité » (*Subjectität*). Cf. M. Heidegger, *Die Metaphysik …*, « *Subiectität und Subjektivität* », GA 6.2, p. 411. Pour une formulation plus fine, englobant à la fois, Kant et Leibniz, cf. *Die Grundprobleme der Phänomenologie*, éd. F.-W. Von Hermann, Francfort,Vittorio Klostermann, 1975, GA 24, p. 178 ; trad. J.-F. Courtine, *Les problèmes fondamentaux de la phénoménologie*, Paris, Gallimard, 1985, p. 159.

Quantum ad naturam rerum attinet, ea videtur pati, ut mens possit esse vel substantia, vel quidam substantiæ corporeæ modus ; vel, si nonnullos *alios Philosophantes* sequamur, qui statuunt extensionem et cogitationem esse attributa, quæ certis substantiis, tanquam subjectis, insunt, cum ea attributa non sint opposita, sed diversa, nihil obstat, quo minus mens possit esse attributum quoddam, eidem subjecto cum extensione conveniens, quamvis unum in alterius conceptu non comprehendatur. Quicquid enim possumus concipere, id potest esse. Atqui, ut mens aliquid horum sit, concipi potest ; nam nullum horum implicat contradictionem. Ergo ea aliquid horum esse potest.	Pour ce qui est de la nature des choses, rien n'empêche, il semble, que l'esprit ne puisse être ou une substance, ou un certain mode de la substance corporelle ; ou si nous voulons suivre le sentiment de quelques *nouveaux philosophes*, qui disent que l'étendue et la pensée sont des attributs qui sont en certaines substances, comme dans leurs propres sujets, puisque ces attributs ne sont point opposés, mais simplement divers, je ne vois pas que rien puisse empêcher que l'esprit, ou la pensée, ne puisse être un attribut à un même sujet que l'étendue, quoique la notion de l'un ne soit pas comprise dans la notion de l'autre : dont la raison est que tout ce que nous pouvons concevoir peut aussi être. Or est-il que l'on peut concevoir que l'esprit humain soit quelqu'une de ces choses, car il n'y a en cela aucune contradiction ; et partant il en peut être quelqu'une.

Cette thèse fait partie, avec la thèse III et la thèse XIII, de celles que Regius (comme annoncé dans sa lettre du 23 juillet 1645 [22]) avait retirées des *Fundamenta Physices* à la demande de Descartes, à savoir :

[22]. Cf. A.-T. IV, p. 254-256 (n° 393) ; A. Baillet, *La vie de Monsieur Des Cartes*, II, 1691, p. 269-271 (n° 34). Sur ce point, cf. J.J.F.M. Bos, *The Correspondence…*, *loc. cit.*, p. 190, avec la note 5.

III. Errant itaque, qui asserunt, nos humanam mentem clare et distincte, tanquam necessario a corpore realiter distinctam, concipere.	III. C'est pourquoi ceux-là se trompent, qui soutiennent que nous concevons clairement et distinctement l'esprit humain comme une chose qui actuellement et par nécessité est distincte réellement du corps.
XIII. Atque ideo omnes communes notiones, menti insculptæ, ex rerum observatione vel traditione originem ducunt.	XIII. Et partant toutes les communes notions qui se trouvent empreintes en l'esprit tirent toute leur origine ou de l'observation des choses ou de la tradition.

Elle fait suite à l'article premier, qui donne la définition de la *mens humana* :

I. Mens humana est, qua actiones cogitativæ ab homine primo peragantur ; eaque in sola cogitandi facultate, ac interno principio, consistit.	I. L'esprit humain est ce par quoi les actions de la pensée sont immédiatement exercées dans l'homme ; et il ne consiste précisément que dans ce principe interne, ou dans cette faculté que l'homme a de penser.

La démarche est biaisée, qui consiste à tirer – ou plutôt à affecter de tirer – d'une définition de l'esprit humain comme *facultas (principium internum) cogitandi*, qui a toutes les apparences du cartésianisme, une position, dite (en français – déjà) des « nouveaux philosophes » (le latin dit simplement : *alios Philosophantes*), qui est si peu compatible avec lui qu'elle revient à attribuer à Descartes une version simplifiée de la thèse de Hobbes, stipulant que, puisque « il n'y a en cela aucune contradiction », il se peut qu'un *même sujet* ait pour attributs la pensée et l'étendue. On voit bien ici quelle est la fonction du « sujet » : enrôler de force Descartes sous la bannière de…l'attributivisme. Il suffit, en effet, de remplacer « sujet » par « corps » dans le schéma imposé par Regius pour obtenir *le contraire exact* de la position des paragraphes 52-53 des *Principes de philosophie*, à savoir :

étendue → sujet (corps) ← pensée

Contrairement à ce que suggère la lecture idéaliste, kanto-schellingienne, du *Ich-denke*, entérinée par Heidegger, le sujet n'entre chez Descartes que pour faire pièce à la distinction de la *res cogitans* et de la *res extensa* ou, si l'on préfère, à la distinction « réelle » de l'esprit *(mens)* et du corps ; il n'a pour fonction, éminemment polémique, que de confronter le cartésianisme au *problème de l'unicité de l'attribut principal* ; d'obliger Descartes à démontrer qu'une substance ne peut avoir *plus d'un* attribut principal, bref à poser comme thèse fondamentale explicite cette « *Attribute Premise* » qui, selon M. Rozemond, manque « en général » à sa défense du « dualisme ».

Or c'est bien au dualisme que s'attaque Regius, et sur deux fronts depuis longtemps ouverts : le premier, dans l'article IV, donnant à entendre que *la distinction réelle n'a pour seul témoin que l'Écriture sainte*, et, douteuse en elle-même, ne saurait être tenue pour certaine que par la foi seule, fait revivre en sous-main, particulièrement dans la version française, plus développée que la latine, le fantasme « averroïste » de la double vérité :

IV. Quod autem mens revera nihil aliud sit quam substantia, sive ens realiter a corpore distinctum, et actu, ab eo separabile et quod seorsim per se subsistere potest : id in Sacris Literis, plurimis in locis, nobis est revelatum. Atque ita, quod per naturam dubium quibusdam esse potest, per divinam in Sacris revelationem nobis jam est indubitatum.	IV. Mais maintenant, qu'il soit vrai que l'esprit humain soit en effet une substance, et qu'il en puisse être actuellement séparé, et subsister de soi-même sans lui, cela nous est révélé en plusieurs lieux de la Sainte Écriture ; et ainsi ce qui de sa nature peut-être douteux pour quelques-uns (au moins si nous ne nous contentons pas d'une légère et morale connaissance des choses, mais si nous en voulons rechercher exactement la vérité) nous est maintenant devenu certain et indubitable, par la révélation qui nous en a été faite dans les Saintes Lettres.

Le second, dans l'article VI, qui, feignant d'accorder la distinction réelle, en limite immédiatement le sens et la portée, en faisant valoir que, quand même elle serait une substance réellement distincte du corps, la *mens* humaine n'en reste(rait) pas moins liée à lui en toutes ses opérations : l'activité de l'esprit étant tout entière modulée sur les « dispositions » du corps – une version radicale, matérialiste, de l'attributivisme.

VI. Mens humana, quamvis sit substantia a corpore realiter distincta, in omnibus tamen actionibus, quandiu est in corpore, est organica. Atque ideo, pro varia corporis dispositione, cogitationes mentis sunt variæ	VI. Quoique l'esprit humain ou l'âme raisonnable soit une substance distincte réellement du corps, néanmoins, pendant qu'elle est dans le corps, elle est organique en toutes ses actions : c'est pourquoi, selon les diverses dispositions du corps, les pensées de l'âme sont aussi diverses.

Si l'on veut répondre efficacement à Regius, c'est donc l'article II qu'il faut réfuter, car en un sens c'est de lui que tout dépend : prouver *l'impossibilité d'un sujet unique de la pensée et de l'étendue*, c'est à la fois réfuter le matérialisme et sauver le dualisme substantialiste, en établissant philosophiquement la vérité de la thèse de *l'unicité de l'attribut principal en chacune des deux sortes de substances dont l'homme est composé*. C'est ce que fait Descartes en discutant pied à pied chaque phrase de son adversaire.

La thèse de Regius (notée ici TR_1) repose sur une erreur grossière : l'affirmation que …

> TR_1 : il ne répugne point à la nature des choses que l'esprit humain puisse être une substance, ou un certain mode de la substance corporelle (« … videri rerum naturam pati ; ut mens humana possit esse vel substantia, vel quidam substantiæ corporeæ modus [23] »).

TR_1 renferme en effet une contradiction, puisqu'elle équivaut à l'assertion que :

[23]. Cf. Alquié, p. 795 ; A.-T., VIII-2, p. 347.

> *TR_1 : il ne répugne point à la nature des choses qu'une montagne soit sans vallée, ou avec une vallée (« … rerum naturam pati, ut mons possit esse vel sine valle vel cum valle [24] »).

Or, il n'est pas de la nature d'une montagne d'être ou de n'être point sans vallée – le thème, *squisitamente* helvétique, est particulièrement cher à Descartes, qui l'utilise en maintes circonstances et dans les contextes les plus divers, la plupart du temps, toutefois, pour en faire un paradigme de l'*impossible* logique. Comme le dit la lettre à Mersenne du 15 novembre 1638, « il n'est pas moins impossible qu'un espace soit vide, qu'il est qu'une montagne soit sans vallée »[25], car, comme le dit la lettre du 9 janvier 1639 au même : « l'idée d'une montagne est comprise dans celle d'une vallée »[26].

Regius commet en fait deux erreurs :

1° il ne distingue pas les *choses contingentes*, pour lesquelles il ne répugne pas à leur nature qu'une chose se comporte ou bien d'une manière ou bien d'une autre (« ut illa vel uno, vel alio modo se habeant ») : « comme que j'écrive maintenant ou que je n'écrive pas », et *l'essence d'une chose*, dont la

[24]. Cf. Alquié, p. 795-796 ; A.-T., VIII-2, p. 347.

[25] . A.-T., II, p. 440.

[26] . A.-T., II, p. 482. L'*exemplum* de la montagne et de la vallée est repris dans la lettre à Gibieuf du 19 janvier 1642 (A.-T., III, p. 476), la V^e *Méditation* (A.-T., IX-1, p. 51-52, notamment : « …l'existence ne peut non plus être séparée de l'essence de Dieu, que de l'essence d'un triangle rectiligne la grandeur de ses trois angles égaux à deux droits, ou bien de l'idée d'une montagne l'idée d'une vallée ; en sorte qu'il n'y a pas moins de répugnance de concevoir un Dieu [c'est-à-dire un être souverainement parfait] auquel manque l'existence [c'est-à-dire auquel manque quelque perfection], que de concevoir une montagne qui n'ait point de vallée ») et dans les *Principes de la philosophie*, §18 (A.-T., IX-2, p. 72). Il est mentionné par Caterus (Jan van Kater, ca. 1590-1655/6) dans les *Premières objections* (A.-T., IX-1, p. 93). Spinoza le reprend à son compte dans son *Court traité [sur Dieu, l'homme et la santé de son âme]*, I, chap. 1, trad. Ch. Appuhn, Paris, Garnier-Flammarion, 1964, p. 44, pour, explicitant le mot "nature", illustrer le principe fondant la première preuve *a priori* de l'existence de Dieu (« Tout ce que nous connaissons clairement et distinctement comme appartenant à la *nature* d'une chose, nous pouvons aussi l'affirmer avec vérité de la chose ») : « Entendez : cette nature déterminée par quoi la chose est ce qu'elle est, et qui ne peut en être en aucune façon séparée, sans que la chose elle-même soit aussitôt anéantie ; c'est ainsi qu'il appartient par exemple à l'essence d'une montagne d'avoir une vallée ou que l'essence d'une montagne consiste en ce qu'elle a une vallée, ce qui est une vérité éternelle et immuable, et doit toujours être dans le concept d'une montagne, même si elle n'a jamais existé ni n'existe. » Ce texte est très proche de l'argument de la V^e *Méditation* affirmant que « de ce que je ne puis concevoir une montagne sans vallée, il ne s'ensuit pas qu'il y ait au monde aucune montagne, ni aucune vallée, mais seulement que la montagne et la vallée, soit qu'il y en ait, soit qu'il n'y en ait point, ne se peuvent en aucune façon séparer l'une d'avec l'autre ; au lieu que, de cela seul que je ne puis concevoir Dieu sans existence, il s'ensuit que l'existence est inséparable de lui, et partant qu'il existe véritablement ».

nature ne saurait au contraire souffrir qu'elle [l'essence] se comporte d'une autre manière que celle dont elle se comporte en réalité :

Quippe distinguendum est inter illa, quæ ex natura sua possunt mutari, ut quod jam scribam vel non scribam, quod aliquis sit prudens, alius imprudens ; et illa, quæ nunquam mutantur, qualia sunt omnia quæ ad alicujus rei essentiam pertinent, ut apud Philosophos est in confesso.	Car il faut bien prendre garde faire distinction entre ces choses qui de leur nature sont susceptibles de changement, comme que j'écrive maintenant ou que je n'écrive pas ; qu'un tel soit prudent, un autre imprudent ; et et celles qui ne changent jamais, comme sont toutes les choses qui appartiennent à l'essence de quelque chose, ainsi que tous les philosophes sont d'accord.
Et quidem non dubium est, quin de contingentibus dici possit, rerum naturam pati, ut illa vel uno, vel alio modo se habeant : exempli causa, ut jam scribam, vel non scribam.	Et de vrai, il n'y a point de doute qu'à l'égard des choses contingentes on peut dire qu'il ne répugne point à la nature des choses qu'elles soient d'une façon ou d'une autre ; par exemple il ne répugne point que j'écrive maintenant, ou que je n'écrive pas.
Sed cum agitur de alicujus rei essentia, plane ineptum est et contradictorium, dicere, rerum naturam pati ut se habeat aliquo alio modo quam revera se habet.	Mais lorsqu'il s'agit de l'essence d'une chose, il est tout à fait absurde, et même il y a de la contradiction, de dire qu'il ne répugne point à la nature des choses qu'elle soit d'une autre façon qu'elle n'est en effet[27].

2° Ne distinguant pas correctement les notions d'attribut et de mode, et ne voyant pas l'homonymie ou l'équivoque qui affecte le mot *attributum*, il croit qu'on peut le prendre en un sens qui convienne univoquement au mode d'une chose, « qui peut être changé », et à l'essence d'une chose, laquelle (essence)

[27] A.-T., VIII-2, p. 347-348 ; Alquié, p. 796.

est au contraire « immuable ». De ce fait, lisant chez Descartes que la pensée et l'étendue sont les attributs principaux des substances en lesquelles elles résident, il ne voit pas que ces attributs-là ne sont pas des modes, mais, dans les deux cas, « une chose qui est immuable, et inséparable de l'essence de son sujet », en tant qu'elle « la constitue, et est, pour cela même, opposée au mode ». Autrement dit, Regius ne saisit pas le coeur même de la doctrine cartésienne, et c'est pour cela qu'il ne voit « aucune contradiction » à ce qu'un *même sujet* puisse avoir pour attributs la pensée et l'étendue. Ce qui donne :

substance corporelle	⇐	étendue pensée
sujet	⇐	attributs (=modes)

La thèse correcte est au contraire qu'il y a *deux sujets pour l'étendue et la pensée* : la substance corporelle et celle qui est le sujet de la pensée, substance-sujet, dont toute la question est de « savoir si elle est corporelle ou incorporelle ».

L'étendue est le sujet de divers modes – comme *être carré* ou *être sphérique* – et elle-même est un attribut de la substance corporelle, attribut (et non pas mode) qui constitue l'essence et (ou) la nature de cette substance [28]. On a donc :

substance corporelle	⇐	étendue		←	être carré
sujet	⇐	attribut	sujet	←	modes

De même, la pensée est le « sujet de divers modes », comme « affirmer, nier, aimer, désirer », le « principe interne d'où ils proviennent », et elle-même est « un attribut qui constitue la nature de quelque substance », dont Regius soutient qu'il se pourrait qu'elle fût corporelle, et dont Descartes soutient qu'elle est nécessairement incorporelle [29]. Tout l'objet du débat est donc de remplir, par la démonstration, la case du sujet de la pensée :

[28]. A.-T., VIII-2, p. 348-349 : « Sic extensio alicujus corporis, modos quidem in se varios potest admittere : nam alius est ejus modus, si corpus istud sit sphæricum, alius, si sit quadratum ; verum ipsa extensio, quæ est modorum illorum subjectum, in se spectata, non est substantiæ corporeæ modus, sed attributum, quod ejus essentiam naturamque constituit. »

[29]. A.-T., VIII-2, p. 349 : « Sic denique cogitationis modi varii sunt : nam affirmare alius est cogitandi modus quam negare, et sic de cæteris ; verum ipsa cogitatio, ut est internum principium, ex quo modi isti exurgunt, et cui insunt, non concipitur ut modus, sed ut attributum, quod constituit naturam alicujus substantiæ, quæ an sit corporea, an vero incorporea, hic quæritur. »

substance <?>	⇐	pensée		←	affirmer désirer
sujet	⇐	attribut	sujet	←	modes

Pour ce faire, il suffit de prouver une seule chose : que la substance qui a pour attribut principal la pensée ne peut avoir d'autre attribut principal – en l'occurrence l'étendue –, autrement dit : que *toute substance a un seul attribut principal* ; car il va de soi que si la substance corporelle a un seul attribut principal, et que cet attribut est l'étendue (ce dont personne ne doute), elle ne pourra être le sujet de la pensée. Le raisonnement peut être conduit ici *a priori.* Il suffit de poser que la substance corporelle a un attribut principal : l'étendue, puis de demander si la pensée peut elle aussi être son attribut principal ; la réponse passe par la solution du problème tantôt évoqué, et qui est au centre de la controverse avec Regius : déterminer si étendue et pensée sont, comme il le soutient, des « attributs qui ne sont pas opposés, mais simplement divers », *non opposita, sed diversa*.

Tout roule donc sur la seconde thèse de Regius (notée ici TR_2), qui est que :

> TR_2 : les attributs pensée et étendue ne sont pas opposés, mais divers

puisque, comme il l'écrit lui-même, s'ils ne sont pas opposés, mais divers :

> TR_3 : rien ne saurait « empêcher que l'esprit ne puisse être un attribut qui convienne à un même sujet que l'étendue, quoique la notion de l'un ne soit pas comprise dans la notion de l'autre » (« nihil obstat quo minus mens possit esse attributum quoddam eidem subjecto cum extensione conveniens, quamvis unum in alterius conceptu non comprehendatur »).

TR_2 comme TR_1 enferme une contradiction. Selon Gueroult, qui reprend ici sa thèse sur la différence entre divers et opposés, Regius ne le voit pas, car il ne comprend pas que les *diversa* « sont infiniment plus que les *opposita* » [30]. Il me

[30]. Dans la *Synopsis* des *Méditations* (A.-T., VII, p. 13, 14) Descartes explique que l'âme et le corps sont, non seulement des *diversa*, mais des choses *quodam modo contraria*, en tant que l'une est indivisible et l'autre divisible. Selon Gueroult le *quodam modo* exprime une restriction : « Il ne

semble plutôt qu'il ne le voit, parce qu'il ne comprend pas que les *opposita* sont infiniment plus que les *diversa*, quand ces *opposita* sont des contradictoires. En tout cas, et à tout le moins, ne distinguant pas l'attribut du mode, il ne voit pas : 1° que, dans le cas d'attributs « qui constituent l'essence de quelques substances », *diversité vaut contradiction*. Dire que l'attribut constitutif de l'essence du corps et l'attribut constitutif de l'essence de l'esprit sont *divers*, c'est dire qu'ils sont *opposés* comme des contradictoires (qui ne peuvent être vrais en même temps). Ou encore, 2° que l'un n'est pas l'autre, ce qui, derechef, signifie qu'ils sont opposés, comme « *être* et *n'être pas* sont opposés » : à savoir comme des contradictoires, entre lesquels il n'y a pas de milieu. On ne saurait donc tirer TR_3 de TR_2 que dans deux cas : (a) si l'on confond attributs essentiels et modes ou (b) si l'on démontre que l'esprit, autrement dit « ce principe interne par lequel nous pensons » est un mode au sens propre du terme. Que l'esprit *soit* un mode, Regius ne le démontre pas. Descartes prouvera, au contraire, que *ce n'en est pas un* en s'appuyant sur « ce qu'il [Regius] dit lui-même dans le cinquième article ». Pour l'heure, il suffit de voir que l'inférence $TR_2 \rightarrow TR_3$ est un paralogisme et d'expliquer *pourquoi.*

On retrouve là ce que j'appelais tantôt, avec M. Gueroult, le problème fondamental de l'anthropologie cartésienne : celui de *l'unité* d'un homme *composé de deux substances hétérogènes.* A ce problème, *L'examen du placard* répond en faisant la synthèse des éléments brassés jusqu'ici dans la discussion de l'article II. Pour ce qui est des « attributs qui constituent la nature des choses » :

---

s'agit pas là, en effet, d'une contrariété entre les substances, mais entre leurs qualités respectives : indivisible – divisible, qui sont des opposés à l'intérieur d'un pseudo-genre : celui du divisible (chez Aristote, par exemple, la puissance est divisible à l'infini, l'acte ne l'est plus) ». Et d'ajouter : « Dans la *VI^e Méditation*, Descartes ajoute que l'indivisibilité est ce par quoi l'âme est *a corpore omnino diversa* (A.-T., VII, p. 86, 13-15), ce par quoi il faut comprendre, non que l'indivisibilité et la divisibilité fondent la diversité de leurs essences, car ce sont seulement des *propria quarto modo* de leurs essences ; mais que, résultant de ce qui en fonde la diversité, elles manifestent de façon particulièrement évidente leur incommensurabilité ». Cette interprétation ne me semble pas entièrement convaincante : qu'est-ce qu'une diversité *absolue*, sinon une diversité *substantielle* (= essentielle, de nature) ?

dici non potest, ea, quæ sunt diversa, et quorum neutrum in alterius conceptu continetur, uni et eidem subjecto convenire ; idem enim est, ac si diceretur, unum et idem subjectum duas habere diversas naturas : quod implicat contradictionem…	on ne peut pas dire que ceux qui sont divers, et qui ne sont en aucune façon compris dans la notion l'un de l'autre, conviennent à un seul et même sujet : car c'est de même que si l'on disait qu'un seul et même sujet a deux natures diverses ; ce qui enferme une manifeste contradiction…

Ce texte contient deux thèses complémentaires (qui ne se peuvent contester, l'une et l'autre se ramenant au *principe de contradiction*) :

> TD_1 : deux attributs essentiels divers ne peuvent avoir le même sujet
>
> TD_2 : deux natures diverses ne peuvent avoir le même sujet

Pensée et étendue étant deux attributs essentiels divers, il en résulte que :

> TD_3 : il n'y a pas de sujet unique de la pensée et de l'étendue

Telle est la première affirmation de Descartes sur « le » sujet. Elle ne prend toutefois son sens anthropologique qu'assortie d'une précision décisive :

subjecto … saltem cum de simplici et non composito quæstio est, quemadmodum hoc in loco.	…au moins lorsqu'il est question, comme ici, d'un sujet simple, et non pas d'un sujet composé.

Il ne faut pas confondre sujet *simple* et sujet *composé*. Sous sa forme développée la thèse cartésienne sur la sub-jecti(vi)té humaine est donc :

> TD_3* : l'homme est un sujet composé de deux substances, l'esprit et le corps, qui sont respectivement les sujets simples, substantiellement distincts, d'attributs essentiels uniques : dans un cas, la pensée, dans l'autre, l'étendue.

Qu'est-ce qu'un être composé ? C'est ce que Descartes explique dans la seconde remarque sur la restriction apportée à TD_3 :

Quippe compositum illud est, in quo reperiuntur duo vel plura attributa, quorum utrumque sine alio potest distincte intelligi : ex hoc enim, quod unum sine alio sic intelligatur, cognoscitur non esse ejus modus, sed res vel attributum rei, quæ potest absque illo subsistere.	Un être est composé dans lequel se rencontrent deux ou plusieurs attributs, chacun desquels peut être conçu distinctement sans l'autre, car de cela même que l'un est ainsi conçu distinctement sans l'autre, on connaît qu'il n'en est pas le mode, mais qu'il est une chose, ou l'attribut d'une chose qui peut subsister sans lui.

Il y a toutes sortes d'êtres composés. Le seul être composé « dans lequel nous considérons l'étendue jointe avec la pensée » est l'homme, « qui est composé de corps et d'âme ». L'erreur de Regius est donc double : la première, qu'il partage en un sens avec le matérialisme hobbésien, est une sorte de réduction métonymique de la nature humaine, qui consiste à prendre l'homme « seulement pour le corps, dont l'esprit est un mode » [31] ; la seconde lui est propre : c'est TR_1, qui consiste à prétendre qu'« il n'y avait point de *contradiction* qu'une seule et même chose eût l'une ou l'autre de deux natures entièrement diverses », à savoir « ou une substance ou un mode ». On peut excuser l'ignorance d'un adversaire [32] ; on peut blâmer son arrogance [33]. En « disant des choses qui se *contredisent* », Regius n'a « fait paraître que l'absurdité de son esprit » [34].

[31]. A.-T. VIII-1, p. 351 : « Unde patet, illud subjectum, in quo solam extensionem cum variis extensionis modis intelligimus, esse ens simplex : ut etiam subjectum, in quo solam cogitationem cum variis cogitationum modis agnoscimus. Illud autem, in quo extensionem et cogitationem simul consideramus, esse compositum : hominem scilicet, constantem anima et corpore, quem videtur author noster pro solo corpore, cujus mens sit modus, hic sumpsisse. »

[32]. Ce que Descartes eût volontiers fait si Regius avait « seulement dit qu'il ne voyait point de raison pourquoi l'esprit humain dût plutôt être estimé une substance incorporelle qu'un mode de la substance corporelle ».

[33]. Ce qu'il eût également fait s'il s'était contenté de dire qu'il « n'est pas possible à la raison humaine de trouver jamais aucune preuve par laquelle on puisse démontrer que l'esprit humain soit l'un plutôt que l'autre ».

[34]. « Si tantum dixisset, nullas se percipere rationes, propter quas mens humana credi debeat substantia incorporea potius quam substantiæ corporeæ modus, posset ejus ignorantia excusari ; si vero dixisset, nullas ab humano ingenio posse inveniri rationes, quibus unum potius quam aliud probetur, arrogantia quidem esset culpanda, sed non appareret contradictio in ejus verbis, cum autem dicit, rerum naturam pati ; ut idem sit substantia, vel modus, omnino pugnantia loquitur, et absurditatem ingenii sui ostendit. »

*

Que reste-t-il du « sujet cartésien » au terme de ce parcours ? Une victoire du patinage français aux « figures imposées », remportée avec la moins cartésienne de toutes : « la » *subiectum*, contre un champion britannique et une étoile batave. Mais aussi, et surtout, une thèse anthropologique forte, qui ne doit rien au magistère scolaire du *cogito*, et guère à la scénarisation post-schellingienne du *Ich-Satz* chez Heidegger : il y a deux sujets simples en l'homme, la pensée et l'étendue. L'homme est un sujet composé de ces deux sujets simples. L'homme n'est ni esprit seulement, ni corps seulement, mais esprit et corps. En d'autres mots : les seuls textes de Descartes où il soit question de sujet ne permettent pas de poser *directement* que c'est chez lui que « la *mens humana*, revendique exclusivement pour elle le nom de sujet de telle sorte que *subiectum* et *ego*, subjectivité et égoïté (*Ichheit*) acquièrent une signification identique » [35]. Il faudrait pour cela une réduction métonymique – et platonicienne – de la nature humaine, antithétique de celle de Regius, *identifiant l'homme* non plus au corps, mais à la *mens*, définie comme sujet unique de la pensée (voire, pis encore, comme sujet unique de la pensée et…de l'étendue). Ce serait une erreur. Il vaudrait mieux, reprenant la question d'Aristote **tiv to ; o[n**, et sa reformulation brunschwigienne (platonicienne) **ti ; ejstiv to ; o[n**, poser à Descartes deux questions : Q1 : Qu'est-ce que l'homme ? et Q2 : Qu'est-ce qui est je (moi) ? Qu'est-ce que c'est qui est je (moi) ? On ne doute pas qu'il répondrait à la première : un être – voire un sujet – composé de deux substances. On peut facilement imaginer qu'il répondrait quelque chose à la seconde, puisqu'il l'a posée lui-même, et l'on ne doute pas non plus de sa réponse, puisqu'elle figure en toute lettre dans les pages les plus commentées de la *Deuxième* et de la *Troisième Méditation.*

> Mais qu'est-ce donc que je suis ? Une chose qui pense. Qu'est-ce qu'une chose qui pense ? C'est-à-dire une chose qui doute, qui conçoit, qui affirme, qui nie, qui veut, qui ne veut pas, qui imagine aussi, et qui sent [36].
>
> Je suis une chose qui pense, c'est-à-dire qui doute, qui affirme, qui nie, qui connaît peu de choses, qui en ignore beaucoup, qui aime, qui hait, qui veut, qui ne veut pas, qui imagine aussi, et qui sent [37].

[35]. Ga 6.2, p. 395, trad. cit., p. 348.
[36]. *Méditations*, II, A.-T., IX-1, p. 22.
[37]. *Méditations*, III, A.-T., IX-1, p. 27.

Resterait à savoir ce qu'*est* cette chose, et en quoi elle garantit, explique ou fonde l'*unité* de l'homme. La question, en ces termes, est pendante : *retour amont* ou *retour aval*, la Querelle d'Utrecht n'est pas terminée.

PART FOUR

Norms and Values

38

Value uncertainty and value instability in decision-making *

GÖRAN HERMERÉN, INGAR BRINCK, JOHANNES PERSSON AND NILS-ERIC SAHLIN

Abstract The purpose of this paper is to clarify the role of value uncertainty and value instability in decision-making that concerns morally controversial issues. Value uncertainty and value instability are distinguished from moral uncertainty, and several types of value uncertainty and value instability are defined and discussed. The relations between value uncertainty and value instability are explored, and value uncertainty is illustrated with examples drawn from the social sciences, medicine and everyday life. Several types of factor producing value uncertainty and/or value instability are then identified. They are grouped into three categories and discussed under the headings 'value framing', 'ambivalence' and 'lack of self-knowledge'. The paper then discusses the role of value uncertainty in decision-making. The concluding remarks summarize what has been achieved and what remains to be done in this area.

Keywords value uncertainty, value instability, decision-making, epistemic indeterminacy, ethics

*The authors wish to thank Paul A. Robinson for valuable help and constructive comments.

1. Introduction

In making decisions on ethically controversial issues it is useful to start with three questions —

1. What do we know?
2. What do we want?
3. What can we do?

— before trying to answer a crucial fourth:

4 What should we do?

Obviously, answers to the first three questions do not settle the fourth, but they provide evidence that needs to be taken into account. Moreover, the set of questions shows that there are interesting parallels between the problems raised by belief-issues and those arising from value-issues. Since Pascal Engel has highlighted the former in an original way in his research on beliefs and epistemic norms, we hope that this attempt to focus on the latter will be of interest to him.[1]

The purpose of this paper is to clarify the role of value uncertainty and value instability in decision-making. When we have introduced definitions of value uncertainty and value instability, and having made some crucial distinctions, we will present a number of illustrative examples drawn from ordinary life, medicine and the social sciences.

As is well known, conclusions or decisions about morally controversial issues are based on premises of several kinds, and an understanding of value uncertainty, as well as of epistemic uncertainty and indeterminacy, is an important asset in the decision-maker. The analysis of these uncertainties will have important ramifications for questions about how to deal with the ethical issues raised by, among other things, new and emerging technologies. In our view, this approach involves a new starting point in ethical analysis.

We cannot take it for granted that all value uncertainty is based on – or is in some other way related to – epistemic uncertainty and indeterminacy. The possibility that two people can agree on all known facts but be genuinely uncertain about their goals and values must be taken seriously. Whereas goals

[1] See, for example, Jérôme Dokic and Pascal Engel, *Frank Ramsey: Truth and Success* (London: Routledge, 2006); Pascal Engel, "The disunity of reason", XVII Congresso Interamericano de Filosofia, Salvador Oct 2013; Nils-Eric Sahlin, *The Philosophy of F. P. Ramsey* (Cambridge: Cambridge University Press, 1990).

pull you in different directions, facts may push you around. Very early on in the discussion of this topic, Levi called attention to the important role of epistemic and value indeterminacy in decision-making.[2] We want to continue to explore these phenomena, and to relate them to earlier discussions of epistemic risk.[3]

The following questions will be discussed: (i) What does 'value uncertainty' mean? (ii) Does it come in different types? Do we, for example, have to make a clear distinction between value uncertainty and value instability? (iii) What factors contribute to, or help bring about, uncertainty? What causes instability? (iv) What role does value uncertainty play in decision-making?

The paper also makes a contribution to the phenomenology of moral conflict, on which there is a growing body of literature.[4] Situations in which each alternative open to the agent "has a high moral cost" (Morris, p. 224) will naturally give rise to uncertainty. We claim that the experience of uncertainty – both epistemological and axiological – is an important aspect of moral engagement generally. In other words, the situation in which there is certainty about the alternative courses of action available, and the probability of their various outcomes, as well as about values, is a special case. Other possible combinations of epistemic (in)determinacy and value (un)certainty also need to be examined.

2. A tentative definition and some distinctions

We begin by proposing the following tentative definition. This pinpoints one type of value indeterminacy. We then seek to gauge whether, and to what extent, it fits the examples we subsequently present. Later, we will introduce another type of indeterminacy.

> *To say that a person, A, is uncertain at time, t, about which values should be the basis for a decision is to say that A has at t a set of val-*

[2] Isaac Levi, *The Enterprise of Knowledge. An Essay on Knowledge, Credal Probability, and Chance* (Cambridge, Mass.: MIT Press, 1980); Isaac Levi, *The Covenant of Reason: Rationality and the Commitments of Thought* (Cambridge: Cambridge University Press, 1997).

[3] See Nils-Eric Sahlin and Johannes Persson, "Epistemic risk: The significance of knowing what one does not know", in B. Brehmer and N.-E. Sahlin, eds., *Future Risks and Risk Management*, (Boston: Kluwer, 1994), pp. 37-62; Peter Gärdenfors and N-E Sahlin, "Unreliable probabilities, risk taking, and decision making", *Synthese*, Vol. 53 (1982): 361-386.

[4] See e.g. Michael K. Morris, "Moral conflict and ordinary emotional experience", *The Journal of Value Inquiry*, Vol. 26 (1992): 223-257; and the literature he refers to.

ues $< a, b, c >$ *and A cannot decide at t (1) how* a, b, c*, are to be ranked and/or (2) what the value distances between* a, b, c *are.*

The type-token distinction applies to values. It is one thing to discuss specific outcomes of preferences, where objects or events are ranked. It is quite another to discuss general types of value which, when used by people in specific situations to make decisions, yield different outcomes.

As to the latter case, consider P's contemplation of abortion. Which value is most important to P in this situation? Justice or the right of women to decide for themselves? The latter, probably – at least in non-Catholic countries, probably the latter. Suppose instead that P is considering the allocation of resources in healthcare. Should the limited resources be spent on uterus transplantation or malaria? Which value is most important in this situation? Justice or the right of the stakeholders to decide for themselves? Probably the former. But there are bound to be murkier situations where there is uncertainty about what is to be preferred, and this uncertainty will be mirrored in differing value orderings with non-identical outcomes.

Clause (1) in the definition above, referring to doubts about how the values are to be ranked, is usually is neglected, but it is very important. The reason is simply that many, if not most, values are recognized in all cultures, but at the same time cultures diverge over their relative importance.

The World Values Survey provides empirical data on differences in values, and on the ways in which values are ranked in different cultures, although it must be said that the values in this project are compared in very few dimensions. Interestingly, Sweden and Zimbabwe occupy very different positions on the world values map, which summarizes some of the findings of this project.[5]

Within Europe as well it is easy to find examples of values which are widely accepted but ranked differently: examples include animal health and welfare, the sanctity of unborn human life, the protection of privacy and integrity (and more generally the rights of individuals in the face of societal interests), and the place and role of women in society.[6]

The concept of value *uncertainty* describes cases where there are doubts about the two dimensions of evaluation we have identified (ranking, and the value distance) at a time. But even if these dimensions are settled at t, they may not remain so: the ranking, or the distances between the values, may

[5] The World Values Survey, see `http://www.worldvaluessurvey.org`

[6] Göran Hermerén, "European Values – and Others. Europe's Shared Values: Towards an Ever-closer Union?", *European Review*, Vol. 16. No. 3 (2008): 373-385.

change between t and a later time t′. Where this happens we propose to talk about value *instability*.

Value uncertainty and value instability Let us, then, distinguish the synchronic notion of value uncertainty from the diachronic notion of value instability. The latter arises where: (i) a person P is either certain or uncertain about his or her values at a given time t (P may, in other words, be in the grip either of value certainty or of value uncertainty); (ii) between t and a later time t′ P changes his or her opinion about (1) the ranking of *a,b,c* and/or (2) the value distances between *a,b,c*; and finally, (iii) when and how the values will change between t and t′ is difficult for P to predict at t.

By eliminating the last clause (i.e. (iii)) about predictive difficulty, it is possible to obtain a simpler theoretical notion. Let us call this *simple value instability*. Where (iii) is included, we shall speak of *complex value instability.* It is the simple concept we have in mind when we refer to value instability without adding a qualifying adjective.

The difference between value uncertainty and value instability is a difference between absolute indeterminacy and underdetermination (as in 'not being fixed'). In the first case there is a genuine uncertainty about the ordering and distancing of the values of a sort that cannot be settled; in the second, the ordering and distance are stable at any given point in time, *but they are not determined/fixed*, and as a result of being underdetermined by the known facts they change from one time to another.

New experiences can help to create value instability, both simple and complex.

Value instability may occur in cases of both epistemic indeterminacy and determinacy. In the latter case we might be inclined to say that the decision-maker has changed his or her values. In the former, we would probably say the decision-maker does not know enough to know what he or she values. Changes of values over time – that is to say, value instability – can occur both ordinally and cardinally.

All combinations of value uncertainty and value instability are possible, and this suggests that value uncertainty and value instability are, in the sense given by this combinatorial potential, independent.

There are cases where value instability seems to presuppose value uncertainty. Whereas value uncertainty can occur both in the individual case and

in connection with a given value at large, generally and across contexts, value instability occurs only concerning values at large, or generally.

If a person is uncertain about the ordering and distancing of certain values at a given point in time, then, obviously, changes will be hard to predict. However, this difficulty is not philosophically interesting. Below we will argue that what matters is the kind of case in which value instability occurs in spite of a limited value certainty.

Now consider an example of what Thomas Magnell has called "collapsing goods".[7] His starting point is recent reports on antibiotic resistance, where we seem to be confronted with the choice between what is good in the short term for particular individuals and what may be good in a longer time perspective for the population at large. Antibiotic resistant bacteria may become a dangerous threat, but depending on the particular situation – the patient's condition, the doctor's relationship with the patient, and so forth – the doctor may be genuinely uncertain about how his or her values are to be ranked in a situation like this. The point is that this uncertainty need not signal uncertainty about the consequences of administering antibiotics.

The example can be developed to illustrate both value instability and value uncertainty. Suppose the values are clear in the sense that there is no value uncertainty in a particular scenario at a particular point in time, but that it is difficult to predict what the values will be at future times and in potentially developing scenarios. This is a case of value instability. In a situation (time, scenario) with uncertainty about values in the sense indicated by the first definition, by contrast, there is genuine value uncertainty.

Value instability concerns the comparison of values at different points of time. It can arise from many different factors, including events occurring during the decision-making process, and it is observable in everyday settings. For instance, suppose that you need to eat, take a shower, and be in time for a meeting. But it is not possible to do all these things. The situation can be complicated by difficulties predicting all manner of things – the mood of the chairman of the meeting if you are late, the traffic, your ability to work well at the meeting without breakfast, and so on. The wisdom of the decisions you make can also be influenced by misfortunes at home or while driving.

However, everyday life often involves a daily routine, and this means that changes over time in an agent's ordering of a given set of values can often be predicted by others with a rather high degree of probability. This has a

[7] Thomas Magnell, "Collapsing Goods in Medicine and the Value of Innovation", *The Journal of Value Inquiry*, Vol. 40, (2006): 155-168.

bearing on some of the conceptions of value instability discussed above, and especially those involving (iii)* and (iii)**.

It may seem that value instability has limited importance because the implementation of values always requires adjustment to circumstance and the problems caused by value instability are therefore negligible. However, in many contexts value instability is a genuine problem. Decisions in economics and politics, at national and international levels, and the implementation of policies and provisions on healthcare, education and the environment, can all be complicated by such instability. In these arenas values are expected to remain relatively stable over time and across similar contexts. In fact, decision-making in politics relies on this expectation. Otherwise, it seems, there would not be any point in introducing policies and rules of conduct.

Consider another type of example, from everyday life: parenting. Mothers and fathers are expected to stick to the same values over time, so as not to confuse their children. They nevertheless sometimes neglect values they genuinely cherish. A bad day at work, a bad night's sleep, or a death in the family may lead the best of parents to apply a simpler strategy than the expected one to get some peace and quiet – to allow the children stay up and watch television too late, and eat too many sweets, or play computer games to see out the day. In hard times, the neglect of family values may become the rule rather than the exception, something that makes it difficult for the children to understand what the family's values really are, or what rules they are supposed to follow.

These examples show that value instability can damage long-term strategies for implementing values. We submit that the problems of value instability and uncertainty need, therefore, to be considered, particularly by regulators. Value instability is a nuisance, especially, for those making new decisions – decisions that are meant to be stable over time, of course – on regulations and policies designed to direct future behaviour. At least, this is the case when we focus on the values of individuals. For then uncertainties are likely to intensify the difficulties of collective decision-making connected with problems raised by the proportionality principle.[8]

Moral and value uncertainty Uncertainty about values of the sort defined above should be separated from uncertainty about which ethical principles or

[8] Göran Hermerén, "Principle of Proportionality Revisited", *Medicine, Health Care and Philosophy*. Published on line Nov 1, 2011.

framework to apply (utilitarian, human rights, human dignity, virtue ethics, and so on). This other kind of uncertainty has been discussed extensively in the ethics literature in connection with copious concrete examples, among them abortion.[9] Following Ted Lockhart, we shall refer to it as moral uncertainty.

The main difference between moral and value uncertainty seems to be this: values are not identical with normative theories, although the latter do promote certain values. A person who is not a utilitarian, or not a Kantian, or indeed somebody who has never heard of these theories, can certainly exhibit value uncertainty in the sense defined above. Thus, there can be value uncertainty without moral uncertainty, but usually not moral uncertainty without some degree of value uncertainty – although sometimes two moral philosophies can be used equally successfully to support a single moral recommendation on how to act.

It may be argued that the position of Lockhart, and the distinction above between value uncertainty and moral uncertainty, ignores particularism in ethics. Particularist positions have been suggested but also criticized by many during the last few decades, in this journal among other places.[10] On a particularist position the differences between the two concepts will, at least, be less clear than is suggested above.

Some illustrations of value uncertainty The definitions of value uncertainty and value instability provided above can be illustrated with concrete examples like these: what do you prefer – tea, coffee or wine? Always in that order? What if the distance between tea and coffee is minimal, whereas the distance between coffee and wine is huge? We have a case of value uncertainty if the preferences of the decision-maker are unclear in the sense either that the ordering of the outcomes is uncertain, unclear or imprecise, and/or that the value distances between them are indeterminate or fluctuating.

Again, consider the following example. A particular Mac is twice as good as a PC, but the Mac is three times the price of a PC. Is the difference in quality worth the difference in price (quantity)? How is this to be analysed? The issue can be understood as an example of value uncertainty. The problem is that the

[9] See Ted Lockhart, *Moral Uncertainty* (New York: Oxford University Press), 2000.

[10] For instance, Jörg Schroth, "Particularism and Universalizability", *The Journal of Value Inquiry*, Vol. 37, No. 4 (2003): pp. 455-461; Maike Albertzart, "Missing the Target: Jonathan Dancy's Conception of a Principled Ethics", *The Journal of Value Inquiry*, Vol. 45, No. 1, (2011): 49-58.

decision maker is weighing up a number of aspects which are very difficult to compare in a non-arbitrary way.

The tentative umbrella definitions proposed above have to be tested against further examples. They may have to be refined later. However, the examples suggest that both value uncertainty and value instability play a significant role in decision-making, that there are several sub-varieties of the two phenomena, and that these are capable of being combined in various ways.

In what follows, we propose to focus on value uncertainty.

3. Examples

In this section examples will be presented and then used to illustrate various ways in which value uncertainty and instability are relevant in decision-making. We have tried to select examples of contrasting types which raise interesting issues, exemplifying (ideally different) types of value uncertainty and value instability.

Example 1 Closing down the nuclear power plant Controversies over the closing down or building of a new nuclear power plant may exhibit value uncertainty within and between stakeholders and decision-makers. The values at stake here include trust, safety, efficiency, industrial competitiveness, cheap energy, and so on. These can be interpreted in more ways than one and ranked differently, in different scenarios, at a single point in time. The result is value uncertainty.

The example illustrates a common feature of value uncertainty. Often it is taken for granted that the values are well-defined, and that uncertainty concerns how the values are to be ranked or the distance between the values. In this example the terms referring to the key values are all vague, and different interpretations yielding different results are possible. This means that uncertainty also concerns how the key terms are to be interpreted and made more precise – in other words, which particular notions of trust, efficiency, and so forth are to be assessed and ranked. This vagueness introduces new types of value uncertainty.

Example 2 Neonatal care In neonatal intensive care controversies over how to treat prematurely born infants are not uncommon. How active should the

treatment be? When should maximal intensive care resources be used and when not? Such controversies illustrate conflicts between well-established values and uncertainties as to how these conflicts should be resolved. The values at stake include saving life, minimizing harm to the child, and optimizing the parents' quality of life. These are capable of being interpreted in more ways than one and ranked differently in various scenarios – what do patients, parents, the wider family, nurses and doctors believe and want? What relative weights should be given to the interests of these stakeholders? What role should societal interest in containing healthcare costs play?

Example 3 Social influence on individual behaviour Some types of value instability, illustrated by experiences from ordinary social life, involve friends and co-workers. In these cases, peers are very influential. Anyone desirous of acceptance by a certain group who wants to achieve as high a status as possible is likely to follow the norms of the group. Very probably, a hopeful friend of this sort will interpret values in line with the scenarios considered acceptable within the group. In the case of co-workers, similar mechanisms seem to be likely to be at work. This is bound to result in value instability, depending on the social influences impacting on the agent's choice of scenarios for each separate decision-procedure (at different times).

Example 4 The Swedish politician Consider, moreover, the following problem – one facing more than one Swedish politician today. In view of the historical connection between labour unions and the social democratic party in Sweden, Swedish social democratic politicians may have several utility functions in the back of their mind: to promote the interests of their party, to act in the interest of the labour unions, and to further their own political careers. These utility functions are not identical. They can also be made more precise. But how should they be ordered?

In different scenarios, this can be interpreted as an example of value uncertainty or as an example of value instability.

Example 5 The suffering child Finally, Anders Castor and Nils-Eric Sahlin ask us to imagine a three year-old child with high-risk neuroblastoma.[11] Her clinical care has followed the standard paediatric protocols: she has been given haematopoietic stem cell transplantation, for example – a treatment not without complications, and one that can involve considerable suffering for the patient. But not all children respond to stem cell treatment, and let us assume this is a case of recurring neuroblastoma. The question is whether to continue treatment or not. How should we decide? Which values, and whose values, should be decisive? In whose interest would it be to continue? To stop the treatment?

This last example will be used to illustrate some important factors contributing to value uncertainty. But interestingly enough, this uncertainty can be paired with value instability. That happens when the parents change their views and reverse earlier decisions

4. Important factors: an overview and classification

In this section we will identify factors tending to produce value uncertainty and instability, and we shall illustrate them using the example chosen above. A complete list of such factors cannot be given. In each particular situation, partly different factors can play a role. Many factors can also be combined. Three main types of factor are distinguishable:

> *A. Value framing,* highlighting the role of external factors in the situation at hand.
>
> *B. Ambivalence,* highlighting ambiguous features either of the case or of the situation itself.
>
> *C. Lack of self-knowledge,* highlighting internal factors, including psychological characteristics of the decision-makers.

Each of these factors can in their turn be divided into subcategories.

[11] Anders Castor and Nils-Eric Sahlin, "Mycket svåra beslut", in Johannes Persson and Nils-Eric Sahlin, eds., *Risk & Risici*, (Nora: Nya Doxa, 2008), pp. 232-248. See also Nils-Eric Sahlin, Johannes Persson, Niklas Vareman, "Unruhe und Ungewissheit – Stem Cells and Risks". in K. Hug and G. Hermerén, eds., *Translational Stem Cell Research: Issues Beyond the Debate on the Moral Status of the Embryo* (New York: Springer/Humana Press, 2010), pp. 421-429 and N-E Sahlin, "Kunskapsluckor och riskhantering", in Göran Stålbom och Birgitta Johansson, eds., *Människan inomhus: Perspektiv på vår tids inneliv*, (Stockholm: Formas 2003), pp. 307-26.

A. Value framing When decisions are made, and values and information about intentions and consequences are taken into account, this never takes place in a vacuum. There is always a context, a background, and several anticipated future scenarios, and these provide the frame of the decision. Depending on how this frame is specified, it may or may not generate value uncertainty and instability.

Let us return to the case of the child with high-risk neuroblastoma. Which values are decisive in this case? Uncertainty about this may reflect epistemic uncertainty, although it does not have to do so. What are the odds that a haematopoietic stem cell transplantation that failed to succeed the first time will work at the second, or the third, or fourth attempt? With the number of refractory episodes, the relevant probabilities become harder and harder, if not impossible, to estimate. As our epistemic status deteriorates, our preferences, desires and values become more and more uncertain, and this value uncertainty may induce value instability. What people see as the 'right' decision may change from day to day, and sometimes even from one hour to the next, although nothing in the situation, or in their information about the situation, has changed.

In this context it is the deterioration of our values that is of particular interest. With each refractory episode it becomes harder to frame the relevant values, both for the patients, the parents and the physicians. There is no familiar structure to rely on – no comforting, mundane fabric of values. Uncertainty about the value structure (the ranking of the values, and/or the distances between them) is a realistic possibility.

This example combines maximal epistemic and value uncertainty (the combinations will be discussed in more detail later). The following reflections on state-dependent preferences/values therefore seem apt. Mark Schervish and his colleagues have shown that all of the classical theories of decision-making have a problem with state-dependent utilities.[12] Theories using (horse) lotteries and prizes to derive probabilities cannot guarantee the existence of unique probabilities. The problem is that the utility of a prize is the utility of that prize *given* a particular state of nature. And even 'constant' prizes might have a different value in different states of nature, which means that the subject's preferences can be represented by far too many utility functions. (Underdetermination may play an important role here.) As a consequence there is no unique subjective probability distribution over states of nature. In the present

[12] Schervish, M., Seidenfeld J., and Kadane, T., "State-dependent utilities", *Journal of the American Statistical Association*, Vol. 85, No. 411 (1990): 840-7.

context, this may seem to be an unnecessary theoretical detour, a superfluous ornament, but it is not. The result tells us something important about the relation between values (and, indirectly, uncertain values) and partial beliefs (expressed as subjective probabilities).

Here we need to distinguish different kinds of situation. On the one hand, we have the ideal decision-maker, who is thoroughly rational and has complete information about all relevant aspects of the situation. On the other, we have various kinds of deviation from this ideal. In the latter cases, the actual situation of the decision-makers and stakeholders – the parents, the doctors, the child – may play a very important role. Do the parents have other children? How old are they? What are the chances that they can conceive another child? What is the previous experience of the doctors? What happens to be the social and political situation in the country where they live (economic recession, war or peace)? Such factors can have an impact on value certainty.

There are several subcategories of value framing. For instance, we can distinguish between value-loaded and selective descriptions. It is possible to be misled by value-laden words referring to values. If the terms used are positively value-loaded, a person may be more inclined to accept the values referred to than he or she would be if they were described less optimistically. Uncertainty can be created by exploiting, or not seeing through, this mechanism.

Certain cases can be interpreted as demonstrating either value uncertainty or value instability, depending on how they are understood. This makes them especially interesting. Consider, for instance, the slogan: 'Yes to Life'. Who would be against life? But those who embrace 'Yes to Life' may not always realize what it means for women's health and quality of life, or for the quality of life of prematurely born children with grave, multiple handicap, or for society as a whole.

This case can be interpreted as one in which the expression 'Yes to life' is simply vague. Those who assert it do not really know what they mean, or are saying, when they do so, with the result that the phrase has different meanings in different scenarios within the same context or at the same time. Analysing the value indicated by the phrase relative to distinct scenarios and a single time/context (one and the same evaluation process) is, it seems, a matter of conceptual clarification. This interpretation, then, represents the case as one of value uncertainty, as a case illustrating the way lack of clarity can give rise to value uncertainty.

However, it seems the case can equally well be interpreted as saying that the meaning of the phrase 'Yes to life' cannot be determined because no fact

of the matter tells us what it means to say yes to life, or how this might be settled. When this is the reason for interpreting the slogan differently in distinct contexts, and so making different decisions about the same issue, we have a case of value instability.

B. Ambivalence Returning again to the three year-old with high-risk neuroblastoma, the serious doubts that this type of situation triggers may lead to value ambivalence: "Are we prepared to put our child through all the suffering once again? Is death after all a better option?" Value uncertainty can trigger ambivalence, but it can also be triggered by it.

Within the phenomenon of ambivalence we can distinguish two kinds of case. In one something can be 'viewed' in more than one way: examples are Wittgenstein's duck-rabbit and (in a rather different way) a jungle location where nothing grows which is regarded as an indication of divine intervention or the result of a chemical accident. The other kind of case is that of ambivalence experiences and insights.

In the first case ambivalence is related to aspect-seeing. In the second case, the experience can be interpreted in more than one way – for example, as the result of haste, or incompetence, or maliciousness. It is possible to realize something without experiencing it, and to experience something without realizing it, so these aspects may need to be separated.

It is not difficult to imagine that certain information given by test results in the case of the three year-old child with high-risk neuroblastoma could be ambiguous, or that the information provided by doctors can be put in different perspectives, making the picture they provide ambiguous in the Wittgensteinian duck-rabbit sense.

C. Lack of self-knowledge (and knowledge of others) Lack of self-knowledge can be exemplified in various ways. It raises questions about the extent to which decision-makers track their own mental states, can influence them (the so-called 'weakness of the will' problem), and can predict their own behaviour and preferences. Lack of self-knowledge may well deepen value uncertainty – something that is also illustrated by the case of the child with high-risk neuroblastoma. But insensitivity to the feelings of other people and lack of first-hand experience can also play an important role.

D. Insensitivity Insensitivity to the suffering of others can be due to limited personal experience. It may be attributable to selective description, however. It may be that when others' sufferings are described in one way, many people are inclined to order their values thus and so, but that when the sufferings are described differently that ordering comes to seem incorrect. Perhaps, when we do not see the face of the other, to use an expression made famous by Emmanuel Levinas, our imagination fails to show us the implications of the way in which we order our values.[13]

Limitations experience can also affect our appreciation of material features of a situation (e.g. economic, socio-cultural, and technological features). Ingar Brinck explains the negative influence of this lack of first-hand experience on decision-making in foreign aid, when ill-judged decisions are made as to how one country should contribute to the improvement of another's agriculture or environmental sustainability.[14]

Suppose you have not seen the suffering and dying people in Darfur: they are just abstract numbers and not quite real. At a general level, all humans are equal, but you know about the fate of these people only via brief notes in the media. Here the tension between abstract principles and lack of first-hand knowledge can help to create value uncertainty. Of course, this is different from having seen the suffering people without being moved to act: a person totally lacking in empathy could not care less.

What do I want to achieve? To avoid? How are my preferences ordered? An individual is not always able to see through his own motives and know what he really wants. 'Know thyself' said the Greeks, and this maxim is as valid today as it was then. Freudian defence mechanisms, active forgetfulness, or bad memory can distort a person's picture of what he or she desires.

Once again, the case of the child with high-risk neuroblastoma is illustrative. To what extent do the parents grasp their own deeper motivations? And what do the doctors know about their own values? The case raises issues of intersubjectivity and our knowledge of other minds. What do doctors know about the needs and wants of the parents? What do parents and doctors know about the preferences of the child? These issues cannot be discussed here, however. They deserve separate treatment.

[13] For an analysis of different kinds of intersubjectivity, and the way these influence our capacity for nonverbal and verbal communication, see Ingar Brinck, "The role of intersubjectivity for the development of intentional communication." In J. Zlatev, T. Racine. C. Sinha, & E. Itkonen, eds., *The Shared Mind: Perspectives on Intersubjectivity* (Amsterdam: John Benjamins Publ. 2008).

[14] Ingar Brinck, "Om riskkommunikation: kartor, klyftor och mål". In I. Brinck, S. Halldén, A.-S. Maurin, & J. Persson, eds., *Risk och det levande mänskliga*, (Nora: Nya Doxa, 2005), pp. 45-78.

What can be done to counter an inability to live in accordance with one's deeper wishes? Sören Halldén discusses what can be done in practice, using a stout man who wants to become slim as an example.[15] Someone with this aim may turn to a psychotherapist for help, expecting to receive moral advice. But as a rule the psychotherapist will decline; his job is to help the person find himself. His task is to be, as Halldén puts it, "a midwife in the moral field". The name of Socrates comes to mind, as do the names of a number of psychologists in the Freudian tradition, like Erich Fromm and Karen Horney.

5. Utilities and value uncertainty: further analysis

Let us now move on and consider preferences, utilities and value uncertainty at the level of the individual. Suppose you prefer vegetarian pizza to beef-steak. But do you prefer fermented Baltic herring to pizza? Or do you prefer steak to herring? Are you indifferent? Do you prefer a completely new type of stem-cell transplantation based on iPS-cells to fermented Baltic herring?

The traditional theory of conjoint measurement assumes a weak ordering of our preferences, i.e. transitivity and totality. If our preferences are weakly ordered and fulfil some other axioms, such as cancellation, it is possible to prove that they can be represented by a utility (or value) function determined up to a positive affine transformation.

Totality means all options are comparable. However, it seems it would be hard to compare something we have not, or have almost never, experienced (how many times have you had fermented Baltic herring?) with something we experience every day, such as a cup of coffee or a mug of tea. Unclear preferences induce value uncertainty – and not in the trivial way implying that we only have utilities (values) determined up to an affine transformation, but in a more serious way: we cannot say whether A is preferred to B or the other way around.

When things are difficult to compare we might have to work with sets of preference orderings. But what do we do when we cannot compare the options? Can a cup of coffee be compared to a stem-cell transplantation? People tend to have very different intuitions here, depending on their circumstances. For parents with a child needing stem cell transplantation, the choice is simple; for others, the very idea that the options are capable of comparison may be alien.

[15] Sören Halldén, *A Socratic Approach to Morality* (Lund: Library of Theoria, vol, 20, 1995), pp. 109 ff.

Is there any total ordering to be found? In this case we seem to have a form of value uncertainty induced, not by lack of experience, but by something slightly more fundamental.

Again, let us assume you prefer coffee to tea, but also prefer a bad red wine to castor oil or cicutoxin. Is the value difference between the first two options greater or less that that between the second two? Tea is almost as good as coffee. The distance between them is almost negligible. But how big is the value distance between bad red wine and castor oil? It depends on how bad the wine is and how keen we are to avoid nausea and emesis.

Are wine and cicutoxin really comparable? They are clearly located on very different scales and in very different contexts: there is quite a contrast between the choice to live or die and the choice to drink tea or coffee. Or we could clarify the situation by specifying the preferences in more detail: I prefer castor oil to read wine for lubrication, even if the wine is bad; but I prefer bad red wine to cicutoxin as a drink.

We can, in other words, feel uncertain about our preferences, but also about the value distances between them. And this second type of doubt too induces value instability. It has been argued that higher-order preferences reveal value distances. If it is better to prefer bad red wine to castor oil than to prefer coffee to tea, the value difference between the first two options is the greater one. This truism means that uncertain second-order preferences induce value instability – unstable value distances.

6. Value uncertainty in decision-making

In this last section we consider the role specifically of value uncertainty in decision-making. Clearly, in each particular case, when there is value uncertainty, we need to investigate the factors producing it – and see what could be done in practice to eliminate, reduce or circumvent it.

The impact of value uncertainty in the decision-process is also modulated by the assumed value of value certainty. Certainty about this value will simplify the process and enable traditional theories of rationality to be applied, which obviously may be a good thing. But the tacit assumption that value certainty is always intrinsically or instrumentally good is neither self-evidently correct nor without danger.

There are situations in which value uncertainty can improve the quality of the decision-making. It can help us avoid pitfalls or serious mistakes. To return to the examples given in Section 3 above, if there is value certainty in

controversies over the closing down of a nuclear power plant or the active treatment of prematurely born infants, serious mistakes can be made which will later be regretted by the stakeholders. In genuinely difficult situations, to proceed as if all stakeholders were certain about their values may be to present a false picture.

Let us pursue this. If the value uncertainty is due to a lack of familiarity with the outcomes of our choices, the obvious strategy is to carry out more research. But if it is arises as a result of our being deceived by the nearness of certain outcomes, and because we are insufficiently sensitive to more distant ones (e.g. events in Africa), we may need to go to the relevant places, read books, study movies, and see for ourselves how and why people are starving and dying in camps. Again, if the value uncertainty is attributable to inconsistent value premises, the inconsistencies have to be made explicit. The agent will have to decide which values are more important.

7. Concluding Remarks

What have we achieved here, and what remains to be done? Two tentative definitions of value uncertainty and value instability have been introduced and tested against examples. Moral uncertainty has been contrasted with value uncertainty. Several types of uncertainty and instability have been distinguished. We have seen that various combinations of the types and versions are possible.

Further conceptual clarification is possible, but for the time being we do not see any need to introduce a set of partially overlapping definitions. For the purposes of the present paper the umbrella definitions proposed here seem to be sufficient.

Various factors which produce or modify value uncertainty and value instability have been highlighted and discussed in this paper. An improved understanding of such factors, and of the relations between them, could be important, even for those who have no wish to replace value instability and uncertainty with their opposites. Knowledge of such factors is essential if one is to get to grips with the situation of decision-making.

Where value *uncertainty* is encountered, what can be done? In Section 4 we recommended further research and more time for reflection. Ideally this ought to improve self-knowledge, identify ambivalence, and clarify the situation and the role of value framing. The prospects for a remedy in the case of value *instability* seem less obvious, since the problem appears to be struc-

turally connected with the decision-making situation and has an ontological basis.

It remains to be said only that we now look forward to research in several areas: the implications of the various definitions proposed here, empirical aspects of the characteristics of decision-makers, the role of value uncertainty and instability in actual decision-making, and the precise consequences of various normative positions.

39

Lessons from Pascal Engel: Achilles, the tortoise and hinge epistemology for basic logical laws

ANNALISA COLIVA

I have known Pascal Engel for about fifteen years throughout which he has been an unfailing source of inspiration, support and amusement. It is therefore a great honor and pleasure to contribute to this Festschrift. Here I will focus on a number of recent writings in which has been concerned with the puzzle raised by Lewis Carroll's celebrated "What the tortoise said to Achilles". I have learnt a lot by reading Engel's work, in particular how to distinguish between the various problems raised by Carroll's paper. In what follows I will briefly summarize Engel's taxonomy and then focus on one specific problem elicited by Carroll's story. I will then consider some prominent solutions to it, which I will find wanting. Hence, in closing, I will sketch my own solution, signaling points of agreement and disagreement with Engel's. First of all, however, let me remind you of the enigmatic situation depicted in Carroll's paper.

1. What the tortoise said to Achilles

Achilles presents the tortoise with the following propositions

> (A) Things that are equal to the same are equal to each other
>
> (B) The two sides of this triangle are things that are equal to the same
>
> (Z) The two sides of this triangle are equal to each other

The tortoise accepts (A) and (B) but refuses to accept (Z), even though she accepts

> (C) If A and B are true, then Z must be true.

She also accepts that if (A) and (B) and (C) are true, (Z) must be true, but she still refuses to accept (Z). And the story suggests that no matter how many more propositions like (C) we could add and have the tortoise accept, she won't accept (Z).

As Engel rightly points out, it is a remarkable fact that the tortoise accepts (C) because she recognizes that (C) is a logical truth and also that she makes her acceptance conditional on entering (C) as a further premise.

2. Engel's taxonomy

Engel insightfully points out that Carroll's story lends itself to a number of different interpretations. That is to say, there are several philosophical problems raised by it. He notices that the most usual moral it elicits – indeed the one that Carroll himself drew – is that the tortoise, who requests that (C) be entered as a further premise, doesn't distinguish between a premise (like (A) and (B)) and a rule of inference (like (C)). Once that distinction is in place, the regress cannot start.

The second moral – drawn by Black (1951) and Stroud (1979) among others – hinges on the epistemology of understanding. If one understands (A) and (B) (and (C)), one can't refuse to accept (Z). So, by contraposition, if one doesn't accept (Z), either one doesn't understand the premises – and, in particular, the conditional *if then* – or else one's acceptance of them is faked.

The third lesson that can be elicited from Carroll's story – in keeping with Quine's (1935) reading of it – concerns the epistemology of logic and, in particular, the justification of basic logical laws. Let us assume that modus ponens

be such a basic rule of inference. How could we justify it? Inevitably it looks as if we would have to appeal to it in our reasoning towards its very justification. Hence, we would presuppose it in our reasoning, thereby providing a circular justification – that is to say, no justification at all – for it.

Finally, according to Engel (2005, 2007, 2009 but also to Blackburn 1995), Carroll's tale can be taken to exemplify the problem of the normative force of logical laws. How can it be that one accepts the premises, recognizes that (C) is a logical truth, and yet refuses to accept (Z)? The question gets traction if one considers logical laws to be external reasons which, as such, on a broadly Humean account of reasons, can't move subjects to behave accordingly. So the issue becomes: what kind of extra fact can move subjects to infer according to modus ponens after recognizing it is a valid law of inference?

Much of Engel's work has been devoted to giving an answer to the last problem, which rightly in my opinion, he sees as connected to the third one. We can bring out the connection following Engel's own discussion. One obvious candidate to the extra element which should move us to infer in accord with modus ponens is *habit*. To put it in Wittgensteinian terms: "that's simply what we do". Yet, the problem arises of justifying why it is correct to infer according to modus ponens and not, for instance, by affirming the consequent, which, after all, is something we do too (at least many subjects are inclined to do that, as anyone who has ever taught undergraduates would know!).

In what follows I will focus mainly on the third problem identified by Engel, even though that will have implications regarding the second and the fourth too. To anticipate a little: I think there is more to the Wittgensteinian answer than Engel makes of it and I will endeavor to show that, properly understood, it will provide us with an answer to both the problem of the justification of basic laws of inference and of the normative force of logic, once these problems get properly into focus.

3. The problem of justifying basic logical laws

When we consider the problem of justifying a basic logical law like modus ponens we can in fact be asking two different questions. The first one is: in virtue of what is *a subject* justified in believing a conclusion, reached by deploying a modus ponens inference, starting from justified premises?[1] To give an example: in virtue of what is my first-year student Emma justified in believing "Anna will take the umbrella" by reaching it through a reasoning procedure

[1] This is the question at the heart of Boghossian 2003.

that starts with justified premises such as "It's raining" and "If it's raining Anna will take the umbrella"?

The second question, in contrast, is: in virtue of what, *as theorists*, can we say that modus ponens in general – or, in other words, the very principle – is epistemically justified? When we raise that question, it should be kept in mind that we are looking for a justification that a theorist can deploy to vindicate the claim that modus ponens is justified.[2] That is to say, we can grant that modus ponens is a valid rule of inference, viz. that it is necessarily truth-preserving; yet we can still raise the question of what makes it that case that it is justified. We will come back to the relevant senses of this question in the following.

The answer I want to propose to our first question – in virtue of what is a subject justified in believing a conclusion such as "Anna will take the umbrella", upon inferring it from the justified premises "It's raining" and "If it's raining Anna will take the umbrella" – is, very crudely, *nothing*, apart from reaching the conclusion as a result of having entertained and understood the premises and having taken them to be justified at least *pro tem*, if only for the sake of argument and, therefore, apart from having in fact inferred to it via an application of modus ponens. In particular, a subject isn't required to know anything about modus ponens, not even that there is such a rule of inference, let alone have a notion of its being a valid rule of inference. Nor does she need to possess anything like a justification for it. That is to say, she need not have an intuition of the validity of modus ponens, if such a thing could ever exist;[3] nor should she be able to provide an argument in favor of modus ponens, for this would prevent many subjects from ever having such a justification, which, moreover, as we shall see at length in the following, would indeed be circular, as it would have to rely, at some point, on modus ponens (or on other basic rules of inference for which it would then be an open question how a subject could be justified in employing them). Finally, unless one were happy with crudely reliabilist notions of justification,[4] one would have to recognize that a subject could perfectly well reason in accord with modus ponens and thereby reach a justified conclusion, without having any justification for it. Alternatively, one would have at least to admit that, even if a subject could have such an externalist justification for a conclusion reached by reasoning

[2] This is nicely brought out in Schechter and Enoch (2006, p. 687 and 2008, p. 552) even though, as we shall see, they don't take full measure of this fact, as far as I can see.

[3] Doubts are forcefully cast on such a possibility by Boghossian 2003 and Wright 2006.

[4] Objections to crudely reliabilist accounts of justification in connection with the issues presently discussed can be found in Boghossian 2003 and 2012, repeated in Schechter and Enoch 2006, 2008.

in accord with modus ponens, she wouldn't have any to offer in response to the question of how she could be justified in holding a belief reached via an application of modus ponens. Either way, her inferring in accord with modus ponens can be as "blind" as one wishes it to be, even though, surely, it will genealogically depend at least on her possession of the relevant concepts – in particular the concept IF THEN. That is to say, since we are considering reasoning on certain propositional contents and their relations, all concepts needed to grasp both the former and the latter will have to be assumed to be at the subject's disposal. Therefore, the whole epistemological issue regarding the justifiedness of modus ponens is moved to the second question: Can we – as theorists – provide a justification for modus ponens?

Recall that the problem isn't that of explaining the validity of modus ponens but only why using it is epistemically justified: why is it the right thing to do, epistemically speaking, to use modus ponens, in our reasoning? Now, the problem is that, on a first-order reading of that question, it seems that we are looking for an argument to show that modus ponens is a right belief-forming method, as opposed to other belief forming-methods, which aren't epistemically kosher, such as affirming the consequent. But, as stressed, this can't be the main issue we are going to address, or that is in fact at the heart of the literature we are rehearsing here. For, presumably, the answer to that question is that modus ponens, as opposed to affirming the consequent, say, is a valid rule of inference and, moreover, it is basic, inasmuch as it is presupposed by all other forms of reasoning. So, surely we can't go astray by using it. If that were the whole issue, the answer would be simple, I think.

There is, however, a second-order reading of that question in the offing. Namely, let us grant that modus ponens is a valid rule of inference, and that we have formal means to prove that it is. How can we claim to possess that knowledge? The trouble here seems to be that any reason we might want to provide to that end will itself rely on the application of modus ponens. So it would be circular and therefore unsuitable as an account of how we can claim to know that modus ponens is valid. To see the point more clearly, consider Carroll's story again. Since any proof of the validity of modus ponens we might give will presuppose reasoning in accord with it, how could we use such a proof to convince someone who – like the tortoise – weren't already disposed to infer according to modus ponens? That proof would be able to convince only the converted.

4. The meaning-constitutive solution

Keeping in mind these qualifications on the understanding of the very problem we are addressing when we consider the issue of the justification of modus ponens, let us move on to examine some current answers to it. Notice, however, that in the existing literature there is always an oscillation between various issues – i.e. to prove that modus ponens is epistemically kosher, that we know that it is and to explain why we are required to abide by it.

Recent attempts in this field have seen supporting the view that its justification depends on meaning-constitutive considerations regarding the concepts involved in basic inferences deploying modus ponens, such as IF THEN (Boghossian 2003).[5] At first approximation, but there will be more on the issue in the following, the idea would be to say that such a belief-forming method is justified because the ability to reason in accord with it is constitutive of the understanding of the concept IF THEN, that figures in it. So grasp of the latter concept requires being prepared to reason in accord with modus ponens. What would be wrong with the tortoise, therefore, is that, insofar as she possesses the concept IF THEN, she is required to infer in accord with modus ponens. By contraposition, if she doesn't reason in accord with modus ponens, this shows that she doesn't really have the concept IF THEN in the first place.

These considerations have been opposed by several theorists who have pressed the point that grasp of the concept of the conditional isn't sufficient to provide a justification for modus ponens because one can grasp the former while sensibly wondering whether modus ponens is indeed valid, at least globally (McGee 1985 and Williamson 2003). So one could possess that concept while not being willing to reason in accord with modus ponens, at least in some cases.

Another line of attack, pressed by the late Paolo Casalegno (2004), is to say that there could be someone who could be said to grasp the concept IF THEN and yet be prevented by whatever causes ever to draw an inference in accord with MP.[6]

I personally think that these aren't fatal objections to the view, since, arguably, the cases invoked by McGee and Williamson aren't *basic* instances of

[5] Even though Boghossian takes himself to be answering the question of how a subject could be justified in believing a given conclusion reached through a specific application of modus ponens, rather than the question of the justification that we, as theorists, can give of that very principle.

[6] In fact Casalegno's example involves the concept AND, but to stick to our leading example in this section I have taken the liberty to change it to IF THEN.

modus ponens, for they involve embedded conditionals and depend on complex contextual information which could explain why we think it intuitive to reject a conclusion reached via an application of modus ponens.[7] Presumably, a supporter of a meaning-constitutive account of the justifiedness of modus ponens should qualify her claims so as to be able to single out those instances of reasoning in accord with modus ponens that are actually constitutive of having the concept IF THEN. Just to help the reader better see the point: those inferences couldn't plausibly be ones which require entertaining indefinitely long premises, etc. So, such a theorist would surely be within her rights in confining her claims regarding the justification of modus ponens to basic instances of it, involving atomic sentences, once all potentially confusing contextual elements have been avoided or disambiguated. Then it would certainly become much more plausible that unless one were prepared to infer Q, given P and "If P then Q", one wouldn't so much as have the concept IF THEN.[8]

As to the objection raised by Casalegno, I think it remains an entirely open question, one which is hard to see how it could be settled, whether a subject unable to infer Q, given "If P then Q" and "P", could really be said to possess the concept IF THEN. For, *ex hypothesi*, such a subject would be able to utter sentences containing "if then", but since he would never be in a position to use it as a premise in an actual chain of reasoning, it would be unclear what evidence there could ever be to show that he does indeed have the corresponding concept.[9]

In this connection, however, I think Engel makes an important observation

[7] A contrived McGee's counterexample, I draw from Cariani 2013, is the following. We know that Mario, an Italian, doesn't like travelling at all and if he travels at all, he likes going as close to home as possible. His parents are trying to convince him to visit some of his relatives, some of whom live in Paris and some of whom live in LA. So Lucia, Mario's mother, may reason thus: (1) If Mario doesn't go to Paris, if he travels at all, he will go to LA; (2) Mario won't go to Paris; hence (3) If Mario travels at all, he will go to LA. According to McGee, while the premises are intuitively acceptable, the conclusion isn't. Personally I never found this claim intuitive. Be that as it may, the point remains that the reasoning involves embedded conditionals and that the intuitions it should elicit depend on contextual information about Mario and his psychology.

[8] For a similar kind of reply, though applied to the case of AND, see Boghossian 2012, pp. 232-3. Williamson 2012 takes issue with Boghossian's reply. As far as I understand the reply, though, it seems to concede that the deviant subject would have to behave like non-deviant ones in non-contentious cases and yet deviate in contentious ones (cf. p. 243). But, as far as I can see, this would mean to concede the basic inferentialist point: there are some instances of basic inferences one has to be willing to make in order to count as having the concept (or the same concept) AND (or IF THEN) we do.

[9] For a similar reply, but addressed to Casalegno's original example (cf. fn. 6), see Boghossian 2012, p. 228-9. Williamson 2012 objects to this line of reply.

in his discussion of Carroll's original story. For he notices that the tortoise accepts (C) and so she seems to have a perfect grasp of the meaning of "if then", yet she refuses to reason in accord with it and thus to infer (Z). Still (C) doesn't contain any embedded conditional. So it seems possible to have a grasp of IF THEN and still not infer in accord with it. This puts pressure on an account of the justifiadness of modus ponens based on meaning-constitutive considerations.

There are other objections. For instance, there are concepts whose introduction and/or elimination rules give rise to invalid inferences. So how could subjects be ever justified in inferring in accord with them? Alternatively, if they refused to infer in accord with such concepts, would they count as not having the relevant concepts? Consider TONK and the rules of inferences constitutive of it.[10]

If A, then A TONK B If A TONK B, then B

Where the problem would arise when B = not-A, for from A its negation would follow. Thus it would seem that we would have concepts that license certain inferences, which would have the remarkable consequence of leading us to unwarranted conclusions. Moreover, there is no intuitive sense in which we should be compelled to reason in accord with them, even though, on a conceptual-role semantics, possessing those very concepts would require at least having the disposition to infer in accord with their constitutive rules of inference.

As is well-known, some theorists, like Christopher Peacocke (1992), would say that TONK isn't a genuine concept and would thus be able to stop one possible counterexample by defusing it this way. However, there are cases that give rise to equally invalid inferences and yet where it is difficult to deny that genuine, albeit obnoxious concepts are at issue. A case in point is BOCHE.

If A is German, A is BOCHE If A is BOCHE, A is cruel

From which it would be inferred that all Germans are cruel. Hence the fact that reasoning in accord with those rules is needed in order to possess the relevant concepts (e.g. BOCHE) doesn't guarantee that the form of inference utilized is valid. Furthermore, we would face the problem that in order to

[10] Prior 1960.

possess that concept we ought to be disposed to make the relevant inferences, while, intuitively, non-racists could have the concept BOCHE, while not at all being disposed to conclude that all Germans are cruel.

This poses the problem of clarifying better under which conditions a rule of inference is justified, according to meaning-constitutive accounts. Yet, to solve the problem posed by BOCHE, it isn't enough to require that the rule of inference under scrutiny should be necessarily truth-preserving. The following example, due to Boghossian (2003), makes this point vivid.

If x is an elliptical equation, x is FLURG	If x is FLURG, x can be correlated with a modular form.

The Taniyama-Shimura conjecture, proved in 1999, states that all elliptical equations can be correlated with modular forms. Hence, inferences licensed by FLURG are necessarily truth-preserving. Yet, clearly, inferences like the one just stated don't seem to be justified. They don't seem to put a subject in a position to draw a justified conclusion, starting from the allegedly justified premise that a given equation is elliptical, for they introduce the non-existing entity flurg which, like phlogiston, can hardly give rise to warranted conclusions. Furthermore, like in the previous case, it seems quite intuitive to hold that one could have the concept FLURG without thereby being disposed to draw the relevant inferences.

It is on the basis of considerations like the ones just explored that Boghossian proposes that, whenever available, only the conditionalized versions of the relevant rules of inference would be justified. Hence, the correct conditional stipulation for FLURG would be as follows (Boghossian 2003, p. 247):

> If there is a property which is such that, any elliptical equation has it, and if something has it, then it can be correlated with a modular form, then if x has that property, x is flurg.

In the case of IF THEN, however, we can't conditionalize it without circularity, for plainly the conditional would be needed to perform such a conditionalization. Hence, according to Boghossian, the application of modus ponens is justified because reasoning in accordance with it is constitutive of possessing the concept IF THEN.

We have already seen, however, that, as Engel reminds us of, the tortoise accepts (C) and so she seems to have a grasp of IF THEN, even though she

isn't disposed to infer (Z). Hence, it remains difficult to see how meaning-constitutive considerations could provide a justification of modus ponens capable of overcoming the eventual doubts of the unconverted. Presumably a supporter of this strategy would have to deny that the tortoise really accepts (C), but this would imply that Carroll himself was unclear about his own tale, which is a difficult consequence to swallow.

5. The pragmatic solution

Be that as it may, I think it is instructive to consider another strategy recently adopted to justify modus ponens, proposed, in slightly different fashions, by Wright (2004b) and Schechter and Enoch (2006, 2008). Engel himself (forthcoming) seems drawn to this solution. Schechter and Enoch call it the "pragmatic" strategy – and rightly so, as we shall see – even though, in my opinion, they don't really take the measure of this fact. They also rightly point out (2008, p. 548) that, if it works at all, it provides merely an entitlement for a second-order claim, i.e. "We know/justifiably believe that modus ponens is valid" (cf. also Wright 2004b, p. 158), even though they don't seem to take full measure of this – to my mind – crucial fact either.[11]

So here are the backbones of their proposals. According to Schechter and Enoch (2008, p. 554):

> If a belief-forming method is such that it is possible to successfully engage in a rationally required project by employing it, and such that it is impossible to successfully engage in the project if the

[11] As we shall see, they oscillate between aiming to provide a justification of modus ponens itself and recognizing that the problem is at second-order. An important clue of this conflation is the fact that they feel the need for their account to provide a distinction between valid and invalid rules of inference (cf. Schechter and Enoch 2008, p. 548, where they contrast modus ponens, for instance, and affirming the consequent). But that isn't the main problem, as we have already remarked upon several times. Let us grant that there are basic rules of inference and that some of them can ostensibly be proven to be valid, i.e. necessarily truth-preserving. The issue we face, then, is: are these belief-forming methods justified and, if so how? Where this should be taken as a quest for a reason that we, as theorists, can produce in order to vindicate the epistemic legitimacy of our employment of such valid and basic rules of inference. As we saw, a way of making the problem vivid is to consider what we might say to someone who weren't already inclined to reason in accord with it. Once we are clear about the kind of question we are asking, we can, I think, more easily see, why indeed there is no justification for modus ponens itself and why all we can do is to provide ourselves with an a priori reason to believe the following "To reason in accordance with modus ponens is rational (even if there is no justification of modus ponens as such)".

> method is ineffective, then we are prima facie epistemically justified in employing that method as basic, even in the absence of a justified belief concerning the method.

Consider for instance modus ponens. The idea is that it is a belief-forming method that is necessary to successfully engage in the rationally required project of reasoning. Hence, we are prima facie epistemically justified in employing it, even if we don't have any justified belief regarding it. That is to say, even if we can't provide a non-circular justification of it, which could prove its correctness.

Schechter and Enoch introduce a number of qualifications, which clarify and sharpen their proposal. For instance, they tell us that success in engaging in the project need not involve achieving perfection (ivi, p. 559). They also remark that it is conceivable that in distant possible worlds the same project is successfully accomplished by applying totally different methods. Therefore, the notion of "impossibility" appealed to in the previous definition is to be understood as relative to sufficiently close possible worlds (ivi, p. 562).

Let us now turn to Wright's proposal, whose closeness in spirit to Schechter's and Enoch's will be apparent. Say that

> P is a presupposition of a particular cognitive project if to doubt P (in advance) would rationally commit one to doubting the significance or the competence of the project. (Wright 2004b, p. 163).

An entitlement of cognitive project can then be defined as follows:

> (i) There is no extant reason to regard P as untrue and
>
> (ii) The attempt to justify P would involve further presuppositions in turn of no more secure a prior standing, ..., and so on without limit; so that someone pursuing the relevant enquiry who accepted that there is nevertheless an onus to justify P would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessor (ibid.)

Wright's claim is that modus ponens is a presupposition of any cognitive project involving reasoning and that any attempt to justify it would presuppose it, since it would involve reasoning (cf. Wright 2004b, p. 166). Furthermore, he thinks there cannot be basic counterexamples to it (Wright 2004b, p. 171). For, when P and Q are atomic sentences, if, given P and "If P then Q",

one isn't prepared to infer Q, this "convicts the thinker of misunderstanding of the conditional" (Wright 2004b, p. 170). Hence, we possess an entitlement of cognitive project for it.[12]

Now, clearly, neither proposal provides us with a first-order justification for modus ponens. So if the task was to tell apart good and bad basic inferences, like modus ponens on the one hand and affirming the consequent on the other, nothing of consequence will follow from the considerations just advanced.[13] As is well-known, entitlements of cognitive project don't speak to the likely truth, or in this case, to the validity, of a given rule of inference, however basic that might be.

One could then think that, at least on the assumption that modus ponens is valid, the proposals presently on the table will allow us to claim that we know or can justifiably believe – in some sense of "justifiably" – that modus ponens is valid. Now, the trouble, as Wright himself recognizes, is that they don't achieve even that much (Wright 2004b, pp. 168-9). For all these strategies attain is to make apparent why it is natural, or even indispensible for us to apply modus ponens,[14] but they do not give us an epistemic reason to vindicate the second-order claim that we *know* (or justifiably believe) that it is valid. To see the point more clearly, consider that if we were to propound the considerations advanced by these authors to someone like the tortoise in Carroll's example – someone, that is, who weren't already prepared to infer in accord with modus ponens – they would certainly not convince such a subject that modus ponens is a valid rule of inference. They would merely make it apparent to her why we find it unavoidable to employ it. Hence, it seems to me that the proposals presently under consideration cannot be seen as providing an answer to the question "How can we claim that we know/justifiably believe that modus ponens is valid?" they aimed to answer.

Nor do they solve the other problem of providing a diagnosis of what

[12] Here I won't consider Wright's final move in his 2004b paper to boost that entitlement to get knowledge of modus ponens out of it, for it relies on a form of "alchemy", which is extremely suspect, as he himself somehow recognizes by allowing that people may well think that his rule-circular account of our knowledge of the validity of modus ponens would "prove to founder" (p. 174).

[13] Surprisingly, Schechter and Enoch (2008, p. 564), at least on one possible understanding of the phrase "there must be some substantive criterion that distinguishes epistemically justified basic methods from the rest", miss that much and go on to say "this is where the pragmatic account fits in. It provides a general, principled explanation of in virtue of what certain basic belief-forming methods are justified".

[14] Surprisingly, Schechter and Enoch (2008, p. 564) take the indispensability for us to reason in accord with modus ponens as an epistemic justification for it.

would be wrong with the tortoise in Carroll's story. For surely they would return the verdict that the tortoise, by refusing to draw the conclusion that Q, given P and "If P then Q", would be prevented from engaging in a valuable cognitive project, i.e. that of reasoning. But why would this be a *rational* deficiency on her part? It isn't very informative to be told, in this regard, that reasoning is a "rationally required" project (cf. Schechter and Enoch's definition above).[15] In what sense is that project required by *rationality*? And why does rationality require it to be executed on the basis of modus ponens rather than on the basis of affirming the consequent, say? To make headway with respect to the diagnosis of what's wrong with the tortoise, we need an explanation that ties reasoning in accordance with modus ponens to the very notion of logical rationality.

6. An alternative solution: hinge epistemology for basic logical laws

Let me introduce the backbones of an alternative solution by considering a parallel case, which I have discussed at length in other writings of mine (cf. Coliva 2012a, b, 2014, forthcoming). That is to say, the role that a proposition like "There is an external world" plays with respect to *epistemic* rationality. We may say, in a Wittgensteinian spirit, that the notion of epistemic rationality depends on a practice of forming, assessing and withdrawing from beliefs about physical objects in our surroundings, such as "Here is my hand", on the basis of perceptual evidence, whose role is precisely that of making certain beliefs, or their negations, more likely true. Such a practice has as a *hinge*, i.e. as one of its constitutive assumptions, "There is an external world" (other ones may be "My sense organs are generally reliable", "I'm not the victim of a lucid and sustained dream", etc.).

For reasons which I can't possibly rehearse here (but see Coliva 2012a, b, 2014, forthcoming), I don't think we can provide any kind of justification for it. For the attempt to derive it from the perceptual justification we have for specific empirical propositions via correct inferences, like G. E. Moore's celebrated proof of an external world, would actually depend on taking for granted such an assumption. A priori justifications would be hard to come

[15] Schechter and Enoch (2008, p. 558) attempt an answer in terms of a project which is such that "a particular agent rationally ought to engage in it given the facts of her constitution and general abilities". They do recognize, however, the uninformativeness of their characterization.

by and entitlement strategies, such as Wright's (2004a), would actually fail to deliver the intended goods, in my opinion.

If that's the case, "There is an external world" turns out to be unjustifiable. Yet, to assume it is constitutive of epistemic rationality, for otherwise we couldn't engage in the practice which is itself constitutive of our notion of epistemic rationality. The reason being that, without that hinge, we couldn't take our perceptual experiences to bear on beliefs about mind-independent objects. Hinges such as "There is an external world" are therefore needed to overcome our "cognitive locality".[16] That is, they are needed in order to be within our epistemic rights in forming beliefs about physical objects based on our current sensory experience. Hence, we are *mandated* by epistemic rationality itself to make that assumption, for without it, there would be no possible perceptual justification for or against a given belief about specific material objects. As a consequence, there would be no practice and therefore no notion of epistemic rationality either, at least as we – and a skeptic – usually understand it.

In connection with modus ponens, I think we can proceed in a similar fashion. First of all, we could point out, in a Wittgensteinian spirit, that the notion of *logical* rationality doesn't hang in the air, but depends on a practice of reasoning by employing certain *basic* patterns of inference which are *valid*, i.e. necessarily truth-preserving.[17] Basic instances of modus ponens are among these patterns of inference (others may be conjunction-elimination and disjunction-introduction). Hence, to reason in accord with modus ponens, at least in certain basic cases, is itself constitutive of logical rationality. That is to say, if one didn't do it, one wouldn't count as rational, at least on the notion of logical rationality we do have. Therefore, we are mandated by logical rationality itself to reason in accord with modus ponens. This is no proof of the validity of modus ponens, though. That proof will be provided differently, yet obviously in a rule-circular fashion, since whatever logical method we will apply to that end, it will presuppose reasoning in accord with modus ponens. Still, given that we can actually prove that modus ponens is necessarily truth-preserving, I think we can grant, in a somewhat externalist spirit, that we know that modus ponens is valid. But we can't *claim* that knowledge, for we can't prove having it to someone who didn't already reason in accord with modus ponens, thereby being implicitly willing to admit its validity. In

[16] This is Wright's (2004a) phrase.

[17] Such a qualification should allow us to dispense with McGee's alleged counterexamples altogether, for even if they were genuine (but see fn 7), they would clearly not be basic instances of modus ponens.

this respect, we can't but accept or take for granted that modus ponens is valid, without being in a position to prove, in a non-circular way capable of persuading the unconverted, that it is. To notice that modus ponens is constitutive of logical rationality, however, gives us an a priori justification for the proposition "To reason in accordance with modus ponens is rational", for it tells us that to do so is constitutive of logical rationality itself.[18]

So, now, suppose we met a tribe who, in basic cases, didn't reason in accord with modus ponens, after we have ascertained that they do mean "if then" the way we do and that no contextual factors intrude in such a way as to make it understandable why they might seem to deviate from modus ponens.[19] Or suppose we met the tortoise of Carroll's example. Namely someone who, *ex hypothesis*, understands "if, then" like we do, but who is simply unwilling to infer (Z), after admitting that both (A) and (B) (and (C)) are justified. Surely we couldn't do much to remove their stubbornness, but we could conclude, with Frege, that by doing so, they would simply be outside the scope of logical rationality. As he writes:[20]

> What, however, if beings were even found whose laws of thought directly contradicted ours, so that their application often led to opposite results? The psychological logician could only accept this

[18] Interestingly, Wright 2012 puts forward the view that to reason in accord with modus ponens is constitutive of rationality as an explanation of what an inference – in fact a basic inference such as a basic instance of modus ponens – is, but doesn't endorse this as an explanation of why modus ponens is justified, which he still thinks can be afforded by means of his entitlements-of-cognitive-project strategy. To repeat a point worth stressing, in order to avoid potential confusions, I don't think we can justify modus ponens, unless by this we mean giving a proof of its validity. But we can surely provide ourselves with a justification for the second-order belief that to reason in accordance with it is rational. As stated, that justification is given by the argument we have proposed in favor of the view that to reason in accordance with it is constitutive of logical rationality.

[19] Notice that the alleged ethnographical counterexample of the Azande, due to E. Evans-Pritchard, is no such case, for it were based on a mistaken translation. In particular, the first ethnographers hadn't paid attention to the fact that for the Azande only the sons of a witch-doctor who are "hot", whatever that might mean, are witchdoctors in their turn, despite the fact that the Azande believe that witchcraft transmits patrilinearly by means of a magic substance all sons of a witchdoctor inherit from their father. A similar mistake in translation can be found in ethnographic early reports about the tribe of the Kassena (Mangiameli 2010) who live in northern Ghana. They were reported to believe that baobabs are sacred and therefore intangible, while they were witnessed cutting them with no specific sense of guilt. Further studies revealed that the Kassena think that only after a certain age baobabs are sacred and therefore that they can innocently cut them before that age.

[20] Frege GGA, XVI.

> and say: for them, those laws hold, for us these. I would say: here we have a hitherto unknown kind of madness.

I wouldn't ban them as necessarily mad, in the sense of being insane, but I would deem them as logically irrational. Let me also note that concurring with Frege's judgment in this case doesn't depend on embracing either his unshakable faith in classical logic, or his overall conception of logic. To stress, I do agree with him that logic is normative and non-descriptive. However, the provenance of these norms, according to the Wittgensteinian perspective I am concerned to develop, lies in our communal practices. It isn't written in stone in a third-realm of abstract entities.

Notice that the point isn't that such a deviant community would be inconceivable. Of course it is conceivable and we've just envisaged it, provided sense could be made of someone who had the relevant concept but weren't prepared to reason in accord with modus ponens. The point, rather, is that it would not show that the notion of "logical rationality" is relative – i.e. that abiding by different and incompatible basic logical laws could qualify as being equally logically rational. For as long as the meaning of "logical rationality" stays put, that requires people to engage in forms of reasoning governed by basic rules of inference, which are valid, so that if one refused to apply them, or even followed different, invalid ones, we could convict them of logical irrationality. Hence, the diagnosis of what's wrong with the tortoise, isn't that she wouldn't know the meaning of "if then", or that she would be merely prevented from taking part in a project which is extremely valuable or even indispensible to us. Rather, it would be that she would be irrational, as she refuses to take part in an activity which is constitutive of logical rationality itself.[21]

One might then say that these people would have their own, equally legitimate notion of rationality, even though incompatible with ours, characterized by appeal to different rules of inference. But in order to take this possibility seriously, i.e. as a legitimate alternative to ours, we should look into their rules of inference. The options, I think, would be as follows: (1) their rules are basic but invalid (like affirming the consequent); or (2) they are valid but not

[21] It is then to be assessed whether she could also be convicted, as Frege thought, of not being capable of thinking at all. That should be assessed, presumably, by observing her behavior and see whether her refusal to abide by basic logical laws is widespread and consistent. Surely, in the kind of normative perspective I've depicted, were she consistently and widely irrational, she would end up being deemed unable to think. But this is clearly not the case of the tortoise in Carroll's story.

basic. In the former case, we could still convict them of logical irrationality, for to count as logically rational in general (or to exhibit a kind of rationality which could be a serious alternative to ours) these rules should be at least necessarily truth preserving. In the latter case, in contrast, it could presumably be shown that these valid but non basic rules of inference presuppose valid and basic ones, and, in particular, ones we abide by, such as modus ponens, that are constitutive of our notion of logical rationality. Hence, in neither case would have we found a notion of rationality, determined by the observance of different rules of inference, which, while incompatible with ours, could be taken as a serious alternative to it. As I have argued elsewhere (Coliva 2010), I think that both options would in fact be in keeping with a Wittgensteinian treatment of similar cases. So, far from merely observing that reasoning in accord with modus ponens is simply what we do, and from allowing for different, incompatible and yet perfectly legitimate inferential practices, I think we can find in Wittgenstein's writings the seeds of a much more theoretically robust and of a much less concessive line of reply.

7. Conclusions

Following Engel's taxonomy of the problems raised by Carroll's story, I have focused on the issue, central to the epistemology of logic, of justifying basic laws of inference such as modus ponens. I have claimed that it should be carefully distinguished both from the problem of determining the validity of modus ponens and from the problem of what justification is needed by a subject in order to be justified in believing a particular conclusion reached through a specific application of modus ponens. I have claimed that it must be interpreted as the problem of providing ourselves, as theorists, with a justification to believe that it is a valid rule of inference. Once looked at in this way, it becomes apparent that it is indeed a hard problem. For any account we would be in a position to offer would presuppose applications of modus ponens and, thus, it could only convince the converted.

I have then introduced two prominent solutions to it, to be found in the recent literature on the topic. Namely one that appeals to the meaning-constitutive role of modus ponens with respect to "if then" and one which appeals to its indispensability to us, given our rational projects. I have found both of them wanting, even though for different reasons.

I have then sketched a third, Wittgensteinian, solution. The crucial claim is that, with respect to logical rationality, modus ponens plays a role similar

to the one played by certain "hinges" with respect to epistemic rationality. Namely, that law is constitutive of the very practice which is itself constitutive of the notion of logical rationality. My Wittgensteinian solution, therefore, recognizes a central role to use and practices but only because they determine norms of rationality, i.e. norms that make the very exercise of reason possible. In this sense, I think there is more to Wittgenstein's position than a mere reminder of what we do, or can't help doing, given the kind of creatures we are, or if we want to take part in cognitive projects that are valuable to us. These naturalist, Humean readings of his later philosophy,[22] as well as their pragmatist cousins, seem to me to miss the crucial aspect of his mature thought. That is to say, that uses and practices give rise to norms. So, in my view, Kripkenstein[23] – the famous, infamous skeptic about rules – isn't Wittgenstein. The real guy was, in my view, fully convinced of the existence of norms, but he was neither a Platonist, nor a full-blown conventionalist, about their provenance – whence his insistence on use and human practices, which, however, when it comes to the basics are, for him, part of a *shared* form of life. Yet, if we keep in mind the crucial normative import of Wittgenstein's considerations, we do have the means to provide ourselves with an a priori justification for holding that reasoning in accord with modus ponens is rational and to diagnose what would be wrong with those who refused to abide by it, like the tortoise in Carroll's story. For we could actually convict them of irrationality. So, contrary to Engel's reading of Wittgenstein,[24] the point isn't that of convincing or forcing those who don't, or even refuse to reason in accord with modus ponens. That's beyond logic and rationality. Rather, it is to be within our rights in passing a normative judgment with respect to their behavior, instead of being forced simply to accept it as one possibility among many equivalent ones.

8. References

Black, M. 1951 "Achilles and the tortoise", *Analysis* 11/5, pp. 91-101.

Blackburn, S. 1995 "Practical tortoise raising", *Mind* 104/416, pp. 696-711.

[22] To which Engel (forthcoming, p. 15) seems to give in.

[23] See Kripke 1982.

[24] Engel (forthcoming, p. 6) writes: "Could Achilles have retorted to the Tortoise: 'But don't you know what the practice of inferring according to Modus Ponens is?' Perhaps Achilles could have bitten him with a stick until he inferred, as the Wittgensteinian line suggests. But it's likely that the Tortoise would have still resisted, probably by re-entering his head under his shell".

Boghossian, P. 2003 "Blind reasoning", *Proceedings of the Aristotelian Society, Supplementary Volume* 77, pp. 225-248.

Boghossian, P. 2012 "Inferentialism and the epistemology of logic: Reflections on Casalegno and Williamson", *Dialectica* 66/2, pp. 221-236.

Cariani, F. 2013 "Modus ponens", *Aphex* 7, pp. 1-32.

Carroll, L. 1895 "What the tortoise said to Achilles", *Mind* 4/14, pp. 278-280.

Casalegno, P. 2004 "Logical concepts and logical inferences", *Dialectica* 58/3, pp. 395–411.

Coliva, A. 2010 "Was Wittgenstein an epistemic relativist?", *Philosophical Investigations* 33/1, pp. 1-23.

Coliva, A. 2012a "Varieties of failure (of warrant-transmission–what else?!), *Synthese* 189/2, pp. 235-254.

Coliva, A. 2012b "Moore's proof, liberals and conservatives. Is there a (Wittgensteinian) third way?" in A. Coliva (ed.) *Mind, Meaning and Knowledge. Themes from the Philosophy of Crispin Wright*, Oxford, Oxford University Press, pp. 323-351.

Coliva, A 2014 "Moderatism, transmission failures, closure and Humean scepticism", in E. Zardini and D. Dodd (eds.) *Contemporary Perspectives on Scepticism and Perceptual Justification*, Oxford, Oxford University Press, forthcoming.

Coliva, A. forthcoming *Extended Rationality. A Hinge Epistemology*, ms.

Engel, P. 2005 "Logical reasons", *Philosophical Explorations*, 8/1, pp. 21-35.

Engel, P. 2007 "Dummett, Achilles and the Tortoise", in R. Auxier and L. E. Hahn *The Philosophy of Michael Dummett, The Library of Living Philosophers* XXXI, Chicago and La Salle, Open Court, pp. 725-752.

Engel, P. 2009 "Oh! Carroll! Raisons, norms et inference", *Klesis. Revue Philosophiques* 13, pp. 21-39.

Engel, P. 2012 "Achille, la tortue et le problème de la connaissance logique", *AL-MUKHATABAT. A Trilingual Journal For Logic Epistemology and Analytical Philosophy* 1, pp. 61-71.

Engel, P. forthcoming "The philosophical significance of Carroll's regress", ms.

Evans-Pritchard, E. E. 1937 *Witchcraft, Oracles and Magic among the Azande*, Oxford, Oxford University Press.

Frege, G. 1893/1903 *Grudgesetze der Arithmetik*, Jena, Verlag Hermann Pohle. English Translation in P. Geach and M. Black (eds.) *The Philosophical Writings of Gottlob Frege*, Oxford, Basil Blackwell, 1960, pp. 137-244.

Kripke, S. 1892 *Wittgenstein on Rules and Private Language*, Oxford, Basil Blackwell.

Mangiameli, G. 2010 *Le abitudini dell'acqua*, Milano, Edizioni Unicopli.

McGee V. 1985 "A counterexample to modus ponens", *Journal of Philosophy* 82, pp. 462–471.

Peacocke, C. 1992 *A Study of Concepts*, Cambridge (MA), MIT Press.

Prior, A. N. 1960 "The runabout inference ticket", *Analysis* 21, pp. 38-39.

Quine, W. v. O. 1935 "Truth by convention" reprinted in *The Ways of Paradox and other essays*. Revised edition. Cambridge (MA), Harvard University Press, 1976, pp. 77-106.

Schechter, J. & Enoch, D. 2006 "Meaning and justification: The case of modus ponens", *Noûs* 40/4, pp. 687-715.

Schechter, J. & Enoch, D. 2008 "How are basic belief-forming methods justified?", *Philosophy and Phenomenological Research* 76/3, pp. 547–579.

Stroud, B. 1979 "Inference, belief and understanding", *Mind* 104, pp. 179-196.

Williamson, T. 2003 "Understanding and inference", *Proceedings of the Aristotelian Society*, Supplementary Volume 77, pp. 249-293.

Williamson, T. 2012 "Boghossian and Casalegno on understanding and inference", *Dialectica* 66/2, pp. 237-247.

Wittgenstein, L. 1969 *On Certainty*, Oxford, Blackwell.

Wright, C. 2004a "Warrant for nothing (and foundations for free?)", *Proceedings of the Aristotelian Society*, Supplementary Volume 78/1, pp. 167–212.

Wright, C. 2004b "Intuition, entitlement and the epistemology of logical laws", *Dialectica* 58/1, pp. 155-175.

Wright, C. 2012 "Meaning and assertibility: some reflections on Paolo Casalegno's 'The problem of non-conclusiveness'", *Dialectica* 66/2, pp. 249–266.

40

Why Ought We to be Logical? Peirce's Naturalism on Norms and Rational Requirements

JEAN-MARIE CHEVALIER

How should we think? "It behooves a man first of all to free his mind of those four idols of which Francis Bacon speaks in the first book of the *Novum Organum*. So much is the dictate of Ethics, itself. But after that, what?", Peirce asks (5.593, 1903). Bacon's overrated work does not divulge very much (W2.311). Surely we ought to be logical, but why? The question is strange, because its answer seems self-evident: we just ought to be. Logical laws are held as norms of our thought. But there are also rules of formation and justification of our beliefs, a whole art of thinking, which play a normative role on our minds. We have cognitive dispositions and intellectual virtues as well. And as mentioned in the quote, there is ethics, too.

This brings about at least three sets of questions. First, how can norms guide our behaviors? In particular, does normativity necessarily imply prescriptions? This is a concern of a psychological kind. Next, an ontological approach: what are norms? Are they real, fictions of our minds, or ideal descriptions? Is every norm equivalent to a value? And finally, there is the epistemological question of our access to norms.

The first set of questions is psychological, but it is not certain that psychology has the means to solve it. The concept of 'normativity' is expected to overcome the limits of an approach which would reduce thought to a knowledge of our contingent minds. The problem consists of determining whether our mental activities are normative, in the sense that they can be evaluated

as good or bad, or natural facts resulting from causes. Or is it possible to accommodate 'transcendentalism' with a form of naturalism? Such has been the attempt of many Peircean scholars, from (Goudge 1947) to (Lane 2009). In order to clarify the problem, we need to take a close look at such oppositions as the natural and the normative, the descriptive and the prescriptive, the objective and the subjective, fact and value.

1. Anti-psychologism within a naturalistic frame

There is a strong contrast between the highly problematic claims of Peirce concerning the role of psychology (and even physiology) in logic, and the very few attempts to disentangle the matter[1]. There have been mainly three attitudes in Peircean scholarship: either it focused on Peirce's early explicit statements on his unpsychological conception of logic, or on the late, as explicit statements about his normative conception of logic, or on the apparent psychological tincture of his theory of inquiry. The latter position, among which (Kasser 1999) is probably the most influential, intended to show that grounding logic on a theory of doubt and belief does not commit to psychologism.

This is very partial when compared to Peirce's abundant corpus on logic and the theory of reasoning, and on psychology. Furthermore, Peirce was himself a logician and a mathematician, and contributed significantly to experimental psychology, so that one can expect from him a broad conception of the relations between logical and psychological 'knowledges'. Nevertheless, one could show that his positions, which obviously evolved over time, are quite puzzling. I will take the easy way out in quoting a few surprising remarks from an anti-psychologistic logician:

> the analysis of conceptions will be psychology (W1.64)
>
> some anthropological facts have a great bearing upon logic (W1.362)
>
> [the three categories] may indicate an anthropological fact (W1.524)
>
> [logic] is bound, by its very nature, to push its research into the manner of reality itself, and [...] must inevitably consider how and what we think (W2.165)
>
> psychophysical laws will not fail to shed a strong light upon the theory of logic (W4.40)

[1] Cf. mainly (Colapietro 2003), (Dougherty 1980), (Hookway 2000), (Hookway 2010), (Kasser 1999).

> In order to gain a clear understanding of the origin of the various signs used in logical algebra [one has to show that] Thinking, as cerebration, is no doubt subject to the general laws of nervous action. (W4.163)
>
> Now modern logic enables us to show that three conceptions are really essential in formal logic; so that they are three fundamental categories of thought. Furthermore, reasons can be given for holding that these three conceptions are due to the three fundamental faculties of the mind, these again to three fundamental functions of the nerves; and finally these to three elementary constituents of the physical universe. (W5.237)
>
> We find the ideas of First, Second, Third, constant ingredients of our knowledge. It must then either be that they are continually given to us in the presentations of sense, or that it is the peculiar nature of the mind to mix them with our thoughts. Now we certainly cannot think that these ideas are given in the sense. [...] They ought therefore to have a psychological origin. (W6.182)
>
> Reasoning is performed by the mind. Hence, the logician must not be entirely neglectful of the science of mind. (W6.418)
>
> [There are some] psychological truths needed in logic (MS 400, 1894)
>
> Something like psychological association certainly appears in logic (2.45, 1902)
>
> [logic] rests on certain facts of experience among which are facts about men (5.110, 1903)

A task for the commentator would be to take a detailed look at each of these occurrences in the context of its theoretical background and explain why it does (or does not) cohere with Peirce's "assumed anti-psychologism"[2]. It would be required to examine the many ways of psychology-making, and the various acceptations of 'psychologism'[3]. It would probably give the supporter

[2] I thereby mean that, especially after J. Kasser's paper (implicitly directed against C.J. Dougherty's thesis that Peirce became anti-psychologist after 1896), most of the commentators agree that Peirce did not transgress his anti-psychologistic program. Cf. especially (Hookway 1992) and (Hookway 2000: 8) which contends that "Peirce consistently attacked psychologicism in logic".

[3] Cf. (Rath 1994). (Kusch 1995: 93-119) identifies no less than eleven psychologistic schools for the period 1866-1931 in Germany.

of the Peircean antipsychologism thesis quite a hard time: for instance, one definition of weak psychologism, that "psychological investigation into actual human thought processes constitute necessary though not sufficient conditions for enquiring into the foundations of logic" (Mohanty 1985: 2), is almost a paraphrase of the third quote above (namely, knowledge of psychophysical laws would shed light upon logic's theory). Of course Peirce assumes that "all attempts to ground the fundamentals of logic on psychology are seen to be essentially shallow" (5.28, 1903), but this does not mean that psychological information may not be useful for classifying arguments as valid or invalid (against Hookway 1992: 16-17).

Peirce's declarations on psychology do not sound accidental. It is as if commentators decided that either Peirce kept on the right, anti-psychologistic track, or sometimes unfortunately slipped aside. But rather than accidents, those so-called falls may be symptoms of a naturalistic framework yet coherent with his unpsychological view of logic nonetheless. Indeed, throughout the years, Peirce's observations build something essential, if not to his conception of logic, to his overall project as a logician, a scientist and a metaphysician, that is to say essential to his conception of the philosopher's task. For if he were as clear with his practice as he seems to be in his claims about the independent status of logic, why would he be flirting so dangerously with naturalism ? For instance, it is beyond doubt that physiology provides him with a strong model: the general science of signs is a "physiology of forms" (MS 478, 1903), just as "Psychology Proper" relates to a kind of "physiology of the mind," "meaning an account of how the mind functions, develops, and decays, together with the explanation of all this by motions and changes of the brain" (8.303, 1909). The phrase "physiology of the mind" (e.g. MS 741, c.1867; 1.579, 1902) was very common in the 19th century, and generally implies a causal account of the mental processes that justify our knowledge, in the same way that Kant used to refer to "the physiology of the human understanding of the celebrated Mr. Locke"[4]. So it is likely that Peirce inherits something from a British tradition that regards psychological and even physiological analyses of the mind as tools for building an epistemology (in the contemporary sense) and (in Peirce's case) for catching the real categories of the world.

Even the normative content of ideals and the ultimate ends of mankind will be expressed in those very same terms: "That ought to be done which is conducive to a certain end. The inquiry therefore should begin with searching

[4] A ix, cf. A 85-7 / B 117-9

for the end of thinking. What do we think for? What is the physiological function of thought?" (5.594, 1903) As for the notion of habit, so fundamental to the fixation of belief theory and the pragmaticistic maxim, despite its frequent use by scholars eager to dismiss a psychologistic interpretation of the theory of inquiry, it is in fact deeply rooted in physiology too. Habit is a general rule operating in the organism (W4.249). It means that the law of habit is a law of empirical thinking (not of pure thought), and that it acts over the body, especially the nervous cells (W4.39). Not only is this notion inspired by Alexander Bain but also by an American naturalist, John Murphy, the author of *Habit and Intelligence in their Connection with the Laws of Matter and Force*. The fifteenth chapter of this work, "The Laws of Habit," establishes that all mental and motor actions are habitual (except those under the control of the will). Peirce adopts this principle that all the vital operations are subjected to a (unique) law of habit. So if habit is expected to wash belief of its psychological mud, much water will be needed.

I do not claim that Peirce yielded to psychologism, but that showing that he did not would probably require many more arguments than the commentators' minimalist defense has developed so far. The program just sketched would not deserve an article but a whole book. That is why my strategy will be different. The rest of the paper will not examine the seeming "accidents of Peirce's anti-psychologism" that occurred between the early unpsychological view period and the late normative sciences period, but wishes to compare those two periods in and of themselves. As intimated above, the strategy of Peirce's readers who did not focus on his belief-and-doubt-based conception of logic was to insist on his explicit definition of logic as an unpsychological, normative science. The point I will be developing from now on is that Peirce's unpsychological conception of logic is *not* the same as his normative conception of logic.

2. Peirce's non-normative conception of logic

The two following questions need a distinct treatment. First, did Peirce sin against anti-psychologism? And second, did he argue for anti-psychologism? The second question may be made more precise: did he simply advocate anti-psychologism, or did he propose arguments to prove that psychologism was wrong? Having sketched out the requirements of a proper answer to the first question, I will now linger on the second one.

Let us start with Peirce's early conception of logic. As it has been the ob-

ject of many a study, I will recall its main characters as briefly as possible, in order to come to the very point I want to stress. The concept of logic is systematically exposed. "Logic has nothing at all to do with the operations of the understanding, acts of the mind, or facts of the intellect." (W1.164) The reason for this is that its object is not psychological thinking, but formal thought[5], "a study of forms, not a study of mind" (MS 350). Logic "does not deal with the matter of thought, but then it as certainly deals with thought as having matter that is as being a representation –true or false" (MS 741, 1864). Kant failed to understand that one never studies the abstract logos: thing and form are always known together through a representation, which, contrary to the *Vorstellung*, is not necessarily mental (W1.257). In other words, according to the unpsychological view, the laws of logic "apply not merely to what can be thought but to whatever can be symbolised in any way" (MS 340). Logic is "the science of the conditions which enable symbols in general to refer to objects" (W1.175).

Thus, the task of logic is to discover and present the forms of possible symbolization. Logic reveals and describes the various ways a symbol[6] can refer to an object. As a consequence, logic is not prescriptive, a prominent point that very few scholars have stressed[7]. It would indeed be nonsensical to command signs to respect some rules, or to force them into some particular behaviors. The laws of logic are not positive laws (in a juridical sense), but known by observation. Thus, the science of logic describes states of affairs, namely the real relations within ideal thought. It means that assenting to logical 'laws' or principles does not imply an action but a belief. They are not injunctions, do not convey any 'ought.' Peirce clearly states it in the first Harvard Lecture (W1.166):

> It has been supposed that the laws of logic might be broken. That they say "Thou ought" not "thou shalt," that in short they are

[5] Cf. (Colapietro 2003: 166-168).

[6] In his early papers, Peirce uses 'symbol' as a general term for 'sign'.

[7] Among the happy few, cf. (Levi 1997). It should be noticed that descriptivism in logic is not particularly original, though: for instance, it is Kant's conception too, and was interpreted as such by Hamilton and his followers, Thomson and Bowen (cf. Michael & Michael 1979: 85). I hereby oppose (Dipert 1994: 57): "In the loss or rejection of the normative approach to logic, Frege is particularly the villain here. In his shrill attacks on Boolean 'law of thought' and related attacks on psychologism, he never seems to have considered the possibility that the 'laws' in question might be normative/ legislative rather than descriptive (as in 'laws of nature'). For Kant and neo-Kantians, logical rules have a similarly mixed normative and descriptive function. There were probably few if any pure 'descriptivists' in logic and mathematics in the nineteenth century –not even Mill– and so it is difficult to see what Frege was ranting about."

> statements not of *fact* but of *debt*. But what page of man's ledger does this "ought" refer to? Thought *debtor* to what? It is impossible to say.

Two arguments are conveyed in this passage. The first one states that laws of logic cannot be violated, for it is impossible to think without them. They define the frame outside of which there is no thought. They display a field of possibilities, without forcing us into anything. In a way (and only in a way), they are like physical laws, for they are known by observation, and do not prescribe any conduct but rather describe how thought works. Drawing a difference with the laws of physics would only be a matter of generality: logic states the most general empirical formulæ for describing thought in any possible world. This absolute universality very much reminds us of Frege's own conception of logical laws, although he introduces a prescriptive element: *in a way*, geometric and physical laws are laws of thought just like logical laws, for "Any law that states what is can be conceived as prescribing that one should think in accordance with it, and is therefore in that sense a law of thought." (Frege 1893: xv) Logical laws "then only deserve the name 'laws of thought' with more right if it should be meant by this that they are the most general laws, which prescribe universally how one should think if one is to think at all."(*idem*) Indeed, you cannot express your refusal of the laws of logic without using them yourself.

The second argument mentioned above regards the use of "ought." "Here is an allusion to an entry on the debtor side of man's ledger. What is this entry? What is the meaning of this *ought*?" (W2.99) Peirce is right in registering that it expresses an obligation *toward* someone. As Wittgenstein would later confirm, the other person is obliged to do something (Wittgenstein 1984: 118). But in the case of logic, there is no "other person"! Husserl also took advantage of this semantic remark to argue against a prescriptive conception of normativity: a command presupposes an authority who issues it, and in the case of logic there is no such authority (Husserl 1970, vol. I §14). When a child ought ('*soll*', in Wittgenstein's example) to do something it means that if he does not, something unpleasant will happen. From the point of view of meaning, an ought-sentence is not complete: for instance, you say that something goes against my lying, but what is this thing? Peirce has the clear awareness that answering this demand would be meaningless for logic. Indeed, if logic were stating the principles we ought to follow, what would be the source of this 'ought'? What does the epistemological 'ought' mean? There is no reason to accept Leibniz's analysis of *debitum* as what is necessary for a *good*

man to do: deduction, induction, abduction and probability are not matters of morals (W2.100). Thus, the only conceivable answer is rationality. But it would hardly hide a vicious circle, for rationality is defined by its conformity to logical laws. And in any case it would lead to the pernicious consequence that one ought to follow the laws of logic as statements of debts and not as statements of (ideal) facts.

To sum up, Peirce defends the idea that thought does not imply ought. But all the difficulties are far from being solved. If one now turns towards thinking and the psychology of women and men, is there no normativity? Once we have understood the principles of logic, there certainly is something that prescribes us to act as rationally as possible, and to think in a consistent way. Whatever logic may be, Peirce should admit that everybody ought to be logical, and that it is possible (and common) not to be so, as is evident not only from fools and madmen but the irrational behaviors and beliefs of every one of us.

It seems that Peirce draws too strong a separation between the objectivity of logic and the subjective attitudes of agents, making it all the more difficult to explain the rational requirements of psychological thinking. This reproach was often held against Frege, who advises: "Always separate sharply the logical from the psychological, the objective from the subjective" (Frege 1884: x). How, then, to justify the idea that we ought to be logical nevertheless? The only possible answer is to deduce a rule of behavior from the objectivity of logic, in other words, to draw a rational 'ought' from the logical 'is'. This however is a flagrant naturalistic fallacy. Will not Peirce's head be cut off by "Hume's Guillotine"? If he is guilty of deducing a prescription from a fact, so are Frege and Husserl, as shown by (Philipse 1989: 58-59). Indeed, their anti-psychological position did not spring from tracking the naturalistic fallacy down, but on the contrary, from deriving logical norms from non-normative propositions.

At issue here are the subtle relations between objectivity, normativity, rationality and (moral) prescription. In fact, drawing on a distinction hinted at by (Korsgaard 1996), one should perhaps say that Frege and Husserl did not derive logical but *rational* norms, for the kind of necessity embodied in the application of 'ought' to our beliefs is *rational* necessity[8]. Does Peirce derive rational prescriptions from logical norms as well? He stresses the objective vs. subjective dichotomy previously mentioned to the point that it is not clear if such a "naturalistic deduction" is still possible. If Peirce cannot use Frege and

[8] Cf. (Korsgaard 1996: 226, n11).

Husserl's model, how should our necessity to act and think in a rational way be accounted for?

Peirce did overtly consider the question: "Why ought we to be logical?" (W1.166) But his solution may disappoint the reader. In the prescription to conform to logical rules, the origin of the 'ought' is to be found, he states, in the fact that "we wish our thoughts to be representations or symbols of fact." (*idem*) In the course of his characterization of logic, what Peirce intends to emphasize is that the true objects of logic are signs. But for our purpose, the relevant element is a "wish for rationality". This is no satisfactory response for explaining our rationality, and sounds like a convenient phrase rather than a "self-controlled" theory. I will come back to this matter in the last section of this article.

For now, it might be an indication that our question, the justification of the rational 'ought', does not have an answer. But let us not block the road of inquiry. It may also signify that the question is meaningless: the distinction between the 'is' of logical laws and the 'ought' of rational principles would eventually be misleading, because it introduces a difference between ideal, objective symbols on the one hand, and our thoughts on the other hand, which have indeed a psychological, that is factual and subjective, component, but *can* and *should* be viewed as symbols as well. But even in this latter interpretation, Peirce skates over the matter a little too neglectfully: how does the transition from psychology to symbolics operate?

In order to clarify the relation, if any, between logical forms and rational thinking, it could be helpful to locate *the normative* in Peirce's early texts. There are four possible options to be examined: normativity may lie in logical laws, in the rational 'ought', in both, or in neither. Let us review them. The first answer seems to be adopted (at least implicitly) by (Michael & Michael 1979: 88 Fn9), but is explicitly denied by Peirce himself: the idea that logic is composed of normative laws is false (W1.166; W4.378). The second answer, namely the normativity of what Peirce takes for an "ought to be logical" – quite misleadingly, for it should rather be called an "ought to be *rational*"– seems quite natural. The problem is that rationality does not depend on features of the world but on the relations amongst an agent's mental states and their contents, whereas norms are dependent on (ideal or actual) facts. In other words, "the rational supervenes strictly on the mental, and this is not the case for the normative" (Reisner, forthcoming). This argues for a strong distinction between rationality and normativity, which needs a more objective

base[9]. Having the normativity of rational prescriptions bear on our mental states would be a form of psychologism too. In consequence, the third answer to the question about normativity, namely that it lies both in objective logic and in human rationality, is excluded. There remains only the proposition that logic is no more normative than the prescriptions to follow its rules.

Is Peirce's early unpsychological view of logic and rationality norm-free? If not, where is normativity hiding? The case is not desperate. It is not because rational prescriptions are not normative that they do not either entail or have an impact on some norms. The entailment solution is adopted by Derek Parfit, a supporter of the distinction between rationality and normativity. He warns that we must be careful to distinguish between the view that rational requirements are themselves special instances of 'oughts' or reasons and the view that they give rise to 'oughts' or reasons. The latter view is the most plausible. For him, our rational faculty gives birth to some imperative judgments. It opposes a form of internalism according to which there is some intrinsic relation between norms and wise choices or apt beliefs. But an alternative to both (moral) internalism and Parfit's radical externalism could be the "imperativist" view that normative judgments are or involve imperatives. In the latter case, the requirements of rationality expressed by Peirce would not be norms but *consequences* of certain normative, non-logical judgments.

What are those mysterious judgments, if they exist? The young Peirce does not provide any substantial clues. We nonetheless find an indication in a famous text comparing man and word. Therein, morality is said to be "the conformity to a law of fitness of things, -a principle of what is suitable in thought, not in order to make it true but as a prerequisite to make it spiritual, *to make it rational*, to make it more truly thought at all" (W1.496, my italics). Why ought we to be logical, then? Because in being so, our thought comes closer to its law of internal determination, its proper use. This latter law acts as a kind of meta-prescriptive principle, which is itself normative: we ought to be rational because there is a norm of good thinking which requires us to be so. This morality of thought, so to speak, is analogous to the functioning of a grammar. And indeed, after beauty and truth, "(t)he third excellence is morality on the one hand, Grammar on the other" (*idem*). Thus, our rational imperatives would come from some normative grammatical-like commands.

[9] This is originally attributable to (Parfit 2001), who argued that the requirements of rationality do not have the force of normative reasons or 'oughts'.

Such a comparison may have been inspired by Kant, who states that the rules of the understanding may be compared to those of a grammar[10]. Logic is almost grammatical in the sense of providing formal rules to our inferences. How far should the parallel between grammar of thought and morality be taken? Peirce does not place much stress on it in his first period. This may be all the more surprising since its probable Kantian source is generally held to stress the role of categorical imperatives. The understanding is the best candidate for giving its duties to thought, and Peirce could easily have followed such a conception to show that all signification bears on some laws of conduct, and that mental normativity depends on certain theoretico-practical rules expressing a "You must".

However, such a 'pragmatism,' in the sense of a doctrine conflating theoretical prescriptions on practical norms, is not truly Kantian enough to satisfy Peirce, for at least two reasons. First, Kant was always careful not to infer the deontic character of epistemic norms from the legislative abilities of our understanding. Logic is normative according to him, but its norms are not prescriptive by themselves: their normativity stems from our becoming reflexively aware of logic's rules. That is how natural, *a priori* rules of the understanding become necessary laws of its conformity to itself (cf. Anderson 2005). Second, and most importantly, Kant never based theoretical norms on morality. Logic is not originally prescriptive, nor does it become so in being inferred from moral imperatives[11]. In the 1860s, and apparently until the 1880s at least, Peirce agreed on such a strict separation between the rational 'ought' and moral prescription. And, as previously explained, Peirce also maintains a strict separation between logic and our theoretical duties. Being *sui generis*, the theoretical 'ought' therefore becomes unintelligible. Peirce cannot break the deadlock without revising his whole conception.

3. Peirce's normative conception of logic

The picture so far is as follows. In his early years, Peirce insisted that logic as a theoretical science is descriptive of the ideal categories of being, and not normative. On the other hand, we ought to behave logically, that is, our thought

[10] Cf. (Jaesche 1819: 11-12).

[11] Thus, for Kant, logic neither describes a fact of psychology nor a categorical imperative, contrary to what (Haack 1978: 238) surprisingly claims: according to Kant, "logic is descriptive of mental processes (it describes how we *do*, or perhaps how we *must*, think)". Literally, in logic, the inquiry is after "not how we think, but how we are to think" (Jaesche 1819: 14).

is subjected to some requirements of rationality. But this 'ought' is not normative either, at least if we take normativity in its stronger sense. Nevertheless, this rational prescription seems to have been vaguely conceived as a consequence of some normative grammar of thought.

Over the years, and under the pressure of many other parameters including the development of experimental psychology, the birth of phenomenology and a better understanding of the mathematical continuous, Peirce's systematic conception of the relations between normativity, rationality, morality and logic evolved and were clarified. This allows us to overcome some shortcomings of his early positions.

First, Peirce shows a manifest interest for the notion of norm and normativity. The word *normative* –"Überweg's adjective" to designate what is 'directive' (2.7, 1902)- was invented in the school of Schleiermacher (2.575, 1902), and Lalande's dictionary attributes its introduction into common speech to Wundt (2.7 Fn1, 1902). It was first used in the context of the German *Psychologismus-Streit*. Whatever its origin, and however short it may fall of being "particularly pleasing" on philological grounds, "the twentieth century would laugh at us if we were too squeamish about the word's legitimacy of birth" (*idem*). Nevertheless, Peirce would still be using 'critical' as a synonym for it as late as 1909 (EP 2.459).

To cut a long story short, Peirce now considers that logic is normative. In the first years of the 20th century, this is integrated in his complete classification of sciences, at the heart of which lie the normative sciences. One can legitimately wonder why Peirce changed his mind, and what are the consequences of this on our epistemic duties. Strikingly, his last conception of logic has been widely viewed as mirroring the early one, despite a complete shift. Explaining the reasons of this change would take us too far away from our stride[12]. The second question is closer to our purpose. The apparent consequence of acknowledging the normativity of logic should be to solve all the perplexities about our rational conduct: logic, with all its normative power, forces us to act according to its rules.

Unfortunately, this is not Peirce's conclusion. For he is still attached to the motives that used to prevent him from reducing logic to prescriptions: logic is not a set of principles describing a possible way of reasoning, but a statement of the conditions of reasoning *in general*. He even most paradoxically extends his anti-prescriptivism to ethics, which should not stick "to the obsolete pretense of teaching men what they are 'bound' to do." (EP 2.459, 1909) If they

[12] For several reasons, cf. (Burks 1943: 189).

are really bound to act in certain ways, people have no choice, and need no science to teach them what to do! In this case, what is required is only a clear acknowledgment of the facts of the matter and how categories and essences are related to each other. In the same spirit, logic does not command what one ought to think:

> Logical treatises never say anything about what 'ought to be thought' as long as there is any compulsion of thought or reflection. In those cases they only speak of how the facts are. (2.50, 1902)

So norms should not be understood as guides (in the sense of a "leading principle" of inference). Rather, they are aims. Understanding this distinction requires us to come back to the semantic analysis of 'ought'. In addition to its connotations of debt, it also evokes an aim or an end. Peirce even decides that such a teleological sense is the only way to save it from nonsense: "The question is what theories and conceptions we ought to entertain. Now the word 'ought' has no meaning except relatively to an end. That ought to be done which is conducive to a certain end. The inquiry therefore should begin with searching for the end of thinking." (5.594, 1903) And as Peirce would develop at length, the science of such ideals is esthetics. Thus, Peirce's agreement with the normativity of logic forbids it to be prescriptive but does imply that it is based on moral.

Once more, a comparison with Husserl may clarify the situation. The latter, in his contemporary *Logical Investigations* (1900-1901), pays lip service to a normative view of logic. For him like for Peirce, the first requirement of a theory of logic is to preserve the latter's objectivity and the ideality of its objects. To this purpose, Husserl needs to defend a dualistic conception of logic, as a normative science, and as a purely theoretical discipline which provides the former with its theoretical basis. He attempts to define normative requirements in terms of value judgments. To him, there are some essentially normative sciences relying on evaluative definitions, sciences to which logic only belongs in part. But "[t]he issue is not whether logic is a normative discipline, but what kind of science provides normative logic with its theoretical basis," Philipse (1989: 62) writes. To be true, logic is normative, but the *Prolegomena to Pure Logic* show that more essentially it is a theoretical, non-normative science. Normative logic, as a practical technology, is founded on the *a priori* science of pure logic. The latter itself rests on a *Grundnorm*, which needs to be the logical excellence of correctness in inference. This conception presupposes that there is a basic norm telling us to reason correctly, which is completely separated

from the theoretical content of the norms of logic. Conversely, the laws of pure logic in themselves are free from any normative connotation.

The limits of this Husserlian conception can by contrast show Peirce's relevance. One such limit is that nowhere does Husserl characterize logic's *Grundnorm*, that is, the "good from the logical point of view." A definition of the aims of logic is lacking. It is likely, of course, that normative logic essentially consists of the norms for valid deductions, namely to infer correctly, that is in non-evaluative words, to conduct truth from premises to conclusion.

As for Husserl's dual view of logic, it opposes a normative, practical technology with a pure logic expressible in completely non-evaluative terms. Peirce's early view of logic was not dual: it was strictly non-normative, even though we should follow principles of rationality. Peirce's new conception of logic seems unitarily normative, though not practical. Contrary to Husserl, it is not logic as an art but logic as a theory which is normative: like aesthetics and ethics, logic is a purely theoretical science which nevertheless sets up norms (2.156, 1902). (Mullin 1966) failed to understand that 'normative' does not have the same meaning for the two authors: the "obstetric methods" (as Mullin rightly says) of theoretical logic is normative according to Peirce in providing ideals for practice, while Husserl calls practice normative because it is *subjected* to norms. In other words, whereas Husserl sees logical norms as rules for our factual reasoning, Peirce conceives them as ideals, strictly separate from our subjective thoughts. That is why he endeavours to investigate on logic's '*Grundnorm*', the source of its fundamental 'goodness', a point missing in Husserl because it only pertains to practical logic according to him.

But on the other side, Peirce's position suffers from a major flaw: it runs the risk of inconsistency, for normative logic, as a theoretical science, is supposed to assert non prescriptive norms ('normative' having indeed the advantage over 'directive' of avoiding an apparent implication "that logic is a mere art, or practical science," 2.7 Fn1, 1902), while its ideals obviously are the norms that guide our thinking. What indeed would be the principles that we ought to observe, if not the norms of logic? It means that there must be a practical side of logic governed by the norms of its theoretical side. Peirce's first conception actually left the art of logic totally in the shade. Logic as an art (Aristotle's "organon," though he himself did not consider it an art but a science) was adopted by most logicians, from the Stoics to the British old rock logicians, *via* the scholastic doctors opposed to Duns Scotus (MS 606, 1906). Such a view particularly displeases Peirce for its essential psychologistic tone: if logic applies principles from another science, it can only be from psychol-

ogy[13].

Overcoming his reluctance, Peirce was urged by his theory of inquiry and his method of fixation of beliefs to regard logic as the art of finding methods of research (W4.378). He eventually found a way to incorporate the practical side of logic without reducing it to an art, thanks to a distinction between practical science and art paralleling the *praxis* versus *poiesis* dichotomy: The latter teaches us to make something, whereas the former only teaches us to act or do something. (MS 607, 1906) There should admittedly exist a practical science, or rather a group of a least twelve distinct sciences, following the principles of methodeutics (MS 603, 1906), but one may wonder whether those sciences still belong to logic.

Thus, like Husserl, Peirce is implicitly committed to a twofold view of logic, which is both a theoretical and practical science. Does it quite nearly replicate the distinction between *logica docens* and *logica utens*, which was invented a few years only before Peirce's system of normative sciences? In this case, Peirce would perhaps agree with Husserl's view "that it is the true sense of our supposed pure logic to be an abstract theoretical discipline providing a basis for a technology [...], its technology being logic in the ordinary, practical sense." (Husserl 1970: 80) And indeed, "A normative science is by no means an art, although it ought to inspire and inform an art" (MS 602, 1906). But contrary to Husserl he would probably not give the name 'logic' to this technology or art: it is barely a handy set of rational principles. That is why *logica utens*, though nearer to actual practice, is not an art of reasoning. It is normative and deals with the distinction of the true and the false (5.108, 1903). But neither is it to be identified with *logica docens*. Therefore, the usefulness of the two *logicæ* is not obvious. Is it that *logica docens*, as a formal, (often symbolic) enterprise, deals with ideal facts? A positive answer would forbid it to be ruled by norms, whereas a negative answer would leave the following question unsolved: what happened to logic as a study of pure forms? Either *logica docens* is normative, and there is no use distinguishing it from *logica utens*, or it is not, which means that logic is not essentially normative, and then the whole system of the normative sciences collapses.

Peirce's stance in this alternative may seem ambiguous. For, to sum up, on the one hand logic, ethics and esthetics are unquestionably normative, in the sense of providing ideals and ends for our actions; but on the other hand,

[13] Cf. (Mill 1865: 359). In that sense, it is true that Peirce's (and Husserl's) attitudes are very similar to psychologism, because they just substitute one science (pure logic) with another (psychology) as the basis of practical logic (Husserl's normative technology).

Peirce has not got rid of his opposition to a prescriptive view of normativity. And third, norms in the sense of final ends can hardly be anything other than imperatives or prescriptions. The key might be found in applying the *utens* vs. *docens* dichotomy to this last point. *Logica utens* comprises the system of norms that we spontaneously use in thinking, without particular investigation on the ends we pursue. "Every agent has a generalized ideal of good reasoning and what is not. We carry more or less distinctly in our minds patterns of good and bad reasoning, which may be called Norms" (MS 453). The texts nevertheless lack clearness, for we also read that *logica utens* is not "subject to any normative laws" (2.204, c. 1901-1902). It is "neither good nor bad; it neither subserves an end nor fails to do so" (*idem*). This apparent contradiction will hopefully be clarified by the rest of the present section and the next one. Roughly, it consists in showing that *logica utens* really *is* normative (against Pietarinen 2005), and as such is neither good nor bad, nor is it guided by *laws*, but just states what is correct thinking, although it is not self-controlled: self-control indeed guarantees normativity, but there also is another form of natural, innate norms.

Logica utens is based on prima facie ends, while *logica docens* depends on all things considered ideals, to use (Ross 1930) famous distinction. An agent's prima facie obligations may conflict, but not her actual, reflected ideals. It is when facing a "normative conflict" (to use a non-Peircean phrase), that is when experiencing a doubt on which (practical or theoretical) option to take, that we do not meet any (normative) compulsion anylonger. Then does the 'ought' find room (2.50, 1902). Every normative science supposes such a dual distinction between a 'may' and an 'ought not'. Thus, logic is a science, but not a science of what is, not yet of what might conceivably be, but of something between these two (MS 602, 1906). This modality between possible firstness and existing secondness typically is the mode of reflection and thirdness. Indeed, it is only when a possible choice happens that the norms we want to follow are in need of clarification, and that we ourselves can decide to follow such or such ends. Then we slide into the uncertain domain of the oughts, and deliberately act according to ethical principles or not. For "the moralist, as far as I can make it out, merely tells us that we have a power of self-control" (1.611, 1903).

From this point, one can imagine two scenarios: either a reflexive inquiry about norms closes the issue, and in ascending (so to speak) from *utens* to *docens* we reach normative logic back again, or even the principles of formal logic are mute on the subject, and we have no solution but try and do what we think we ought to do. The example Peirce gives, the property of elegance of

a system, is enlightening: "we are told that we ought to try simple hypotheses before complex ones." (2.50, 1902) What is the rationale of that sort of Ockhamian principle? Some have tried to show that it is an ultimate a priori epistemic principle that simplicity is evidence for truth. But it does not seem to have been very promising so far, and it is likely that "Just as the question 'why be rational?' may have no non-circular answer, the same may be true of the question 'why should simplicity be considered in evaluating the plausibility of hypotheses?'" (Sober 2001: 19). Nevertheless, it would be for two very different reasons in Peirce's system: the privilege of simplicity would belong to the prescriptive, non normative, uncertain realm of the oughty attempts to achieve what is best, whereas being logical is an ultimate ideal, a norm. In other words, whereas logical rules (like *modus ponens*, conditional proof, universal generalization, etc.) express permissions, the rules containing the language of obligation (e.g. "one must not carry out an inferential step in a deductive system unless it is permitted by one of the rules of the system"; "sets of propositions in a system should be consistent," etc.) are better counted as metalogical principles (cf. Resnik 1985: 236).

The answer to the previous alternative is now obvious: *logica docens* cannot be but normative. What then is the point in contrasting it with *logica utens*? Their difference should probably not be emphasized too strongly: they are not two kinds of logic, only two modes of considering it, as a system of intuitive normative requirements or as a reflection on the ideals of thinking[14]. Do we really need to ask ethics and eventually aesthetics for the ultimate ends in order to be logical? Is not truth a value obviously good enough not to search beyond? In short, is not Peirce's pyramid of the normative sciences a little too aesthetical itself and formal, and mostly vacuous? Peirce faces the objection:

> What, then, is our ultimate aim? Perhaps it is not necessary that the logician should answer this question. Perhaps it might be possible to deduce the correct rules of reasoning from the mere assumption that we have some ultimate aim. But I cannot see how this could be done. (1.611, 1903)

Peirce fears that a vague conception of norms be too weak to prevent us from, for instance, living a life of pleasure which would have us regress to an illogical state. It supposes holism about ideals. There lies the interest of Peirce's

[14] I would not be inclined to call them two 'faculties' as (Pietarinen 2005) does. He tends to present *logica utens* as a kind of rational non-normative instinct, and *docens* as a normative classification of arguments. But as shown above, it is not consistent with all the texts.

two-facet logic: in shifting from *logica utens* to *logica docens*, one criticizes one's own aims and earns more self-control over oneself. It is precisely this kind of reflexive procedure which characterizes rationality –but not normativity strictly speaking, because objective ideals do not need any critical thinking to stand by themselves. (This is why Peirce paradoxically insists on *logica utens* belonging to morality (5.108, 1903), while it is mostly *logica docens* which is self-controlled.)

I can now return to the difficult question of the reasons of Peirce's shift from a non-normative to a normative conception of logic and attempt to draw some conclusions. It seems that the normative view was (paradoxically, at least for our modern minds) rejected because of its collusion with psychologism: it is an art that gives rules and directions to our reasonings. That is why Mill enthusiastically adopts a normative conception of logic:

> Logic is not the theory of Thought as Thought, but of valid Thought; not of thinking, but of correct thinking [...]. Logic has no need to know more of the Science of Thinking, than the difference between good and bad thinking [...]. The properties of Thought which concern Logic, are some of its contingent properties; those, namely, on the presence of which depends good thinking, as distinguished from bad. (Mill 1865: 460)

For Mill, logic expresses what *must* be thought according to certain laws. But that (empirical) psychology is able to provide such laws is doubtful (2.50, 1902). Conversely, Peirce first considers that logic has no normative or evaluative function: it *is*, that is all. But the need to account for our rational duties eventually forces him to endorse a normative view of logic. However, contrary to Husserl, he did not have an originary dual conception of logic, so that, instead of putting normativity in its 'technological' part, that is, on the practical side, as Husserl does, he feels compelled to locate it into its only true logical content. As a consequence, theoretical logic is viewed as providing norms of correctness to our reasonings, though not prescribing anything. Those logical norms are themselves oriented by moral and 'aesthetical' ultimate duties. Logic still is a science of pure form without paradox: in displaying the norm of correctness, it provides a classification of formally sound arguments. A normative science is indeed classificatory, so that Peirce's early concept of logic as a science of classification survives even in his last period[15].

[15] I disagree with (Short 2007: 63) that logic "is normative and is not a study of pure form". Around 1902, Peirce still thinks that "[t]he only concern that logic has with this sort of [mathe-

Logic only observes what should be the relation of a fact, however it can be thought, to another fact, for the truth of the first one to imply the truth of the second (MS 603, 1906). It studies the conditions of *truth*, this kind of excellence which can or cannot belong to the objects considered as representing real objects (HPPLS II, 826, 1904). Logic is not interested in the "psychological dresses" of thinking, except if it allows the discovery of formal equivalences under superficial dissimilarities (N3.298, 1908). It is not even interested in the forms of valid human reasoning, because it would become a natural history of thought –Dewey's program (8.239-242, 1904). So why ought we to be logical? Because correct reasoning "consists in such reasoning as shall be conducive to our ultimate aim" (1.611, 1903).

In short, to the very difficult question of the rationale of rationality for a theory rejecting prescriptive norms, Peirce gives an extremely simple answer: norms state ideals of conduct (respectively, correctness, goodness and ultimate perfection for logic, ethics and aesthetics) and as such do not prescribe anything. As for our reasonings, they are directed toward those ends: we do not have abstract duties toward ideals, but ought to be rational and good *only insofar* as we *actually do* aim at realizing such ends.

4. Why Peirce may derive 'ought' from 'is'

The previous section showed Peirce's elegant solution for keeping away from psychologism without committing himself to an insuperable objective vs. subjective (Frege-like) or ideal vs. factual (Husserl-like) dichotomy. According to authorized scholarship, Frege and Husserl adopted such a (self-destructive) strategy because of their psychological conception of the mental. The *psychical*, characterized by final causality (1.253, 1902), is to be distinguished from the psychological (5.485, 1907).

> Many persons, perhaps most persons have the idea that every observation about the human mind is a psychological observation. They might as well regard the sight or sound of an apple dropping from a tree as an astronomical observation in view of what is said to have befallen Isaac Newton. (MS 614, 1908)

Despite his anti-psychological slogans, Husserl himself applies logic to the human psychological being "*as we find it*" (4.7, 1906), Peirce regrets. The former thus conflates the mental and the psychological, for his phenomenology

matical] reasoning is to describe it." (2.192)

is nothing but another psychology. On the contrary, Peirce's phenomenology does not observe the same facts as psychology. "It looks upon the same world; -the same world that the astronomer looks at. But what it observes in that world is different" (8.297, 1904).

Is not Peirce's solution *too* elegant, and too simple altogether? For the suture between logical facts and following ideals is not as much explained as repelled onto a slippery slope, man's benevolence: we ought to be logical because it is the only way to reach our (logical, and eventually moral) ideals; but those ideals are not commands by themselves, they just are ultimately good, so that we love them. What if we don't? It would reveal bad dispositions: "Bad reasoning is almost as bad as bad morals" (W1.454). Peirce countenanced such a view long before his theory of normativity: to him, man should never be irrational. If representing is the act of a symbol, how is it possible that a representation be false?, he wonders (MS 921, 1860). Error must come from perversion, weakness or passion (W1.5); in all cases it is a miracle (W1.338). Labelling irrationality as supernatural reveals a conception of logicality as a natural fact. It is *normal* to be rational. The pivotal concept of normality in Peirce's early papers has not been much noticed. To a large extent, it plays the part of his absent normativity. Yet, when the normative approach is later admitted, his recourse to the normal does not cease (e.g. 1.662, 1898). Indeed, as previously established, Peirce's norms are lazy: they do not prescribe anything.

That is why logicality needs to be anchored in naturality, or more precisely in human nature: "reasoning power is related to human nature very much as the wonderful instincts of ants, wasps, etc. are related to their several natures" (MS 682, 1913). It has something of an Aquinian flavor, with the difference that it is not an essence of divine origin but a set of instincts resulting from evolution. That is why I claim against (Hookway 2000)[16] that the "non-transcendental alternative to psychologism" *is* defended naturalistically. At the heart of this issue is the possibility of a natural science of man. In avoiding the pitfalls of the *ought*, Peirce manages to preserve normativity without excluding the possibility of a naturalistic approach.

As the relations between habits, instincts and reasoning have been widely studied by scholars, it is no use dwelling on it[17]. But there is an aspect of such a natural inscription that has generally been overlooked. Instincts account for

[16] Cf. especially: 294-297 "Naturalism and the Transcendental Philosophy".

[17] Cf. (Murphey 1991), (Misak 1991), (Hookway 2000: 255), etc., and all the studies on critical common sense.

the wonderful success of many of our spontaneous guesses, not for our actual aiming at normative ends. That is why Peirce also needs a dispositional theory of human nature. Before (Roberts & Wood 2007), who seem to share Peirce's hope that "with the further development of ethics this relation [between good reasoning and good morals] will be found to be even more intimate than we can, as yet, prove it to be" (1.576, 1902), (Zagzebski 1996) was one of the first to relate Peirce's underlying but effective voluntary dispositionalism with 20th century works against Kantian deontism (cf. Anscombe 1958, Foot 2001), in showing that motivation for knowledge is not totally expressed by following reliable well-known rules of belief-forming.

Thus, a theory of epistemic virtues ultimately ties Peirce's system of norms and oughts. More precisely, it explains why we generally tend to do what we ought to do despite the non prescriptive character of norms: such a tendency is inscribed in our nature, not only as attuned to the laws of the world through induction and abduction, but as driven by moral virtues. In sum, Peirce's theory of epistemic virtues answers the problem of our *access* to norms: we hold them to be our aims because we meet them in our virtuous nature. In this respect, it is not so far from Sigwart, so harshly decried for his logical *Gefühl*. (Sigwart 1889: 22) reads: despite the "normative character" that is "essential" to logic, nevertheless "we deny that these norms can be cognized otherwise than on the foundation of the study of the natural forces and functional forms which are supposed to be regulated by those norms." Peirce's parallel position is no form of psychologism, but shows that a normative theory may and must be deeply rooted within nature.

*

I wish to draw two conclusions from this study. The first one is about Peircean scholarship. As Peirce's first and last texts assert a strong claim against psychologism, it has widely been held that his late theory of logic was the one he defended in the 1860s, and that the commentator's task was to discuss the disturbing theory of inquiry within this anti-psychological framework. My purpose has been to show that such an analysis is fiction. Contrary to the normative sciences period, Peirce's first unpsychological view of logic is not normative, and fails to explain why we ought to be logical. Using Bernard Williams' typology, one could say that Peirce's thought developed from Kantian internalism to Lockean externalism. Peirce's internalist period contends that, independently from practical prescriptions, there is an internal contradiction in not looking for truth, because logic describes the laws of good think-

ing. In his normative, externalist period, Peirce would rather regard the search for truth as a moral duty, an ethical rule of belief. In the beginning, error is a miracle; in the end it is a sin against our human dispositions. One could emphatically speak of a translation from a German Peirce (under the influence of Kant, Fechner, Helmholtz) to a British Peirce (following Reid, Bain and Darwin).

My second conclusion is about psychologism and naturalism. In the present philosophical context, the two main traditions of the 20th century, idealistic phenomenology and formal analytic philosophy, are out of breath. It indicates that Husserl and Frege need to make peace with modern psychology and cognitive sciences. Does it mean a revival of psychologism? Not necessarily. On Peircean grounds, Susan Haack advocates weak psychologism, which would not *describe* our actual processes of thought but *prescribe* a correct way of thinking: "Logic, I suggested, is prescriptive of reasoning in the limited sense that inference in accordance with logical principles is safe" (Haack 1978: 238). But what is invariable throughout Peirce's works is his lack of "appetite for oughty things." Even normative logic does not prescribe anything. Against Susan Haack, therefore, I argue that it is not a matter of psychologism but of naturalism, which is very different: we need to take into account some data inscribed in our human nature to make good epistemology. In this respect, Peirce has still much to teach to contemporary philosophy.

5. References

Anderson, R.L., 2005, "Neo-Kantianism and the Roots of Anti-Psychologism," *British Journal for the History of Philosophy* 13: 287-323.

Anscombe, Elizabeth, 1958, "Modern Moral Philosophy," *Philosophy* 33(124): 1-19.

Burks, Arthur W., 1943, "Peirce's Conception of Logic as a Normative Science," *Philosophical Review* 52(2):187-193.

Colapietro, Vincent, 2003, "The Space of Signs: C.S. Peirce's Critique of Psychologism," in D. Jacquette (ed.): 157-179.

Dipert, Randall R., 1994, "Peirce's Underestimated Place in the History of Logic: A Response to Quine," in Kenneth L. Ketner, ed., *Peirce and Contemporary Thought: Philosophical Inquiries*, New York, Fordham University Press: 32-58.

Dougherty, Charles John, 1980, "C. S. Peirce's Critique of Psychologism," *Two Centuries of Philosophy in America*, Peter Caws (ed.), Oxford, Basil Blackwell: 86-93.

Foot, Philippa, 2001, *Natural Goodness*, Oxford, Clarendon Press.

Frege, Gottlob, 1884, *Die Grundlagen der Arithmetik*, Breslau.

Frege, Gottlob, 1893, *Grundgesetze der Arithmetik*, Jena, Verlag Hermann Pohle.

Goudge, Thomas, 1947, "The Conflict of Naturalism and Transcendentalism in Peirce," Journal of Philosophy 44(14): 365-375.

Haack, Susan, 1978, *Philosophy of Logics*, Cambridge, Cambridge University Press.

Hookway, Christopher, 1992, *Peirce*, London, New York, Routledge.

Hookway, Christopher, 2000, *Truth, Rationality and Pragmatism*, Oxford, Oxford University Press.

Hookway, Christopher, 2010, "Normative Logic and Psychology: Peirce on Dewey," manuscript.

Husserl, Edmund, 1970, *Logical Investigations* (trad. Findlay), London, Routledge.

Jacquette, Dale, 2003, *Philosophy, Psychology, and Psychologism: Critical and Historical Readings on the Psychological Turn in Philosophy*, Springer.

Jaesche, Gottlob, 1819, *Logic of Emmanuel Kant*, London, Simpkin & Marshall.

Kasser, Jeffrey, "Peirce's Supposed Psychologism," *Transactions of the Charles S. Peirce Society*, 35(3): 501-526.

Korsgaard, Christine, 1996, *The Sources of Normativity*, Cambridge, Cambridge University Press.

Kusch, Martin, 1995, *Psychologism: A Case Study in the Sociology of Philosophical Knowledge*, London, Routledge.

Lane, Robert, 2009, "Persons, Signs, Animals: A Peircean Account of Personhood," *Transactions of the Charles S. Peirce Society* 45(1): 1-26.

Levi, Isaac, 1997, "Inference and Logic According to Peirce," in *The Rule of Reason: The Philosophy of Charles Sanders Peirce*, J. Brunning and P. Forster (ed.), Toronto, University of Toronto Press: 34-56.

Michael, Emily & Frederick Michael, 1979, "Peirce on the Nature of Logic," *Notre Dame of Formal Logic* 20: 84-88.

Mill, John Stuart, 1865, *An Examination of Sir William Hamilton's Philosophy*, Boston, W. V. Spencer.

Misak, Cheryl, 1991, *Truth and the End of Inquiry: A Peircean Account of Truth,* Oxford, Clarendon Press.

Mohanty, Jitendranath, 1985, *The Possibility of Transcendental Philosophy*, Dordrecht, Martin Nijhoff.

Mullin, Albert, 1966, "C. S. Peirce and E. G. A. Husserl on the nature of logic," *Notre Dame Journal of Formal Logic* 7(4): 301-304.

Murphey, Murray, 1991, *The Development of Peirce's Philosophy*, Cambridge, Harvard University Press.

Parfit, Derek, 2001, "Reasons and Rationality," Egonsson, Dan, et al., *Exploring Practical Philosophy: from Action to Values*, Aldershot, Ashgate.

Peirce, Charles Sanders, 1931-5, *Collected Papers of Charles Sanders Peirce*, C. Hartshorne and P. Weiss (ed.), Cambridge, Harvard University Press, vol. 1-6.

- 1958, *Collected Papers of Charles Sanders Peirce*, A. Burks (ed.), Cambridge, Harvard University Press, vol. 7-8.

- 1982-2010, *Writings of Charles S. Peirce: A Chronological Edition*, Bloomington & Indianapolis, Indiana University Press, vol. 1-6, vol. 8.

Philipse, Hermann, 1989, "Psychologism and the Prescriptive Function of Logic," in Mark A. Notturno (ed.), *Perspectives on Psychologism*, E.J. Brill, Leiden: 58-74.

Pietarinen, Ahti-Veikko, 2005, "Cultivating Habits of Reason: Peirce and the *Logica Utens* Versus *Logica Docens* Distinction," *History of Philosophy Quarterly* 22(4): 357-372.

Rath, Matthias, 1994, *Der Psychologismusstreit in der deutschen Philosophie*, Freiburg, Karl Alber.

Reisner, Andrew, (forthcoming), "Is There Reason to Be Theoretically Rational?" in Andrew Reisner & Asbjørn Steglich-Petersen (eds.), *Reasons for Belief*, Cambridge University Press.

Resnik, Michael, 1985, "Logic: Normative or Descriptive? The Ethics of Belief or a Branch of Psychology?," *Philosophy of Science* 52(2): 221-238.

Roberts, Robert C., & W. Jay Wood, 2007, *Intellectual Virtues: An Essay in Regulative Epistemology*, Oxford, Clarendon Press.

Ross, W.D., 1930, *The Right and the Good*, Oxford, Clarendon Press.

Short, Thomas, 2007, *Peirce's Theory of Signs*, Cambridge, Cambridge University Press.

Sigwart, Christoph, 1889, *Logik. Erster Band: Die Lehre vom Urteil, vom Begriff und vom Schluss*, Freiburg i.B.: J.C.B. Mohr. 2nd ed.

Sober, Elliott, 2001, "What is the Problem of Simplicity?" in Zellner, A.et al., *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*, Cambridge, Cambridge University Press: 13–31.

Zagzebski, Linda, 1996, *Virtues of the Mind: an Inquiry into the Nature of Virtue and the Ethical*, Cambridge, Cambridge University Press.

41

Making Rules Explicit and Following Them *

MATHIEU MARION AND MITSUHIRO OKADA

Let us begin with a little story about logic in Ancient Greece.[1] In *Lesser Hippias*, Socrates is arguing with the eponymous Sophist over his claim that Achilles is better than Odysseus, although the latter is the 'wiliest', in the sense that he lies intentionally, something that Achilles is incapable of. At the conclusion of the dialogue, Socrates will eventually make Hippias concede that someone who does evil intentionally is better than someone who does it unintentionally, and thus that Odysseus, who, as opposed to Achilles, tells falsehoods intentionally, is the better of the two. Towards the middle of the dialogue, at 366c-369b, Socrates is merely trying to elicit from Hippias that the same person can either tell the truth or tell lies, so that, eventually, Hippias will have to concede that Achilles, if he is truthful, also tells lies, and Odysseus, if he tells lies, is also truthful. To do so, Socrates argues for a first case asking Hippias, who was in his days reputed mathematician, if he is experienced at arithmetic and, if so, if he would have 'the most power' at telling the truth about ordinary arithmetical calculations. Upon securing Hippias' agreement, Socrates gets him to agree further that, given his ability to tell the truth, he would be equally able to tell arithmetical lies consistently, while an incompetent arithmetician would simply state arithmetical falsehoods unintentionally. Hippias has then to concede that the same person, namely himself, has "the most power to lie" and "to tell the truth about calculations". The dialogue

*We would like to thank Crispin Wright for conversations on the topic of this paper.

[1] This 'story' is taken from (Marion & Rückert unpublished).

then goes with Socrates making the same point, in the same manner, about geometry and astronomy. Having thus secured Hippias' consent to three cases, Socrates then introduces the universal proposition, immediately asking for a counterexample:

> SOCRATES: Come then, Hippias. Examine all the sciences similarly. Is there any that's different from these, or are they all like this? [...]
>
> [...] tell me, in accordance with what you and I have agreed upon, if you find any case in which one person is truthful and another (distinct, not the same) person is a liar. Look for one in whatever sort of wisdom or villainy you like, or whatever you want to call it; but you will not find it, my friend, for none exists. So tell me!
>
> HIPPIAS: But I can't, Socrates: at least, not offhand.
>
> SOCRATES: And you never will, I think. [...][2]

Unable to provide a counterexample, Hippias has to concede to Socrates, who is thus in a position to infer to the conclusion he was after:

> [...] But if what I say is true, you will remember what follows from our argument.
>
> [...]
>
> SOCRATES: [...]You realize that you said that Achilles was truthful, whereas Odysseus was a liar and wily?
>
> HIPPIAS: Yes.
>
> SOCRATES: You are now aware, then, that the same person has been discovered to be a liar and truthful, so that if Odysseus was a liar, he also becomes truthful, and if Achilles was truthful, he also became a liar and these two men are not different from one another, nor opposite but similar.

Aristotle knew this dialogue, which he explicitly mentioned in his *Metaphysics*, Δ, 29, $1025^{a}6$-13, and he made explicit the dialectical rule that Socrates followed at *Topics*, Θ, 2, $157^{a}34$-$157^{b}2$:

[2] Translations are from Plato, *Complete works*, J. M. Cooper & D.S. Hutchinson (eds.), Indianapolis IN, Hackett, 1997.

> When it happens that, after you have induced from many cases, someone does not grant the universal, then it is your right to ask him for an objection. However, when you have not stated that it does hold of some cases, you have no right to ask 'of which cases does it not hold?' For you must previously carry out an induction to ask for an objection in this way.[3]

This rule can be analysed in terms of a game or interaction semantics.[4] For this we first note that it involves an universal affirmative proposition, of the form '*B* belongs to all *A*' or 'All *A* are *B*', established first by *apagogê*, i.e., induction. Aristotle gave a meaning explanation of the universal quantifier in *Prior Analytics*, *A*, 2, $24^{b}28$-29:

> We use the expression 'predicated of every' when none of the subject can be taken of which the other term cannot be said.[5]

To introduce a bit of type-theoretical notation, we write the universal affirmative $\Pi^{+}(A, B)$. As stated, it is to be asserted only if no A can be taken which is not B, which means that no c of type A or '$c : A$' can be found for which it is not the case that $B(c)$. And the dialectical rule states that, in a dialectical bout, if the proponent puts forward $\Pi^{+}(A, B)$ on the basis of having argued it for a number of cases (usually between two and five), then the opponent should either concede the generalization or challenge it by putting forward a possible counterexample, namely a $c : A$ which is not a B. So, keeping to the type-theoretical language, we could give the following:

Opponent	Proponent
	$a : A$ is B
	$b : A$ is B
	$c : A$ is B
	$\vdots$
	$\Pi^{+}(A, B)$
$d : A$ is not B	

[3] See also *Topics*, Θ, 8, $160^{b}1$-6. We are using here Robin Smith's translation of *Topics Books I and VIII* (Oxford, Clarendon Press, 1997).

[4] See (Marion & Rückert unpublished) for a detailed presentation.

[5] We are using here Robin Smith's translation of Aristotle's *Prior Analytics* (Oxford, Clarendon Press, 1989).

The game would then go on with the players debating if *d* is really a counterexample or not. If the proponent does not wish to concede that *d* is a counterexample, then he must either argue that *d* is not of type *A* or argue that *d* is, contrary to his opponent's claim, a *B*.

The point of telling this story is that it provides an historical example of 'making explicit' a rule which was implicitly followed by Socrates, Hippias, and, presumably, others before Aristotle wrote his *Topics*. An argument can be made that, *qua* inference rules, Aristotle's own syllogistic rules in *Prior Analytics* and the Stoics' 'indemonstrables' were, likewise, rules made explicit that were first implicit in the practice of dialectical bouts.[6] The above story thus exemplifies the idea that some *rules (including rules of inference) are followed before they are made explicit*, so it might be useful to reflect upon examples of this sort from the 'inferentialist' point of view set forth by Robert Brandom.[7] In this paper, we shall limit ourselves to two points. First, we wish to show how the above story involves a philosophical lesson also found in Lewis Carroll's 'paradox of inference',[8] and, secondly, we will argue how this point is related to Wittgenstein's well-known 'rule-following argument'. In our discussion, we will offer criticisms of Pascal Engel's views on Carroll's paradox. He has not only been a tireless promoter of analytic philosophy within the French-speaking world, he has also made significant contributions to it, in particular on the topic of Carroll's paradox.[9] Analytic philosophy certainly shares the interactive spirit of the origins of philosophy and logic in Ancient Greece, exemplified above, so it is only a fitting tribute to Pascal that we controvert his claims. There is at all events no greater homage than to consider someone's ideas worthy of serious discussion, and we are very glad to have this opportunity to celebrate Pascal's contribution to philosophy.

*

That some rules are followed prior to being made explicit sounds like a platitude. Why would it be so? Possibly because the contrary claim, namely that can only be said to follow a rule once it has been explicitly stated, would

[6] For more on dialectic in this context, see (Castelnérac & Marion 2009) and (Castelnérac & Marion 2013).

[7] See (Brandom 1983), (Brandom 1994), (Brandom 2000). See also, for an explanation of the meaning of 'inferentialism' in the context of interaction semantics, the first part of (Marion 2012).

[8] (Carroll 1895).

[9] Pascal Engel has published extensively on Carroll's paradox of inference. For this paper, we consulted mainly (Engel 1998), (Engel 2005), (Engel 2007), (Engel 2009).

sound rather odd. Indeed, this would mean, to put it crudely, that no one ever followed, say, the disjunctive syllogism *before* a Stoic logician first stated the rule. To keep with examples from Ancient Greece, Sextus Empiricus' account of Chrysippus' dog, would make no sense at all:

> And according to Chrysippus, who shows special interest in irrational animals, the dog he shares in the far-famed "Dialectic". This person, at any rate, declares that the dog makes use of the fifth complex indemonstrable syllogism, when, on arriving at a spot where three ways meet, after smelling at the two roads by which the quarry did not pass, he rushes off at once by the third without stopping to smell. For, says the old writer, the dog implicitly reasons thus: "the creature went either by this road, or by that, or by the other: but it did not go by this road or by that: therefore he went by the other."[10]

Still, the idea has been expressed at times, e.g., by Richard Robinson in his study of *Plato's Earlier Dialectic*, when he argued that Plato was not aware of the distinction between what he calls an 'indirect refutation' (his expression for *reductio ad absurdum*) and a 'direct refutation' of the form 'A, therefore B', which would indeed directly refute $\neg B$. Answering the objection that the distinction is too obvious for Plato not to realize it, he wrote:

> This belief is destructive of any true history of human thought, and ought to be abandoned. Evidently there must have been a time when the human race, or its immediate ancestor, possessed no logical proposition at all, true or false. Nor is there any necessity that logical propositions, when they did arise, should at once be those which seem obvious to us. Nor did logical propositions in any scope and abstractness arise with Socrates or with the early Plato, but [...] with the later Plato and his pupil Aristotle. The history of thought cannot succeed if we assume from the beginning that some idea or other is innate and necessary to any human mind.[11]

Setting aside Robinson's argument from the history of thought, we note that the above example from *Lesser Hippias* has shown that, although no explicit rule had been stated, Socrates and Plato already knew how to argue for an

[10] Sextus Empiricus, *Outlines of Pyrrhonism*, I, 69.
[11] (Robinson 1953, 28-29).

universal affirmative proposition. It is rather easy to find instances of disjunctive syllogism or *Modus Ponens* or simply instantiations of game-semantic rules for logical connectives in Plato's dialogues.

Therefore, although it might be true that no "logical propositions in any scope and abstractness" arose with Socrates or (the early) Plato, it does not follow that they did not make any of the relevant inferences, i.e., that they not follow any of the corresponding rules. What seems wrong in Robinson's claim is thus that, in order to be able to follow a rule at all or at least to be recognized as having followed a rule, one needs to entertain an explicit statement of it, i.e., a 'logical proposition' or, more appropriately put, a logical truth. In the language of Wittgenstein's *Philosophical Investigations* §219, it would not be possible to follow a rule "blindly". We will discuss this point later on, but for the moment we would like to point out how Robinson's claim involves a difficulty raised by Lewis Carroll's notorious paradox of inference.[12]

Recall that in his 'What the Tortoise said to Achilles', Carroll describes an imaginary discussion involving a challenge issued by the Tortoise to Achilles. The Tortoise takes three propositions from Euclid:

> (*A*) Things that are equal to the same are equal to each other
>
> (*B*) The two sides of this triangle are things that are equal to the same
>
> (*Z*) The two sides of this triangle are things that are equal to each other

And she issues her challenge: he is to force her 'logically to accept *Z* as true' on the basis of *A* and *B*, i.e., '*A* is true' and '*B* is true'. The paradox of inference occurs in Achilles' attempt at forcing the Tortoise to accept *Z*. Given that *Z* follows from *A* and *B*, if the following is logically true:

> (*C*) $(A \,\&\, B) \rightarrow Z$

Achilles suggests that one includes *C* in the above:

> (*A*) Things that are equal to the same are equal to each other
>
> (*B*) The two sides of this triangle are things that are equal to the same
>
> (*C*) $(A \,\&\, B) \rightarrow Z$

[12] The discussion of Carroll's paradox in the following paragraphs is derived from (Marion to appear).

> (*Z*) The two sides of this triangle are things that are equal to each other

The Tortoise's reaction to this move is to point out that she has been served a further conditional or hypothetical proposition, which would be of the form:

> (*D*) $(A \,\&\, B \,\&\, ((A \,\&\, B) \rightarrow Z)) \rightarrow Z$

And that she would now refuse to grant *D*, which is, incidentally, *D* is an instance of the form of reasoning that medieval logicians called *Modus Ponens*. We can thus modify slightly Carroll's story, and present it under a form under which it is often discussed. Under this new form, the Tortoise accepts *A* and $A \rightarrow B$ but refuses to infer *B*, i.e., she rejects:

> (1) $(A \,\&\, (A \rightarrow B)) \rightarrow B$

Achilles then suggests that *A* and $A \rightarrow B$, and $(A \,\&\, (A \rightarrow B) \rightarrow B)$ together are to yield *B*, but the Tortoise now refuses to grant:

> (2) $(A \,\&\, (A \rightarrow B) \,\&\, ((A \,\&\, (A \rightarrow B)) \rightarrow B)) \rightarrow B$

If we reiterate the move, we will then get the Tortoise to refuse:

> (3) $(A \,\&\, (A \rightarrow B) \,\&\, ((A \,\&\, (A \rightarrow B)) \rightarrow B) \,\&\, (A \,\&\, (A \rightarrow B) \,\&\, ((A \,\&\, (A \rightarrow B)) \rightarrow B)) \rightarrow B)) \rightarrow B$

And so on.

Lewis Carroll offered no solution for his 'paradox of inference'. A common one consists in pointing out that the Tortoise would refuse to infer the conclusion (*Z*) by adding the needed rule of inference (*C*) as an extra premise simply because the rule of inference that she needs should not be added as premise; it is ineffectual as such. This point was made long before Carroll by Bolzano,[13] and it is to be found in Carroll's contemporary at Oxford, John Cook Wilson,[14] but Gilbert Ryle is more frequently cited for having made that very point:

> The principle of an inference cannot be one of its premises or part of its premises. Conclusions are drawn from premises in accordance with principles, not from premises that embody those principles.[15]

[13] In §199 of his *Wissenschaftslehre* (Bolzano 1972, 273-274).
[14] (Cook Wilson 1926, 443-444).
[15] (Ryle 1971, 138).

Ryle also observed:

> 'Well, but surely the intelligent reasoner *is* knowing rules of inference whenever he reasons intelligently.' Yes, of course he is, but knowing such a rule is not a case of knowing an extra fact or truth; it is knowing how to move from acknowledging some facts to acknowledging others. Knowing a rule of inference is not possessing a bit of extra information but being able to perform an intelligent operation. Knowing a rule is knowing how. It is realised in performances which conform to the rule, not in theoretical citations of it.[16]

Thus, according to Ryle, if there is any 'knowledge' involved in following a rule, it would be a 'knowing how', not a 'knowing that' or knowledge of a propositional content, i.e., of a logical truth such as (1). We are of course aware that Ryle's account has been controverted,[17] but we cannot defend it here, otherwise we will not get to the issues we wish to discuss. We wish instead to bring it in parallel with another lesson to be learned from Carroll's paradox hinges on distinguishing between (1), which is an implication that holds between unasserted propositions and an *inference* of the form '. . . , therefore . . . ', which holds between assertions, a distinction first made by Bertrand Russell.[18] With '$\vdash$' standing for the inference relation, while the comma on its left-hand side can be read as 'and', one would write the rule:

(1') $A, A \to B \vdash B$

This being the elimination rule for '$\to$' in Gentzen's natural deduction systems. Recall that these systems (and related sequent calculi) are characterized by the replacement of axioms by corresponding rules of inference, so that focus is shifted away from 'logical truth', since the class of logical truths now becomes merely a 'by-product' of the adoption of these rules. (This last point was already made in the *Tractatus* at 6.126.)[19] One can easily convince oneself that there is no regress if the confusion between '$\to$' and '$\vdash$' is avoided.[20]

In this context, it is worth going back to L. E. J. Brouwer's original idea, which led to the development of Heyting or BHK semantics, that *an inference*

[16] (Ryle 1971, 216-217).

[17] See, e.g., (Stanley 2011, 27*f.*).

[18] (Russell 1903, 35).

[19] See (Hacking 1976, 288*f.*).

[20] This is recognized, for instance, in (Engel 2007, 726 & 729). Somehow, Carroll had himself built this in his parable, given that the Tortoise ends up granting (1), but refusing to apply it.

is a kind of act. Like any action, an inference may be said to be in accordance with a rule or in violation of it, but does one need to entertain the rule of inference, either as a logical truth (1) or under the form (1′), in order to infer? The first thing to be said here is that having 'in mind' an explicit statement of the rule, say, (1′) would amount to *believing* that the corresponding implication (1) in logically true. These two ideas, i.e., that it is appropriate to speak of one believing in a logical truth such as (1) and of one as acting according to a rule such (1′) are seldom kept apart. This might be explained by the existence of what Stewart Shapiro called 'transfer principles',[21] i.e., principles that establish correspondences between rules of inferences and logical truths, e.g.,

(4) $A, A \to B \vdash B$ if and only if $(A \,\&\, (A \to B)) \to B$ is logically true

However, the *act* of inferring a conclusion from some premises in accordance to a rule of inference is not the same thing as a *belief* in the logical truth corresponding to that rule.[22] Belief is a propositional attitude one may have towards sentences such as logical truths, but one only *acts* according to rules or not. Does not have to refer to an 'internalized' version of rule in order to act. In some contexts, e.g., when doing logic exercises, one probably needs to, but it does not seem a necessary condition, as is amply demonstrated in other contexts such as that of dialectic in Plato's dialogues. And what seems the point of Carroll's paradox here, is that having 'in mind' an explicit statement of the rule as a logical truth amounts to introducing it as an extra premise, and it is this move that generates the regress. Therefore, under the view that harks back to Brouwer, there is no regress, while a regress is indeed generated with the view that when one infers, one has to have 'in mind' an explicit statement of the rule in the form of a logical truth, whose requirements one would then merely 'track'.

One can now see the obvious fault committed by Robinson: it is no use trying to find out if Socrates or Plato were 'aware' of this or that logical truth in order to recognize they were able to do this or that inference, because they were already acting/inferring, before one made explicit the rules in accordance to which their actions/inferences were made. The culprit here may be the view of logic as "the systematic study of the logical truths", to quote W. V.

[21] (Shapiro 2000, 337).

[22] This point has been made by many, e.g., (Dummett 1981, 596), (Priest 1979, 291), (Shapiro 2000, 337), and, as (Shapiro 2000, 338-339) notes, also it occurs in a closely related form in (Wright 1986, 192-194).

Quine,[23] which meant that one would look at logic not as a system of deductions of consequences from arbitrary, possibly false, premises, but as a system of proofs of logical truths based on logical axioms and rules.

One should note here a possible link with Wittgenstein's 'rule-following argument', given that it is understood by some as counting among its targets the very idea that in order to follow a rule one needs to track its requirements.[24] But before we should get to this, we would like to address an issue to which Pascal Engel as devoted numerous paper, namely the question 'How can logic move the mind?'. The original motivation for this question comes from Simon Blackburn's 'Practical Tortoise Raising': according to him, the Tortoise's failure to infer B indicates that logic alone does not move the mind, so that one always needs something else – a desire, a disposition or a habit – that *causes* one to infer. As he puts it:

> There is always something else, something that is not under the control of fact and reason, which has to be given as a brute extra.[25]

Engel agrees with Blackburn inasmuch as he takes Carroll's paradox to have refuted a simple version of 'internalism' which we need to reformulate here in terms of belief in (1):[26]

(i) $(A \,\&\, (A \rightarrow B)) \rightarrow B$ is a logical truth

(ii) This conditional is an instance of the form $(A \,\&\, (A \rightarrow B)) \rightarrow B$

(iii) Thus it is valid

One should note that (i)-(iii) is under the form of a *Modus Ponens*, so we will speak of the *Modus Ponens Model*, hereafter MPM, for reasons that will become clear shortly. According to this particular version of MPM, no one recognizing that (1) is a logical truth could fail to infer that B.[27] But Carroll's paradox shows precisely that this is not the case, given that the Tortoise is never moved to infer B. Hence the need to supplement this particular version of MPM with an added ingredient that will ultimately explain why the Tortoise was forced to infer, i.e., to provide a complete, satisfactory solution to Carroll's paradox.

[23] (Quine 1986, vii).

[24] See, e.g., (Wright 2001).

[25] (Blackburn 1995, 695).

[26] Engel puts it in terms of rules of inference in (Engel 2009, 27), but for the reasons just given, we think this is inappropriate.

[27] This is what Engel calls 'logical cognitivism' in (Engel 2005, 24*f.*).

Engel, who calls Blackburn's proposal 'externalism', also sees inference as a kind of act, but defined that act as "the moving from a belief to another", i.e., from one mental state to another, and he shares with Blackburn the idea that we have to explain how we are thus "*being moved* to infer".[28] He also grants to Blackburn that one's following a rule of inference without being conscious of it, because of a disposition or habit.[29] Engel ends up dismissing Blackburn's externalism, however, but this is not the place to enter into a detailed discussion of his reasons. For the purpose of our argument we need merely to emphasize the premises they share. We merely note one objection: to be told that a disposition or habit as a 'brute extra' that forces us to infer is Engel's eyes insufficient, because this much does not at the same time explain why our belief in the truth of the premises and in the validity of the inference both justifies us *and* forces us to infer. For the same reasons, Engel finds it insufficient to be told that one follows the rule 'blindly'.[30] Engel offers instead his own solution, which he calls 'sophisticated'[31] or 'nonreflective internalism',[32] in order to distinguish it from the above 'internalism' that falls prey to Carroll's regress argument. His solution hinges on the analogy between judgements of perception and logical judgements (of which (i), above, is an example). Judgements of perception such as my judging that 'this object is red' are here taken to be non-inferential and as offering us an 'epistemic warrant'. In the case of logical judgements, we are told, one needs the logical rule as some sort of 'norm' or 'law', which is in that sense external, objective, and so forth, but one also needs an internal reason accessible to the agent, which would justify her in making the inference according to it, i.e., would be the

[28] (Engel 2005, 25). (Engel speaks at times as if 'inferring' is a "mental state' (Engel 1998, 48), but the claim would be indefensible.) One should note that we do not understand the idea that inference is a kind of act in the same manner. To begin with, for us an inference is not a relation between beliefs but a relation between assertions. Furthermore Engel seems to be looking for a further mental state to explain the 'moving' from one state to another.

[29] (Engel 2005, 31). This might serve as basis for a critique to the view we put forth above concerning the likes of Socrates and Plato, who followed before they were made explicit, namely that all this would show is that they have internalized these rules. The problem with this view is that one cannot presuppose that what was internalized was first made explicit, so it is not clear what 'internalization' means in this context. The source of the rules of inference is to be located elsewhere, in the very interaction within, say, a regimented dialogue such as dialectic or even an ordinary dialogue, which contains them implicitly.

[30] (Engel 2009, 28). This point could be served to object to our own approach in this paper, as it would appear that we are also not attempting to explain why we are *forced* to infer. Our point is rather that we wish to undermine the need to provide such an explanation.

[31] (Engel 2009, 32).

[32] (Engel 2007, 737).

cause of her inferring, and some sort of equivalent of an epistemic warrant in the case of perceptual judgements for logical judgements would thus provide that extra something needed for the mind to be moved.[33]

One could try and pick apart Engel's solution at its weak spots, e.g., the analogy between perceptual and logical judgements, but we will not engage in this direction (Engel is aware of this objection and provides a rejoinder)[34] or by asking further clarifications concerning the nature of the 'warrant' in the case of logical judgements. We would like first simply to note that our 'inferentialist' solution is neither 'internalist' (simple or sophisticated) nor 'externalist', because we think that, in the end, the fault is with MPM, which we do not presuppose in our solution of Carroll's paradox.[35] Secondly, we would like to argue that the range of solutions marshalled by Engel, including his own solution and related ones[36] are defective precisely in their reliance on some form or the other of MPM, and the request that some extra ingredient be added to it, and for this we would like to propose an argument which we claim can be found in Wittgenstein's remarks on 'rule-following' in *Philosophical Investigations*, hereafter *PI*, §§143-242.[37]

*

To make our point, we need to rehearse the central argument of Wright's most recent paper on rule following: 'Rule-Following Without Reasons: Wittgenstein's Quietism and the Constitutive Question'.[38] It is based on a distinc-

[33] (Engel 2009, 32).

[34] (Engel 2009, 30-31).

[35] As we claimed above, rules of inference as such are merely making explicit features of a prior practice, i.e., in the case of the Ancient Greeks, the highly regimented form of dialogue known as 'dialectic' so our proposal would thus hinge instead on a better understanding of the location and role of rules of inference in *dialogue* itself. This is, alas, not a point that can be argued for here.

[36] For example, a view such as Bill Brewer's appeal to some sort of rational compulsion; see, e.g., (Brewer 1995, 242), which is briefly discussed in (Engel 2005, 29), (Engel 2007, 743, n. 30), (Engel 2009, 30).

[37] Qutoations from *Philosophical Investigations* are from the 4th edition, (Wittgenstein 2009). Engel also saw interesting links between Carroll's paradox and Wittgenstein on rule following, e.g., at (Engel 1998, 49-51) and (Engel 2007, 726), but they are in terms that presuppose the validity of Kripke's reading in (Kripke 1982). We read *PI* §§143-242 otherwise – see the next footnote.

[38] (Wright 2007). In this last section we make use of a variant of the rule-following argument by Friedrich Waismann, which we studied in an hitherto unpublished paper, 'Wittgenstein on Reasons, Causes, And Rule-Following: A Variation by Waismann' (Marion & Okada, manuscript), thus putting greater emphasis that usual on the source of some remarks *PI* §§143-242 in the opening pages of the *Blue Book* and the distinction between reason and cause.

tion between basic and the complex cases of rule-following. To use Wright's own examples, predicating 'red' would be a basic case, while castling in chess would be complex one. Concerning the latter, Wright proposes a "*modus ponens* model of rule-following". This being another version of MPM, we adapt his example:[39]

> (i') If neither King nor one of its Rooks has moved in the course of the game so far, and if the squares between them are unoccupied, and if neither the King nor any of those squares is in check to an opposing piece, then one may Castle.
>
> (ii') In this game neither my King nor this Rook have yet been moved, the squares between them are unoccupied, and ...
>
> (iii') I may castle now.

Wright's argument is to the effect, however, that MPM cannot apply to basic cases. To demonstrate this, he proposes that we entertain what it would look like in the case of predications of red (we adapt again his example):[40]

> (i'') If ...x ..., it is correct to predicate 'red' of x
>
> (ii'') ...x ...
>
> (iii'') It is correct to apply 'red' to x.

MPM requires here that one possesses an anterior concept '...x ...' whose satisfaction in a given situation will determine as appropriate the application of the rule. Wright is quick to point out, however, that the anterior concept in question is no other than the concept 'red' itself. But this looks as if in order to follow any rule one needs a conceptual repertoire anterior to the understanding of that very rule, a bit like the child, in *PI* §32, who learned his first language as if he "came into a strange country and did not understand the language of the country; that is, as if [he] already had a language, only not this one". As Wright points out:

> In short the problem with extending the modus ponens model to cover all rule-following, including that involved in basic cases, is that it calls for a conceptual repertoire *anterior* to an understanding of any particular rule – the conceptual repertoire needed to grasp the input conditions, and the association of them which the rule

[39] (Wright 2007, 490).
[40] (Wright 2007, 495).

> effects with a certain mandated, prohibited or permissible form of response. From the standpoint of the philosophy of thought and language of the *Investigations*, this is an enormous mistake. With respect to a wide class of concepts, a grasp of them is not anterior to the ability to give them competent linguistic expression but rather *resides in* that very ability.[41]

Therefore:

> the modus ponens model *must* lapse for basic cases. Basic cases – where rule-following is 'blind' – are cases where rule-following is *uninformed by anterior reason-giving judgement*.[42]

To see how the point applies to our above discussion, it suffices that we think of applying MPM to *Modus Ponens* itself: this would result in exactly Carroll's regress.

One could reply to this that Engel has avoided precisely this very pitfall with his idea of a non-inferential warrant, in the case of perceptual and logical judgements. Not so, since a problem appears to lie in the vicinity, not for the notion of 'non-inferential warrant' but for the idea that we need one at all. Recall now what Wittgenstein had to say in *PI* §§211-219 about 'blind rule-following':

> 211. No matter how you instruct him in continuing the ornamental pattern, how can he *know* how he is to continue it by himself – Well how do *I* know? – If that means "Have I reasons?" the answer is: my reasons will soon give out, and then I shall act without reasons.
>
> 217. "How am I able to follow a rule?" – If it is not a question about causes, then it is about the justification for my acting in *this* way in complying with the rule.
>
> Once I have exhausted the justifications, I have reached bedrock, and my spade is turned. Then I am inclined to say: "This is simply what I do".
>
> 219. [...] When I obey a rule I do not choose.
>
> I obey the rule blindly.

[41] (Wright 2007, 496).

[42] (Wright 2007, 496). One could even go one step further and point out that complex cases are always reducible to a set of basic cases, a bit like arithmetical calculations can be broken down in a series simple tasks, e.g., programming a Turing Machine to perform addition, so that basic cases in the end the key cases of rule-following.

The idea that "reasons will soon give out" so that, in the end, "I obey the rule blindly" is likely to be misunderstood.[43] We suggest that, to gain a better understanding of its meaning, it be replaced in its context of origin in the *Blue Book*, where Wittgenstein suggests that there are two ways of looking at teaching the meaning of a word such as 'red'. Either (a) teaching is a drill that could be said "to have built up a psychical mechanism", i.e., a 'disposition' or (b) teaching supplies one "with a rule which is itself involved in the process of understanding, obeying, etc.". In other words, it supplies one with a reason. The first alternative is, by contrast, causal:

> The drill of teaching could [...] be said to have built up a psychical mechanism. This, however, would only be a hypothesis or else a metaphor. We could *compare* teaching with installing an electric connection between a switch and a bulb [...] it is the *cause* of the phenomena of understanding, obeying, etc; and it is an hypothesis that the process of teaching should be needed in order to bring about these effects.[44]

Wittgenstein then proceeds to argue against (a), so that it "drops out of our considerations".[45] One of the points made is that (a) would only provide us with a cause, not a reason. He then defends (b) against a possible regress argument:

> Now there is the idea that if an order is understood and obeyed there must be a reason for our obeying it as we do, and, in fact, a chain of reasons reaching back to infinity.[46]

Indeed, Wittgenstein dismisses as based on a wrong analogy with the infinite divisibility of the line:

> [...] the idea of an infinite chain of reasons arises out of a confusion similar to this: that a line of a certain length consists of an infinite number of parts because it is indefinitely divisible, i.e., because there is no end to the possibility of dividing it. [...] If on the

[43] For example, Kripke understands the thought as meaning that applying the rule is "an unjustified stab in the dark", (Kripke 1982, 16).

[44] (Wittgenstein 1969, 12).

[45] (Wittgenstein 1969, 14). The connection with *PI* §§143-242 is obvious, since he is also trying there to undermine the view that understanding is a 'mental process' (§154).

[46] (Wittgenstein 1969, 14).

> other hand you realize that the chain of *actual* reasons has a beginning, you will no longer be revolted by the idea of a case in which there is *no* reason for the way you obey the order.[47]

The upshot is clear: having dismissed causal explanations in terms of training, Wittgenstein saw no valid objection to holding the view that the chain of reasons comes to an end. He immediately entertains the possibility that one carries on asking 'Why?', even after reaching this 'bedrock', i.e., the end of the chain of reasons:

> At this point, however, another confusion sets in, that between reason and cause. One is lead into this confusion by the ambiguous use of the word "why". Thus when the chain of reasons has come to an end and still the question "why?" is asked, one is inclined to give a cause instead of a reason.[48]

So the thought is that, once the end of the chain of reasons is reached, there ought not to be any more 'why-question' asked, because answering them would lead us into causal territory, so to speak, i.e., one would surreptitiously begin to replace rational by causal explanations. Having one's spade turned when reaching bedrock, merely means, therefore, that one should not step into the realm of causal explanation. But that is precisely the sort of extra 'causal' ingredient that the 'internalists' and 'externalists' discussed by Engel are looking for.

What would the argument in favour of this prohibition be? It would simply be the argument against MPM, properly understood. As Wright puts it:

> And basic- 'blind' – rule-following, properly understood, is rule-following without reason – not in the sense of being phenomenologically immediate or spontaneous in the way a good chess player may make a clever move without fully self-consciously rationalising his grounds for it, but in a sense involving the inappropriateness of the modus ponens model.[49]

The upshot is thus the inapplicability of MPM to basic cases, i.e., 'blind rule-following'. Thus the inappropriateness of MPM calls into question the very basis for the range of views marshalled by Engel, including his own solution,

[47] (Wittgenstein 1969, 15). See also (Wittgenstein 1969, 143).
[48] (Wittgenstein 1969, 15).
[49] (Wright 2007, 497).

inasmuch as they purport to find the missing ingredient that will make MPM function.

*

Perhaps we should let Crispin Wright speak one last time, in order to conclude:

> In any basic case, the lapse of the modus ponens model means that we should not think of knowledge of the requirements of the rule as a state which rationally underlies and enables competence, as knowledge of the rule for castling rationally underlies a chess player's successfully restricting the cases where she attempts to castle situations where it is legal to do so. In basic cases there is no such underlying, rationalising knowledge enabling the competence. *A fortiori* there is no metaphysical issue about the character of the facts it is knowledge of, with Platonism and communautarism presenting the horns of a dilemma. The knowledge *is* the competence. Or so I take Wittgenstein to be saying.[50]

As an interpretation of Wittgenstein, this seems on the whole right, especially in its 'quietist' inclinations. But the argument deployed against MPM in basic cases of rule-following to sustain this conclusion is also a powerful one, one that suggests that we abandon the attempt to find a solution of the type envisaged by Engel. We began the paper with an example of a practice, dialectic as a regimented form of dialogue, where logical rules where followed prior to being made explicit. Here too we have basic cases where, so to speak, 'there is no underlying, rationalising knowledge enabling the competence' and 'knowledge *is* the competence'. This, again, is very much in line with the basic lesson to be learned from Carroll's paradox.

1. References

Blackburn, S., 1995, 'Practical Tortoise Raising', *Mind n.s.*, vol. 104, 695-711.

Bolzano, B., 1972, *Theory of Science*, Berkeley/Los Angeles, University of California Press.

Brandom, R., 1983, 'Asserting', *Noûs*, vol. 17, 637-640.

[50] (Wright 2007, 498).

Brandom, R., 1994, *Making It Explicit. Reasoning, Representing & Discursive Commitment*, Cambridge Mass., Harvard University Press.

Brandom, R., 2000, *Articulating Reasons. An Introduction to Inferentialism*, Cambridge Mass., Harvard University Press.

Brewer, B., 1995, 'Compulsion by Reason', *Proceedings of the Aristotelian Society. Supplementary Volume LXIX*, 237-254.

Carroll, L., 1895, 'What the Tortoise said to Achilles', *Mind*, n. s., vol. 4, 278-280.

Castelnérac, B. & M. Marion, 2009, 'Arguing for Inconsistency: Dialectical Games in the Academy', in G. Primiero & S. Rahman (eds.), *Acts of Knowledge: History, Philosophy and Logic*, London, College Publications, 37-76.

Castelnérac, B. & M. Marion, 2013, 'Antilogic', *Baltic International Yearbook of Cognition, Logic and Communication*, vol. 8, URL : http://newprairiepress.org/cgi/viewcontent.cgi?article=1079&context=biyclc

Cook Wilson, J., 1926, *Statement and Inference*, 2 vols., Oxford, Clarendon Press.

Dummett, M. A. E., 1981, *Frege. Philosophy of Language*, 2nd ed., London, Duckworth.

Engel, P., 1998, 'La logique peut-elle mouvoir l'esprit?', *Dialogue*, vol. 37, 35-53.

Engel, P., 2005, 'Logical Reasons', *Philosophical Explorations*, vol. 8, 21-38.

Engel, P., 2007, 'Dummett, Achilles and the Tortoise', in R. E. Auxier & L. E. Kahn (eds.), *The Philosophy of Michael Dummett*, Chicago & LaSalle IL., Open Court, 725-746.

Engel, P., 2009, 'Oh ! Carroll ! Raisons, normes et inférences', *Revue Klesis*, n. 13, URL: http://www.revue-klesis.org/pdf/3-Engel.pdf

Hacking, I., 1976, 'What is Logic?', *Journal of Philosophy*, vol. 76, 296-319.

Kripke, S., 1982, *Wittgenstein on Rules and Private Language*, Cambridge MA., Harvard University Press.

Marion, M., 2011, 'Game Semantics and the Manifestation Thesis', in M. Marion, G. Primiero & S. Rahman (eds.) *The Realism-Antirealism Debate in the Age of Alternative Logics*, Dordrecht, Springer,151-180.

Marion, M., *to appear*, 'Lessons from Lewis Carroll's Paradox of Inference', *The Carrollian*.

Marion, M. & Okada, M., *unpublished*, 'Wittgenstein on Reasons, Causes, And Rule-Following: A Variation by Waismann'

Marion, Mathieu & Rückert, H., *unpublished*, 'Aristotle on Universal Quantification: A Study from the Perspective of Game Semantics'.

Priest, G., 1979, 'Two Dogmas of Quineanism', *Philosophical Quarterly*, vol. 29, 289-301.

Quine, W. V., 1986, *Philosophy of Logic*, 2nd ed., Cambridge MA., Harvard University Press.

Robinson, R., 1953, *Plato's Earlier Dialectic*, Oxford, Clarendon Press.

Russell, B., 1903, *The Principles of Mathematics*, London, Allen & Unwin.

Ryle, G., 1971, *Collected Papers*, London, Hutchinson, vol. 2.

Shapiro, S., 2000, 'The Status of Logic', in P. Boghossian & C. Peacocke (eds.), *New Essays on the A Priori*, Oxford, Clarendon Press, 333-366.

Stanley, J., 2011, *Know How*, Oxford, Oxford University Press.

Wittgenstein, L., 1969, *The Blue and Brown Books*, 2nd ed., Oxford, Blackwell.

Wittgenstein, L., 2009, *Philosophical Investigations*, revised 4th edition, Oxford, Wiley-Blackwell.

Wright, C., 1986, 'Inventing Logical Necessity', in J. Butterfield (ed.), *Language, Mind and Logic*, Cambridge, Cambridge University Press, 187-211.

Wright, C., 2001, 'Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics', in *Rails to Infinity*, Cambridge, MA., Harvard University Press, 170-213.

Wright, C., 2007, 'Rule-Following Without Reasons: Wittgenstein's Quietism and the Constitutive Question', *Ratio (new series)*, vol. 20, 481-501.

42

Taking Norm-Regulation Seriously

DAVIDE FASSIO

Abstract Engel has recently introduced a distinction between *norm* and *norm-regulation*. The regulation of a norm concerns the ways in which agents can follow that norm. In this paper I develop in some detail a particular account of regulation. The notion of regulation that I outline here is non-normative; it consists of a set of descriptive conditions about the agent's epistemic position, intentions, motivations and environment. I also provide an account of rationality as a notion dependent on that of regulation. I characterize rationality as the obtaining of a subset of regulation conditions of some norm. The most striking consequence of my account is that rationality is not a normative notion. I conclude the paper by discussing differences and relations between assessments relative to norms and those relative to norm-regulation.

*

It's a great pleasure and honor for me to contribute to the present collection of papers celebrating the sixty years of Pascal Engel. Engel has been for me a teacher and a friend. I learned a lot from his writings and the discussions I had with him, and I am extremely grateful for his invaluable friendship. In this essay I shall focus on a distinction recently discussed by Engel that deeply inspired some of my ideas and works. The distinction is that between a *norm* and its *regulation*.[1] On the one hand, a norm requires, permits or forbids something to someone, it is addressed to a specific set of agents and involves specific conditions of satisfaction.[2] On the other hand, the regulation of a norm concerns the ways in which agents follow, or can follow that norm; it consists in a set of conditions whose satisfaction is necessary for following a norm, such as the obtaining of certain psychological and environmental circumstances. Engel introduces such a distinction in order to cope with some problems affecting the view that belief is constitutively governed by a truth-norm. According to some objectors, since whether a proposition is true or false is an objective matter not always transparent from a subjective perspective, it is unclear how a norm to believe only the truth can guide and motivate its addressees.[3] The distinction between norm and norm-regulation promises to solve this problem by delegating to the latter the function of enabling the norm's guidance and motivation. However, here my concern will not be restricted to the context in which the distinction has been originally formulated. My main aim in this paper is to develop in some detail a version of this distinction extensible to all normative domains.

The account of norm-regulation that I will present in this paper partially diverges from that suggested by Engel. The main difference is that the notion of regulation that I outline here is non-normative, consisting of a set of

[1] Engel discusses this distinction in several places. See in particular Engel 2007a pp. 160-164, 2008 pp. 56-57, and 2013.

[2] The terms 'normative', 'norm' and 'normativity' are used in two senses. In a broad sense, the normative is contrasted with the descriptive, where the latter includes physical entities and properties and some abstract entities and properties like sets and numbers. This distinction is often drawn on the basis of a distinction between what *is* and what *should/ought to be*, or in terms of natural vs. non-natural facts (where natural facts are the proper objects of natural sciences or mathematics, accountable for in mere descriptive terms). In a narrow sense, the domain of norms is contrasted with that of values. Here I use these terms in the narrow sense.

[3] For a discussion of this type of problem see, for example, Steglich-Petersen 2006, Bykvist & Hattiangadi 2007, Gluer&Wikforss 2009. Engel uses the distinction to address similar problems concerning the knowledge norm of assertion in Engel 2008.

descriptive conditions about the agent's epistemic position, intentions, motivations and environment. This has important consequences for a set of notions definable in terms of norm-regulation. As I will argue, one of these notions is that of *rationality*. This implies that, according to this picture, rationality is not normative. Though in this paper I will limit my discussion to rationality, similar considerations can be extended to other notions that I consider regulation-dependent, such as those of epistemic justification, warrant, responsibility and excusability. Here a significant departure from Engel's views is apparent, since he holds that rationality and justification are normative notions (e.g., Engel 2007b, 2011).

This is the plan of the paper. In §1 I introduce the distinction and develop in detail a specific account of norm-regulation. In §2 I provide an account of rationality based on the account of regulation considered in §1. I characterize rationality as the obtaining of a subset of regulation conditions of some norm. I also clarify some differences between my account of rationality and other well-known contemporary accounts. In §3 I discuss the differences and relations between two types of assessments, one relative to norms, and the other relative to norm-regulation.

1. Norm and Norm-Regulation

In order to clarify the difference between norm and norm-regulation, let me introduce some features of norms. Consider a specific norm by way of example: the law that requires citizens to pay taxes. The law is addressed to a specific set of agents, citizens, and is satisfied if and only if citizens pay taxes. It is supposed to motivate citizens to do what it requires, providing them with reasons to comply with it. It does that by means of praise of those who respect it, justification of the reasons why they must respect it, and the threat of punishment for infractions.[4] The very same features individuated in this law can be found in every other norm. Norms are in force for some purpose or reason (in the above example, the purpose is to finance part of the public expenses). They require, forbid or permit something to someone. They are addressed to a specific set of agents (the *addressees* of the norm). They have *satisfaction conditions* (what must be the case for the norm to be satisfied), and sometimes

[4] The ways in which norms are able to motivate their addressees to comply with them are different. Agents can be motivated to follow norms by the fear of punishment, the desire to respect a common established convention, a self-commitment to the rules, the aversion to negative feelings such as shame, embarrassment and guilt, the criticism of other participants in a practice, the risk of exclusion from a practice, and so on.

also conditions for application (for example, one has to pay taxes for the ownership of an house only if one does own one).

Commonly what norms require, permit or forbid are objective conditions, such as the performance of an action or the obtaining of a certain state of affairs. Such conditions determine in which circumstances a norm is satisfied – in the above example, the norm requires *paying taxes*, and it is satisfied when taxes are paid. These conditions must be distinguished by the conditions necessary for an agent to follow a norm. In order to follow a norm, an agent must acknowledge it (both its existence and what it demands), recognize its normative force; she must be motivated by the norm and form certain intentions, try to realize certain means necessary for norm compliance, and so on.[5] The difference between these two types of conditions – the satisfaction conditions of a norm on the one hand and the conditions necessary for following the norm on the other – is made apparent by cases in which an agent satisfies the former, but not the latter: one could ignore a norm, or not accept or fail to be motivated by it, and still comply with it by chance.[6] This is, for example, the case of someone who takes a plane, ignoring that it's forbidden to smoke on board, but who does not smoke for some other reason (for example because she dislikes smoking on a plane). In such cases, though the agent doesn't intentionally follow the norm and is not guided by it, she complies with the norm by merely fulfilling its satisfaction conditions (i.e., by not smoking on the plane).[7]

The above considerations show that there is an important difference between, on the one side, norms and their satisfaction conditions, and, on the other side, the ways in which agents follow norms and the conditions enabling such agents to comply with them. This is precisely the difference between norm and norm-regulation introduced by Engel in many of his works. Consider the following passage in which Engel states the distinction:

> "[...] there is no reason why we should not distinguish two levels: (a) The statement of the norm [...]; (b) How the norm is regulated (its regulation). It is one thing to say what the norm is, that is what

[5] I will say more on these conditions later in this section.

[6] For obvious reasons this happens more often with norms of permission than with requirements.

[7] In such cases, if there is some blame or negative assessment of the conduct of the agent, it does not concern the violation of the norm (in fact the norm is not violated), but the inappropriateness of the ways in which the agent complied with the norm – ways not in conformity with how norms are supposed to guide and that could have easily brought about an infraction if circumstances had been slightly different. For more on this type of assessment see §3.

> kind of truth (analytic, or essential) is expressed by it, and it is another thing to say how the norm is regulated, and realised in the psychology of the believers. ... The distinction between the statement of the norm and the conditions of its regulation is reminiscent of the distinction between the formulation of a general norm on the one hand, and its conditions of application, or between the law and its decrees of application" (2007a, p.163).

Elsewhere, Engel describes norm-regulation as the subjective conditions under which the satisfaction conditions of a norm are accessed by a given individual and are implemented in his psychology (2008, 56-57). In short, we can say that the domain of regulation of a norm includes the set of conditions that allow an agent to follow a norm, from the epistemic access to the norm to the appropriate psychological attitudes and practical conditions that allow the agent to follow it.

There are two differences between the interpretation of norm-regulation suggested by Engel and the one that I consider here. The first is that, though Engel in his works describes this notion in non-normative terms, as concerning mere psychological features of agents, he seems to follow other philosophers in identifying norm-regulation with a number of second-order subjective norms. According to this view, an agent can follow an objective norm by following a set of subjective norms related in some way to the objective one. For example, in some papers Engel considers as responsible for the regulation of the truth-norm of belief other derived norms of evidence and rationality.[8] This strategy has been adopted by many other philosophers (e.g., Boghossian 2003, p. 39, Gibbard 2005, p. 343, Shah 2003, p. 471, Wedgwood 2002, p. 282). I consider this view problematic for several reasons, some of which will be mentioned in the next section. In contrast, here I will take 'seriously' the descriptive characterization of norm-regulation suggested by Engel. As I conceive it, the regulation of a norm is a non-normative matter, a matter of psychological (cognitive, volitional and motivational) conditions and external environmental conditions. This account of norm-regulation is not normative in the sense that it does not involve any further commitment binding the agent to whom the norm is directed; there is not a further 'ought' on addressees of a norm beyond that norm itself.

[8] An exception is Engel (2013), where he considers the regulation of the specific truth-norm constitutively governing beliefs and argues that such a norm is regulated through the phenomenon of doxastic transparency in intentional processes of deliberation.

The second difference between the notion of norm-regulation suggested by Engel and the one considered here is that I include in the conditions of norm-regulation not only psychological features of the agent, but a series of conditions external to agents necessary for enabling them to follow a norm, such as environmental conditions necessary to become aware that one is under a normative commitment. I've to admit that it is unclear to me whether Engel would disagree on extending norm-regulation to these factors external to the psychology of agents, or whether he just does not mention them because he is concerned with issues for which external conditions of regulation are irrelevant, such as how an agent from her own subjective perspective can follow a norm involving objective satisfaction conditions.

Let me now consider in more detail the various conditions of norm-regulation according to my account of such a notion. I have already mentioned some of them above. These conditions can be distinguished as *internal* and *external* to the psychology of an agent,[9] and in *preliminary* and *core* conditions. Internal conditions can be further distinguished as *voluntary* and *involuntary*. I will present the various conditions in the order in which they must be satisfied for an agent to be able to follow a norm.

1. *Preliminary conditions (internal and external).* To follow a norm, an agent must satisfy a number of preconditions: she must acknowledge the existence of the norm, that the norm is supposed to guide and motivate some addressee to act in some way (i.e., that it is reason providing), that she is one of these addressees, that the norm has such and such satisfaction conditions, that she has at least a rough idea of how to comply with it, and so on.[10] These conditions are necessary for the agent to be able to realize that she is committed to a norm, even before her decision to accept the norm and her attempt to comply with it. They are preliminary conditions for the agent to be at least minimally responsive to normative demands. Some of these conditions are *external* to the agent's psychology, mainly concerning the accessibility of various features of a

[9] I am aware of the difficulties in drawing a neat distinction between what is internal or external to one's psychology. Here I don't want to enter into such deep issues. After all, the present distinction can be conceived as partially stipulative.

[10] Some philosophers called such preliminary conditions 'enabling conditions' (e.g., Dancy 2000, p.127, Menzies 2004, Steglich-Petersen 2010). Notice however that enabling conditions, as discussed by these philosophers, include also some of the conditions that I included in the set of external core conditions, such as access to whether the satisfaction conditions of the norm actually obtain.

norm to its addressees. For instance, a condition for following a norm is that there are no physical obstacles to a possible acknowledgment of it: the law requires drivers to stop when traffic lights are red, and an external condition for the regulation of this law is that no obstacle precludes a driver from seeing a traffic light. Other preliminary conditions are *internal*. For example, an agent could be unable to acknowledge a norm in a specific circumstance because of some cognitive defect.

2. *Internal voluntary core conditions*. Once the above preliminary conditions (both internal and external) are satisfied, an agent must satisfy other conditions in order to be able to follow a norm. The satisfaction of some of these conditions depends on factors under the voluntary conscious control of the agent. The most important of these conditions is the *acceptance* of the norm.[11] Once an agent has acknowledged that there is a norm supposed to provide her with reasons to perform a certain action, she must *accept* it. This means that she must take the norm as something she ought to follow, as providing *all things considered* reasons for her to act as it requires, forbids or permits. When an agent doesn't accept a norm, she either does not take the norm as authoritative and forceful enough to provide her with reasons to act, or takes the reasons provided by the norm as only *pro tanto* and outweighed by other stronger reasons. For example, someone could know that the law requires her to pay taxes, and nevertheless decide not to pay them, consciously violating the law. Though the law provides *pro tanto* reasons for paying taxes, there are other reasons outweighing them, such as the desire to be richer and the thought that it's extremely improbable that her tax evasion will be detected.

3. *Internal involuntary core conditions*. Once one accepts a norm, if all goes well, one will be motivated by the norm to act as it requires. The agent's motivation will be accompanied by an intention that will lead to an action. However all does not always go well. Sometimes an agent can accept a norm, but still fail to be motivated to act as the norm requires. This happens when the connection between taking oneself as having *all things considered* reasons to F and being motivated to F fails. Or one may be motivated and intend to F, and still fail to act as intended because the motivation has a force insufficient for acting as wanted. These are cases

[11] On the notion of norm-acceptance see for example Boghossian 2008 and Gluer & Wikforss 2009.

of akratic behavior due to internal factors out of the control of agents, such as psychological processes not operating in the normal way. Examples are pathological cases of dependence: an agent could be motivated to stop smoking, but be unable to refrain, irrationally acting against her own will.

4. *External core conditions.* An agent satisfying all the above conditions could still fail to follow a norm because of the lack of favorable environmental conditions enabling her to be adequately responsive to the normative demand. For example, the satisfaction conditions of a norm could not be fully transparent to an agent due to contingent environmental circumstances. We can try to comply with a norm, be aware of what its satisfaction conditions are, and nevertheless be unable to follow that norm because we fail to recognize whether these conditions actually obtain or not. Similarly, an agent could not be in a position to see whether the conditions for the application of a norm obtain. For example one could be motivated by the law to pay taxes, but fail to pay all of them due to the complexity of the procedures of payment, or because she ignores the fact that she is committed to pay certain taxes. Of course, in some such cases (though not in all) one can be excused for violating a norm, but this does not change the fact that one violated it. Furthermore, an agent could try to follow a norm but, for contingent reasons, be unable to act in ways appropriate to the satisfaction of the norm. For instance an agent can fail to follow a norm because that would require a type of ability that she has still not acquired. Consider the biblical precept not to desire the things that belong to others. If this norm can be followed, it requires an indirect control of certain desires. Presumably, this requires certain acquired abilities than not all agents have yet developed.

2. Rationality as a regulation-dependent notion

In this section I provide an account of rationality. I suggest that rationality is strictly related to norm-regulation. In my view, whether an agent is rational depends on the obtaining of a subset of regulation conditions relative to some norm. Though here I will limit my discussion to the notion of rationality, I think that analoguous considerations can be applied to other notions such as those of justification, warrant, responsibility and excusability. In my view,

all these notions, that I call *regulation-dependent*, can be defined in terms of the presence or absence of a subset of regulation conditions of some specific norm.

An important consequence of my account is that, since regulation is not normative (at least in the sense specified in the previous section that it does not involve commitments saying what agents ought or are permitted to do), and rationality, justification and other regulation-dependent notions are defined in terms of subsets of regulation conditions, these notions also are not normative in this sense. There are no norms of rationality or justification. However, there is a sense in which such notions can be said to be *norm-relative*, for they cannot be defined or characterized without making reference to some norm (as with the notion of norm-regulation itself). For example, as I conceive the notion of epistemic justification, a justified belief is a belief satisfying all the internal descriptive conditions necessary for the regulation of a truth-norm constitutive of belief. These conditions involve the possession of a set of non-normative properties. This set of properties is characterized by reference to the regulation of the constitutive norm. Therefore, that norm plays a role in individuating the set of properties necessary for epistemic justification. However, this does not entail that for being justified one must satisfy some norm, or that one is under a requirement to be justified in addition to being committed to the constitutive norm of belief.

Rationality

An agent is rational when she satisfies a specific subset of regulation conditions of a norm. Rationality is primarily a property of agents following a norm, and derivatively a property attributed to attitudes relevant for the regulation of the norm. Not all regulation conditions are relevant for rationality. External conditions are irrelevant for whether an agent is rational or irrational. An agent cannot be deemed irrational for not having epistemic access to a norm, or because she cannot satisfy a norm on account of environmental conditions independent of its psychology. For example one trying to comply with a norm but unable to comply with it because one fails to recognize whether the satisfaction conditions of the norm actually obtain is still rational. Internal preliminary conditions of regulation are also irrelevant for rationality, but in a different sense. An agent who doesn't follow a norm because she doesn't acknowledge it or doesn't recognize its normative force is neither rational nor irrational. Rather, she cannot be judged according to a standard of rationality. In this sense, preliminary regulation conditions of a norm work as preconditions for the attributability of rationality or irrationality to an agent or attitude.

Some internal voluntary core conditions are relevant for rationality, though not all of them are. Sometimes an agent does not accept a norm and still is rational, for example when she takes the norm as providing reasons only *pro tanto*, outweighed by other stronger reasons. An agent consciously violating a norm is not always irrational in circumstances in which she has *all things considered* reasons to do that. For instance, suppose I am driving in my car to an important meeting; it is a matter of life or death that I arrive on time and I am late; in this circumstance it's rational for me not to respect the speed limits. However, an agent who i) knows that she is committed to a norm, ii) recognizes its normative force, and iii) takes the norm as providing *all things considered* reasons, but iv) does not accept it as reason providing, is irrational. Acting in this way would denote a form of unresponsiveness or insensitivity to normative reasons.

The satisfaction of internal involuntary core conditions is always necessary for rationality. If one takes a norm as providing *all things considered* reasons to do something, and still is not motivated to do that thing, or doesn't intend to do it, because of weakness of the will or some sort of cognitive failure, then one is irrational. Cases of akratic behavior are typical instances of irrationality.

In sum, the regulation conditions relevant for rationality are the internal involuntary core conditions plus some internal voluntary core conditions. Furthermore, for being rational, it's not sufficient that the above conditions are satisfied. These conditions must also be connected in the right way. For example, an agent may accept a norm, take it as providing all things considered reasons to act in a certain way, and also be motivated and intend to act in that way, but the connection between reasons and motivation could be of the wrong kind. For example, the motivation could be caused by some abnormal psychological process rather than stemming from an appropriate consideration of reasons.[12]

The above characterization of rationality provides a simple explanation of many "requirements of rationality" discussed by philosophers.[13] Consider,

[12]Characterizing a connection of the right kind is notoriously difficult. The force of a norm is supposed to determine the agent's motivation *for the right kind of reasons*. On the appropriate ways in which norms are supposed to motivate their addressees see Glüer e Pagin (1999), p. 208. On the inappropriateness of deviant causal chains in the explanation of normative guidance see Railton (2006) and Schroeder (2008).

[13]Here I frame the discussion maintaining the terminology commonly used by philosophers. However a consequence of my view is that there are no "requirements" of rationality, except in the very weak sense in which whatever necessary or sufficient condition can be said to be a requirement. In other words, such "requirements" would be instances of what are often called *anankastic conditionals*, that is, conditionals expressing a necessary condition for a certain fact or

for example, the requirement to take the means to satisfy one's ends. If we assume that a practical norm is that one ought to satisfy one's ends,[14] an agent who does not take the means to satisfy her ends also fails to satisfy some of the regulation conditions considered above. For example, an agent intending to pursue an end, but failing to take the means necessary to that end, does not satisfy some regulation conditions relevant for being rational in that circumstance (either because of unresponsiveness to normative reasons or for some weakness of the will). Similar considerations are valid for the requirement to try to do what one believes that she ought to do. Someone believing that she ought (*all things considered*) to ϕ but who does not try to ϕ, does not satisfy some regulation condition necessary for rationality.

A specific type of rationality is *epistemic rationality*. Epistemic rationality concerns the regulation conditions of a norm constitutive of belief that requires believing only the truth.[15] An agent failing to satisfy the regulation conditions of this norm necessary for rationality is epistemically irrational (and by extension also the beliefs responsible for this failure are). Consider a specific requirement of epistemic rationality: if S believes that p and that p implies q, then S should not believe that not-q. My characterization of rationality explains why an agent believing that p, that p implies q and that not-q, is irrational. According to the truth-norm of belief, for any ϕ, one ought to believe that ϕ only if it is true that ϕ. However, if p and p implies q, then q. Therefore, if a subject S believes that p, that p implies q, and that not-q, she believes some falsity. If S recognizes and accepts the truth-norm of belief, then she violates some regulation conditions of the norm: S is unresponsive to normative reasons, or unable to reason as she intends and knows she should do.[16]

---

event being the case. Anankastic sentences are commonly used for expressing logical and natural (causal or physical) necessities. Such claims state mere conditions for the happening of some fact or event. An example of an anankastic conditional is "In order to go to Paris, one must take the 12:27 train".

[14] It could be argued that this norm is constitutive of what an end is. An end is in part something that ought to be satisfied, at least *pro tanto* and according to a pragmatic standard.

[15] Notice that not every norm whose condition of satisfaction is the possession of true beliefs is an epistemic norm. As some philosophers have shown (Owens 2003, Kelly 2003), there can be practical reasons for having true beliefs. In my view, epistemic rationality is related to a specific truth-norm constitutive of belief. This norm defines the limits of the epistemic domain, in the sense that an epistemic notion can be defined by means of some specific relation with this norm. Unfortunately I cannot develop this view here. For a similar view, see for example Wedgwood 2002. I think that substantially this is also the view of Engel (though he didn't explicitly argue for it).

[16] I am aware that the approach discussed here is just sketched and needs some important

Differences from other accounts of rationality

My account of rationality is substantially different from other well-known accounts of this notion. Some philosophers hold that rationality is a matter of norms, or normative reasons. For example, according to Wedgwood (2002, 2003, 2007, 2013), there are subjective norms of rationality. These norms would derive from objective norms of correctness governing mental attitudes and would allow regulating the latter. Also Engel seems to accept a similar view in some articles (e.g., Engel 2007a, 2007b, 2011). Other views do not rely on the distinction between subjective and objective norms, but still maintain that there are genuine norms of rationality separated by and independent of other practical and epistemic norms (e.g., Broome 1999).

Such views are affected by several problems that, for reasons of space, I cannot mention here.[17] I shall focus on some issues that make manifest the advantages of my account of rationality over these other views. These issues boil down to the intuition that rationality and norms are related in a peculiar way: the former seems to depend, to be secondary or parasitic on the latter. Here is an example: on the one hand the law requires one to pay taxes; on the other hand an agent is rational only if she tries to pay taxes when believes that she should pay them. These two claims seem to be related in obvious ways: the latter seems to suggest a means to satisfy the former. However, if we conceive standards of rationality as constituted by a set of norms, there can be genuine conflicts between the norms of rationality and other norms; consider again the example above: if someone wrongly believes that she should not pay a tax, she ought both to pay and not to pay that tax. In this case there should be an overt conflict between a norm of rationality and law. However, it seems clear that such a conflict of obligation is not the case: what someone in the described situation ought to do is to pay the tax, no matter what she believes. Of course, *for being rational*, one should do what one believes; however this *should* is not the expression of a norm. Rather, one should do what one believes *in order to* satisfy a condition necessary for rationality, in the same way in which temperatures should be close to zero degrees *in order to* snow. In general, we are never faced with the choice between acting as we "primarily" or "objectively" ought and acting as we rationally ought. We always ought to act as the unique norm requires, and we are rational if we satisfy certain

refinements. The present discussion aims to provide only a rough and tentative picture of how to conceive epistemic rationality in the present framework.

[17] For some of these objections see Kolodny 2007, Gluer &Wikforss 2009 pp.44-45, Dancy 2009 and Parfit 2011.

conditions necessary for acting as that norm requires.[18]

The account of rationality sketched here has some similarities with those suggested by Scanlon (1998), Dancy (2000, 2009), Kolodny (2005, 2007) and Parfit (2011). According to these philosophers, normative reasons and rationality are related in the following way: on the one hand reasons are not dependent on the subject's perspective, they are facts or true propositions. On the other hand what it is rational to do depends on what appears to be a reason from one's perspective. One is rational if one does what one believes there are reasons to do, or what one would have reason to do if one's beliefs were true. Like my account, these also assume that there are no independent norms of rationality, or normative reasons to be rational. Rather, for them rationality is a matter of conditions related to the perspective of agents engaging (or believing themselves to be engaged) with norms. My account agrees with these other accounts that rationality is a matter of doing what one believes that ought to do, or what one ought to do from her own perspective. But my account goes further, specifying that once an agent believes that she has a reason to *F*, the rationality of her response does not depend uniquely on whether that agent also *F*-es, but also on several other conditions such as her belief appropriately motivating and bringing her to the formation of an intention to fulfil the norm, the cognitive system functioning appropriately, the absence of akratic behaviours and judgments, and so on. In fact, believing that one has a reason to *F* and *F*-ing are not jointly sufficient for being rational; also the other conditions listed above must obtain.[19]

[18] For a similar objection see Dancy 2009.

[19] My account of rationality can also explain the seeming normative force of rationality in a way similar to that described by Kolodny 2007: when an agent believes to have a reason, as it seems to her, she has a reason to have that attitude. The normative pressure an agent feels to act rationally derives from how things seem to her – the reasons that, as it appears to her, she has. My account, differently from Kolodny's, can also account for the normative force of third person criticisms of irrationality not consisting in advices but in external assessments. If all the seeming normative force of rationality were reduced to an internal phenomenon, to what it appears to one "from the inside", it would be hard to account for the supposed normative force of judgments of irrationality about people that from their perspective do not feel any pressure at all (for example, because akratic). On the contrary, according to my account, one's perspective is only one of many factors relevant to determine rationality; other ones are the appropriate responsiveness to the normative force of norms and reasons and the absence of akrasia. Norms give reasons to act in certain ways, and a condition for agents to be able to comply with these norms is that they feel normative pressure when they take these norms as reason-providing, are motivated by this pressure in the right way, and act accordingly.

3. Normative and norm-relative assessments

To the distinction between norms and their regulation corresponds a distinction between two types of assessment. A first type is relative to the satisfaction conditions of a norm. On the one hand an agent violating a norm is, for this very reason, subject to a criticism, even if she was motivated to follow the norm and tried to do that. For example, an agent who recognized and accepted a norm and tried to comply with it, but failed to comply because she ignored whether the satisfaction conditions of the norm obtained, though excusable for her infraction, is still criticizable for violating the norm. Similarly, an agent not paying some tax because she didn't know that she must do it, is maybe excusable and not blamable; nevertheless, she violates a norm and is potentially punishable for this infraction. On the other hand, an agent complying with a norm is free from criticism and punishment for norm-infraction, regardless of how she complied with the norm – whether she acknowledged the norm, accepted it and tried to comply with it, or she complied with it by mere chance. A distracted driver who doesn't see a red traffic light, but stops at the light because she sees a friend on the side of the street and wants to talk with him, doesn't violate the norm and is not subject to criticism or punishable for norm-infraction.

The second type of assessment is relative to the conditions of norm-regulation (or to a proper subset of them). Consider again the example of the distracted driver who stops at the traffic light but ignores that the light is red. Though she is not criticizable for violating a norm, she can be subject to criticism and blame for not stopping for the right reason (i.e., for not having paid attention to the light). Criticism and blame here concern the non-satisfaction of some regulation condition of the law.

Here an obvious question arises: if the agent did what the norm requires, why should she be subject to any criticism and blame at all? After all, norms aim at being satisfied, and as long as one satisfies them, there should not be room for criticism or blame. This question can be answered in different ways. My favorite answer is the following:[20] though in such cases an agent does not violate a norm, it could have easily happened that she violated it. Criticisms for not complying with regulation conditions are evaluative judgments of the ways in which an agent acts in circumstances in which she is binded by a norm. These ways may be evaluated as inappropriate if they are not conducive to the satisfaction of a norm in normal circumstances, or in close

[20] For similar approaches see Dancy 2009 and Millar 2009.

possible worlds where something is slightly different from how it is in the actual world and the agent could have easily failed to satisfy that norm. In the example of the driver, it could have easily been the case that the driver didn't see her friend on the side of the street and passed through the red light. The driver is therefore criticizable and blamable because her behavior was imprudent. Her distraction could have easily lead to the violation of the norm. In general, it is justified to blame someone for not satisfying certain conditions of norm-regulation (at least those conditions whose satisfaction is under the voluntary control of the agent), regardless of whether the norm is satisfied or not.

Both types of assessments depend in some measure on some norm: the assessment relative to satisfaction conditions is obviously and directly dependent on a norm. But also the assessment relative to regulation conditions is dependent on a norm, even if indirectly, for i) it properly bears on the satisfaction conditions of a norm in close possible worlds and ii) it concerns regulation conditions, and, as said in §2, there is no norm-regulation without some norm. In this respect, we can see norm-regulation, regulation-dependent notions such as rationality and justification, and assessments relative to regulation, as byproducts of some norm, though not normative features in themselves.

A consequence of the above considerations is that, though normative assessments (criticisms, justifications, judgments, excuses, and so on) depend on some norm, the object of these assessments (what is criticized, excused,...) does not necessarily bear on the satisfaction conditions of a norm. A criticism for not *F*-ing is not an argument for the claim that there is a norm requiring one to *F*.[21] That criticism only shows that *F*-ing is either a satisfaction condition of a norm, or one of the necessary conditions or main sufficient conditions for the regulation of a norm. A driver passing through a red light can be criticized and blamed for not seeing the light. However, the law doesn't require one to see a red light; it requires one to stop when the light is red. Seeing the red light is a regulation condition, not a satisfaction condition of the law; the driver is criticized for being careless, for not having fulfilled a regulation condition – and only indirectly for having violated the law, to the extent that the claim involves a presupposition that the driver also passed through the red light. Similarly, consider the distracted but lucky driver criticized for not hav-

[21] This argumentative strategy has been widely used recently by philosophers to argue that assertion and action are subject to epistemic norms of knowledge or justification. See, for example, Williamson 2000 and Hawthorne and Stanley 2005. In my view, the present considerations partially weaken the force of this strategy.

ing seen the red light, even if for some other reason she stopped at the light. In this case, though the driver complied with the norm (and, consequently, cannot be criticized for a norm infraction), she is subject to criticism for not satisfying some regulation condition.

4. Conclusion

In this paper I proposed a specific account of the distinction introduced by Engel between norm and norm-regulation. Then I provided an account of rationality as the obtaining of a subset of regulation conditions of some norm. I argued that, according to these accounts, both norm-regulation and rationality are not normative notions. Finally I discussed the nature of assessments relative to norms and norm-regulation. The accounts introduced here are only sketched. My aim here was just to develop some thoughts inspired to me by the works of Engel. Though only sketched, I consider the approaches to normative guidance and rationality described in this article particularly promising and deserving of further consideration in future works.

5. References

Boghossian, P. A. (2003). The Normativity of Content. *Philosophical Issues*, 13: 31–45.

Broome, J. (1999). Normative Requirements. *Ratio*, 12(4): 398–419.

Bykvist, K. & Hattiangadi, A. (2007). Does Thought implies Ought? *Analysis*, 67: 277-285.

Dancy, J. (2000). *Practical Reality*. Oxford University Press.

Dancy, J. (2009). Reasons and Rationality. In Robertson, S., editor, *Spheres of Reason*, pp. 93–112. Oxford University Press.

Engel, P. (2007a). Belief and Normativity. *Disputatio*, 23 (2): 153–177.

Engel, P. (2007b). Epistemic Norms and Rationality. In W. Strawinski, ed. *Festsschrift for Jacek Jadacki*, Springer.

Engel, P. (2008). In What Sense is Knowledge the Norm of Assertion? *Grazer Philosophische Studien*, 77(1):45–59, 2008.

Engel, P. (2011). Epistemic Norms, in S. Berneker & D. Pritchard, eds, *The Routledge Companion to Epistemology*, London.

Engel, P. (2014). In Defense of Normativism about the Aim of Belief. In T. Chan (ed.), *The Aim of Belief*. Oxford: Oxford University Press, pp. 43-85

Gibbard, A. (2005). Truth and Correct Belief. *Philosophical Issues*, 15: 338–350.

Glüer, K and Pagin, P. (1999). Rules of Meaning and Practical Reasoning. *Synthese*, 117(2): 207–227.

Gluer, K. & Wikforss, A. (2009). Against Content Normativity. *Mind*, 118: 31-70.

Hawthorne, J. and Stanley, J. (2010). Knowledge and Action. *Journal of Philosophy*, 105 (10): 571-590.

Kolodny, N. (2005). Why be rational? *Mind*, 114(455): 509–563.

Kolodny, N. (2007). Ix-how does coherence matter? *Proceedings of the Aristotelian Society*, 107: 229–263.

Menzies, P. (2004). Difference-Making in Context. In Collins J. D., Hall N., and Paul L. A. (Eds.) *Causation and Counterfactuals*. MIT Press, pp. 139–80.

Millar, A. (2009). How Reasons for Action Differ from Reasons for Belief. In Robertson, S., editor, *Spheres of Reason*. Oxford University Press.

Owens, D. J. (2003). Does Belief Have an Aim? *Philosophical Studies*, 115(3): 283–305.

Parfit, D. (2011). *On What Matters*. Oxford University Press.

Railton, P. (2006). Normative Guidance. In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*, Vol. 1. Oxford University Press.

Scanlon, T. (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Schroeder, M. (2008). Value Theory. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.

Shah, N. (2003). How Truth Governs Belief. *Philosophical Review*, 112: 447-82.

Steglich-Petersen, A. (2006). The Aim of Belief: No Norm Needed. *The Philosophical Quarterly*, 56(225): 500-516.

Steglich-Petersen, A. (2010). The Truth Norm and Guidance: A Reply to Gluer and Wikforss. *Mind*, 119(475): 749–755.

Wedgwood, R. (2002). The Aim of Belief. *Philosophical Perspectives*, 15: 267-297.

Wedgwood, R. (2003). 'Choosing Rationally and Choosing Correctly', in *Weakness of Will and Practical Irrationality*, Sarah Stroud and Christine Tappolet (ed.). Oxford University Press, pp. 201-229.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford University Press.

Wedgwood, R. (2013). Doxastic Correctness. *Proceedings of the Aristotelian Society*. Supplementary Volume, 87: 217-234.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press.

43

What does intentional normativism require? *

DANIEL LAURIER

Abstract Most people who have discussed the question whether attributions of intentional attitudes or contents are normative have assumed that this question boils down to the question whether such judgements have normative force "by themselves", or as it is often put, to the question whether they are "intrinsically" or "non-hypothetically" normative. I take issue with this and argue that a judgement can be normative, in the sense of essentially involving a normative concept, even if its normative force is "extrinsic", and even if it lacks normative force altogether. The result is that most attempts to show that attributions of attitudes or contents lack normative force, even if successful, could not count as refutations of intentional normativism.

Keywords Intentionality, normativity, attitude, content, normative force, normative content

* A previous version of this paper has been read in Prague, at the conference Normativity and Meaning: Sellarsian Perspectives, in May 2011.

1. Introduction

Following the works of Kripke and Davidson, there has been much controversy around the claim that intentional/semantic facts/judgements are "intrinsically" or "constitutively" normative. People have wanted to determine, not merely whether it is true or false, but also (and perhaps mainly) whether it is compatible with the program of naturalizing intentionality. It is fair to say that, on the whole, and unsurprisingly, naturalists have tended to reject this claim[1], and their opponents, to endorse it. Given that there is much disagreement, in the field of metaethics, on whether the normativity of moral properties is or isn't an obstacle to moral naturalism, it seems however premature to take any definite stance on whether intentional normativism is or isn't compatible with intentional naturalism before the exact content of the former has been clarified[2]. It is to this preliminary task of clarification that I hope to contribute in this paper.

Given that it is generally admitted that there is a distinction to be made between linguistic and non-linguistic intentionality (as well as between "mood" and content), it should first be observed that the debate about intentional normativism actually covers four different questions: (i) whether illocutionary acts are normative, (ii) whether linguistic meaning (i.e., meaning, as it attaches to words and sentences) is normative, (iii) whether intentional (non-linguistic) attitudes are normative, and (iv) whether intentional content (i.e., content, as it attaches to intentional, non-linguistic, attitudes) is normative. Until recently (Boghossian 2003, 2005, Bykvist and Hattiangadi 2007), discussion of these issues has mainly focused on linguistic intentionality (and even more narrowly, on linguistic meaning) but I will here be concentrating on non-linguistic intentionality.

There are two reasons to put non-linguistic intentionality at the forefront. One is that I'm assuming that language depends on thought in a sufficiently strong sense for it to be reasonable to expect that non-linguistic intentional attitudes/contents could not turn out to be normative without illocutionary acts and linguistic meanings also turning out to be normative. The other stems from the fact that, in my opinion, it is a mistake to think that Kripke's sceptical problem about rule-following is exclusively, or even primarily, con-

[1] There are exceptions. For example, Boghossian (2005) argues that naturalists should actually welcome the normativity thesis.

[2] This is not to say that intentional naturalism doesn't stand in need of any further clarification. But there seems to be less disagreement about what natural facts are than about what normative facts are. In any case, the focus here will be on normativity.

cerned with the question: what does the fact that certain words mean certain things in some *public* language consists in? Kripke is indeed quite explicit that, to suggest that someone's meaning addition by '+' consists in her/his having the intention, when using this symbol, to apply the addition function would immediately raise the objection that it has not yet been said what it is for an intention to be an intention to apply the addition function rather than the quaddition function. Thus, what is supposed to be normative, in his discussion, is not the fact that the symbol '+' represents the addition function in the language of some community, but the fact that it represents it *for the speaker*. It would seem to follow that the (often made) observation that the fact that a word means a certain thing in the language of some community can have normative force only for those who are already motivated to communicate in this language, though quite right (and perhaps sufficient to show that linguistic meaning facts aren't "intrinsically" normative[3]), is also quite besides the point.

Intentional normativism has been interpreted in a variety of ways, which I am not going to rehearse, but one thing which can be taken for granted is that, since the normative/non-normative[4] distinction is primarily a distinction between sorts of judgements (propositions) or states-of-affairs, and secondarily a distinction between sorts of concepts or properties, to say that intentional attitudes/contents are normative can only be a relaxed way of saying that intentional judgements (that is, attributions of intentional attitudes and/or of intentional contents), or the corresponding states-of-affairs, are normative, and/or that the concepts of intentional attitudes (such as the concepts of belief, of desire and of intention) and of intentional content, or the corresponding properties[5], are normative. Obviously, it would not make much sense to suggest that intentional attitudes themselves are normative, unless they are taken to be properties (which is in accordance with the previous remark), while to hold that intentional contents themselves are normative would be tantamount to claiming that all propositions are normative, and would thus abolish the contrast between the normative and the non-normative.

[3] Actually, I argue below that it can be sufficient only on the assumption that attributions of motivations themselves are normative.

[4] I prefer to contrast the normative with the non-normative, and not with the descriptive, insofar as it would be odd to talk of "descriptive facts" which could be opposed to normative facts, not to mention the fact that I don't want to exclude the possibility that some evaluative judgements (which are a kind of normative judgements) count as descriptive.

[5] For the sake of readability, I will henceforth drop all talk of properties and states-of-affairs, as nothing I will say depends on whether normativity is construed as pertaining primarily to judgements and concepts, or to states-of-affairs and properties.

Normativity is a huge and notoriously complex subject, about which there is much controversy; and I am far from being in a position to say exactly what we are claiming, when we claim that something is normative. On the other hand, we need to have some more definite idea of what is at stake in the dispute about intentional normativism, if it is to be amenable to rational adjudication.

It is fairly common to acknowledge that there are two basic varieties of normative judgements/concepts, namely, the deontic and the evaluative (or axiological). The deontic judgements deal with such things as obligations/permissions, oughts/mays and duties/rights, while the evaluative judgements are concerned with the Good and the Bad (and values in general). Both the deontic and the evaluative can be deployed along multiple and various dimensions: there are not only moral obligations/permissions, but also legal, prudential and perhaps rational ones, and there are not only moral values, but also prudential, aesthetic, and perhaps rational ones. There is much dispute about how these two basic sorts of normativity relate to each other, but the only point I want to make in this connection is that, although it may be possible to construe some evaluative judgements in such a way that they don't carry any deontic implication, discussions of intentional normativism have tended to focus exclusively on deontic normativity, and insofar as evaluative concepts have been appealed to, they have generally been construed as involving a deontic dimension (though not necessarily as being reducible to deontic concepts)[6]. I will conform to this practice in what follows, since I find it hard to think of anything that could be distinctive of evaluative judgements/concepts, once they are stripped of any deontic dimension (except perhaps that they are polar and scalar: they involve an opposition between a positive and a negative pole, and they concern quantities). Accordingly, from now on, I will use 'normative' to mean "having a deontic dimension".

2. Normative Force vs Normative Subject-Matter

Now, suppose you are given a list of judgements, and told that they are normative. You will naturally ask what is it that makes them normative, and probably feel less than satisfied if you are told that they are normative because they involve concepts from a certain list which is then handed to you. You will want to know what is it that makes these concepts normative, that is, whether they share any feature which gives them the "power" to make certain

[6] The concept of justification (or rational justification) may be a case in point here.

judgements count as normative, but is lacking in concepts which don't have this power. But it is hard to think of any such feature, besides their having this very power.

Now look at it from the opposite perspective. You are given a list of concepts, and told that they are normative. You ask what it is that makes them normative and you are given the answer that their normativity comes from the fact that some of the judgements in which they are involved (as constituents) have a certain special feature called 'normative force'. This is of course less than fully satisfactory, until we have been told what having this feature actually amounts to, which I am unfortunately unable to do. Yet I do submit that this kind of approach (which takes normativity to be, in the first instance, a feature of judgements/states-of-affairs) is more promising than the one contemplated in the previous paragraph and puts us on the right track.

The bottom line is that certain judgements just have (and strike us as having) normative force (what Millar 2004: 92-99 calls 'normative import'), in the sense that they cannot be true unless some people have what (taking my inspiration from Brandom 1994) I will call a normative status; that is to say, unless some people are obliged/permitted to do/think (or not to do/think) certain things, or unless there are (normative) reasons for some people to do/think (or not to do/think) certain things. Moreover, and most importantly, their having such a force cannot be completely accounted for by the mere fact that they involve certain specific concepts. This is one of the lessons of the familiar Frege-Geach point: the conditional judgement that if you ought to make it the case that P then Q doesn't have any normative force, even though its antecedent, when used on its own, normally has such a force. Likewise, and even more obviously, the judgement that 'X says/believes that you ought to make it the case that P' lacks normative force, even though the embedded judgement normally has one[7]. It is worth pointing out that insofar as 'X says

[7] However, the judgement 'X is a genuine authority and X says that you ought to make it the case that P' would seem to have normative force, probably because to say that X is a genuine authority is tantamount to saying that you ought to do/think whatever X says you ought to do/think, i.e., to saying that if X says that you ought to make it the case that P then you ought to make it the case that P. This makes the concept of "being a genuine authority" a normative concept, though not a normative force conferring one (see below). By the same token, it makes it plain that a conjunctive judgement (or a pair of judgements) can have normative force without any of its conjuncts (members) individually having normative force. Perhaps some will be tempted to suggest that 'X is a genuine authority' actually has normative force, and not merely normative subject-matter, on the ground that it entails that you ought to make it the case that (if X says that you ought to make it the case that P, then you ought to make it the case that P), or (more plausibly) that you ought top make it the case that (if X says that you ought to make it the

that you ought to make it the case that P' may sometimes be used interchangeably with 'X requires that you make it the case that P', not all uses of the latter have normative force, despite the fact that it doesn't overtly contain any embedded judgement which would have normative force when used on its own[8]. This clearly shows, not merely that some judgements may involve normative concepts without having normative force, but that no concept can be such that all judgements involving it have normative force. Following Millar (2004: 95), I will say that a judgement which involves some normative concept (without necessarily having normative force) has a normative subject-matter, or that it is normatively contentful.

Since the judgement that X believes that you ought to make it the case that P obviously attributes an intentional attitude, the foregoing also illustrates the fact that some intentional judgements unquestionably (and trivially) have normative subject-matter. But let's not jump to the conclusion that intentional normativism should therefore not be concerned with establishing that intentional judgements have normative subject-matter, but with establishing that they have normative force. As I will shortly be arguing, this conclusion must be resisted. It should instead simply be reminded that intentional judgements *attribute* attitudes (or contents): they are *about* them. The fact that the attitude (or content) attributed by some intentional judgement has normative subject-matter clearly is sufficient for the latter to also have normative subject-matter, but gives no support to the conclusion that the attitudinal concept involved (or the concept of content) is normative or contributes in any way to make this intentional judgement normatively contentful. That is to say, it gives no support to the conclusion that *all* intentional judgements which attribute the same "intentional mood" (or which attribute some intentional content or other) are normatively contentful, which is at least part of what intentional normativism requires. In a word, attributions of normatively contentful attitudes are irrelevant to this discussion and should simply be kept out of the way.

To say that no concept is such that it confers normative force to all judgements involving it is not, however, to deny that some concepts do contribute to make it the case that some of the judgements involving them have normative force. Clearly, the judgement that you ought to make it the case that P

case that P, then you make it the case that P). But it doesn't entail any such thing. If it did, then you could discharge your obligation just by preventing X from saying that you ought to make it the case that P!

[8] As far as I can see, the judgement that 'X requires that you make it the case that P' will only have normative force when 'X' stands for something like 'The fact that Q'. John Broome sometimes, but not always, uses 'require' in just this way.

would lack normative force if it didn't contain the concept 'ought'. This provides one reasonably clear sense in which a concept may count as normative, namely in virtue of the fact that it confers normative force to *some* of the judgements involving it. But, interestingly, the Frege-Geach point also suggests that this may not be the only way for a concept to qualify as normative. For just as a judgement which normally has normative force may be embedded in a judgement lacking normative force (which thereby counts as normatively contentful), a concept which normally confers normative force to the judgements involving it may itself be embedded in a concept which doesn't confer normative force to any of the judgements involving it (but which thereby counts as normatively contentful). Clear (if contrived and artificial) examples of this are provided by 'being such that if you ought to make it the case that P then Q' and 'being told that you ought to make it the case that P'.

By extension, it seems reasonable to hold (i) that a judgement which either has what some like to call an "internal relation" to a judgement having normative force, or couldn't be explained except in terms of some such judgement, also counts as normative in the subject-matter sense, and (ii) that a concept which either has an internal relation to a normative force conferring concept, or couldn't be explained except in terms of some such concept, also counts as normative in the subject-matter sense.

The important thing to bear in mind, for our purpose, is that a judgement which lacks any normative force may nonetheless "essentially" involve a normative force conferring concept, and that a concept which is not itself normative force conferring may nonetheless "essentially" involve a normative force conferring concept. This means that an intentional judgement could be normative in the subject-matter sense without having normative force, and that an intentional concept could be normative in the subject-matter sense without ever contributing to confer normative force to the judgements involving it (or in other words, without any of the judgements involving it having normative force, except in virtue of their involving some other normative concept). It is thus somewhat disappointing to have to record that nearly all discussions of intentional normativism that I know of have focused on the question whether intentional judgements have normative force[9]. For, even if having normative force or being normative force conferring is clearly sufficient for a judgement or concept to be normative, it's by no means necessary[10].

[9] A relevant sample would include Kripke (1982), Gibbard (1994), Hattiangadi (2006, 2007), Boghossian (2003, 2005), Glock (2005), Gluër (1999), Gluër and Pagin (1999), Horwich (1998, 2005), Whiting (2007) and Bykvist and Hattiangadi (2007).

[10] I owe special thanks to XXX for having forced me to clarify this point.

Clearly, I'm swimming against the tide (see footnote 9) in allowing intentional normativism to be read as a claim that intentional judgements/concepts are normative merely in the subject-matter sense, and it will probably be complained that I'm being too liberal. The only ground I can see for such a complaint is a widespread tendency to see normativism as being primarily a weapon against naturalism, and to assume that it could threaten the latter only if it is read as a claim that intentional judgements have normative force[11]. For my part, I take the question of normativism to be one which naturally arises when reflecting on the nature of intentionality and which is interesting in its own right, quite apart of the question of naturalism. But even from this point of view, it is hard to see why normativism couldn't be a threat (or at least, a challenge) to naturalism if it were understood as claiming that intentional judgements/concepts "merely" have normative subject-matter. How could anyone who thinks that the judgement that you ought to make it the case that P raises a problem for the naturalist also hold that the judgement that if you ought to make it the case that P, then Q (or its converse) raises no such problem? I would have thought that any full account of what it is for the latter judgement to be (non-trivially) true must involve an account of what the truth of the former consists in. In other words, it would seem that any full account of a normatively *contentful* judgement (or concept) is bound to rest in part on an account of the normatively *forceful* judgements (or normative force conferring concepts) that it involves and must ultimately involve.

3. Intrinsic and Categorical Normativity

There seems to be a consensus that the normativity thesis wouldn't have much interest if it said merely that intentional judgements are such that, with the help of appropriate auxiliary premises, they entail some overtly (or "basic") normative judgements. This, it is contended, would at most show that intentional judgements are "extrinsically" normative, while what has to be shown is that they are "intrinsically" (or "constitutively") normative, in the sense that they entail basic normative judgements "all by themselves".

[11] For example, Boghossian (2005: 217) writes that 'the philosopher with the most reason to believe [...] in the normativity of content is, ironically enough, the naturalist about content. But if this is the only way in which the normativity of content can turn out to be true, *it shows what an uninteresting thesis it is*' (my emphasis). See also Whiting (2007: 135) who clearly suggests that what is "required" is to show that intentional judgements are normative 'in a way that might pose problems for naturalism'.

Since most of the people involved in this debate have restricted their attention to the question whether intentional judgements have normative force, it could at first seem that this contrast (between intrinsic and extrinsic normativity) is meant to be between two different ways of having normative force. But as I have been using these notions, both normative force and normative subject-matter are "intrinsic" features of judgements. It may thus be somewhat unclear, at first sight, exactly how the intrinsic/extrinsic distinction, as applied to normative judgements, is supposed to relate to the force/subject-matter distinction. It seems, however, reasonable to expect that all and only judgements having normative subject-matter will count as being either intrinsically or extrinsically normative (and in particular, that a judgement could be extrinsically normative only if it is normatively contentful). But it seems that there is no way of drawing the intrinsic/extrinsic distinction which will ensure that this is the case.

Let's grant (with, I think, a majority of philosophers) that some judgements qualify from the start as manifestly having normative force, and thus count as "basic" normatively forceful judgements. Intuitively, these will be judgements which overtly report the fact that someone or other has a certain normative status, such as 'S ought/may make it the case that P', 'It is justified/unjustified for S to make it the case that P', 'There is reason for S to make it the case that P', or 'The fact that Q is a reason for S to make it the case that P' [12], etc., and they will automatically count as intrinsically normative. The general idea, then, might be that a judgement is intrinsically normative (i.e., intrinsically normatively forceful) if and only if it either is such a basic normatively forceful judgement or entails one *without the help of any auxiliary premise whatsoever*. Now, in my language, to say that some judgement entails another, but only with the help of some further premise, is tantamount to saying that it doesn't entail it (but the conjunction of this judgement and this further premise does), and to say that it entails it without the help of any further premise is tantamount to saying that it entails it. That being so, it is easy to see that (on this way of construing intrinsic normativity) a judgement will be intrinsically normative if and only if it has normative force.

[12] Clearly, the last two judgements are meant to be about normative reasons. It is beyond doubt that there is such a normative concept, and that it is normative force conferring, though there may be some uncertainty as to whether it belongs to the deontic or the evaluative. The fact that one's reasons for doing or not doing something may have more or less strength seems to bring the concept of a reason closer to the evaluative side; while the fact that what one ought to do could be defined as what one has most reason to do suggests that it has a deontic aspect.

Now, there are two different ways of defining extrinsic normativity which are both compatible with the foregoing construal of intrinsic normativity, but only one of which is really appealing, and it turns out to be such as to ensure that a judgement will be extrinsically normative only if it has normative subject-matter without having normative force. The first option would be to say that a judgement is extrinsically normative if and only if (i) it doesn't entail ("by itself") any basic normatively forceful judgement (i.e., it is not intrinsically normative) and (ii) there are judgements such that, in conjunction with them, it entails some basic normatively forceful judgement. Such judgements could then be said to have "extrinsic" normative force.

If extrinsic normativity is construed in this way, it is easy to understand why it has been thought that the claim that intentional judgements are extrinsically normative is uninteresting, since it is obvious that every judgement which is not intrinsically normative will then qualify as extrinsically normative. Consider the judgement that the sky is blue. It is not intrinsically normative, but in conjunction with the judgement that if the sky is blue, then Socrates ought to wash the dishes, it entails the intrinsically normative judgement that Socrates ought to wash the dishes. The trouble is that there obviously is no intuitive sense in which the judgement that the sky is blue could be said to be normative, while (I would have thought) the intrinsic/extrinsic distinction is meant to be a distinction between two sorts of normativity. On this reading, every judgement is normative, if not intrinsically, then extrinsically.

The second option would be to say that a judgement is extrinsically normative if and only if (i) it doesn't entail ("by itself") any basic normatively forceful judgement (i.e., it is not intrinsically normative) and (ii) there are judgements *which don't involve any normative concept*, such that, in conjunction with them, it entails some basic normatively forceful judgement. Clearly, on this construal, the judgement that the sky is blue no longer qualifies as extrinsically normative, since (as far as I can see) it is only in conjunction with judgements which have normative subject-matter that it could entail any normatively forceful judgement. On the other hand, the conditional judgement that if the sky is blue then Socrates ought to wash the dishes will still count as extrinsically normative, since it lacks normative force (i.e., it is not intrinsically normative), and in conjunction with the perfectly non-normative judgement that the sky is blue, it entails the normatively forceful judgement that Socrates ought to wash the dishes. This certainly is a more satisfying result. Since (as far as I can see) only a judgement with normative subject-matter could possibly entail a normatively forceful judgement with the help of judgements which don't have normative subject-matter, *only* such judgements will qualify

as extrinsically normative. However, on this construal of extrinsic normativity, not all judgements with normative subject-matter (and no normative force) will count as extrinsically normative[13]. For example, it is hard to see how the judgement that if Socrates ought to wash the dishes then he ought to beat his wife could be made to entail a normatively forceful judgement by the addition of premises lacking normative subject-matter.

Hence, this second construal of extrinsic normativity has the drawback that some judgements with normative subject-matter will be neither intrinsically nor extrinsically normative. Yet all extrinsically normative judgements will have normative-subject matter, and if what has been said above is correct, then this should mean that there is no more reason to refuse reading intentional normativism as a claim about extrinsic normativity than there is to refuse reading it as a claim about subject-matter normativity. In the end, then, the distinction between intrinsic and extrinsic normativity proves to be useless, and I recommend to simply drop it.

Here, I must pause to consider a potentially powerful objection to the claim I have made, in the course of the foregoing argument, to the effect that only a judgement with normative subject-matter could possibly entail a normatively forceful judgement with the help of further judgements none of which is normatively contentful. On some ways of construing the notion of entailment, a necessary truth is entailed by any judgement. If the claim that X entails Y is construed in some such way, e.g., as meaning that it is necessary that if X then Y, then the judgement that the sky is blue, which lacks normative subject-matter, will trivially entail the judgement that if S ought to make it the case that P then S ought to make it the case that P, which *has* normative subject-matter.

At first sight, it looks as if it could simply be replied that the latter judgement still lacks normative force, and the claim is that no *normatively forceful* judgement can be inferred from a set of judgements none of which has normative subject-matter. But this will work only insofar as it can be maintained that no normatively forceful judgement is necessarily true. As far as I can see, this can however be maintained, without having to deny that there are necessary truths of the form "For all S, S ought to make it the case that P". For such a universal judgement can be a necessary truth only if the variable S is understood as a restricted variable ranging over rational or human agents.

[13] Even if this is wrong, the point to be made in the next paragraph will stand, since the intrinsic/extrinsic distinction will then be equivalent to the distinction between being normatively forceful and being *merely* normatively contentful.

In which case, it really is a generalized conditional, which doesn't entail that there is someone who ought to make it the case that P, and thus strictly speaking lacks normative force.

Unfortunately, admitting that no normatively forceful judgement is necessary is not quite enough to answer the objection. For if it is necessary that for all S, if S is a rational/human agent then S ought to make it the case that P, and it is granted that the judgement that the sky is blue entails all necessary truths, then it will follow that, in conjunction with the premise that Socrates is a rational/human agent, it entails that Socrates ought to make it the case that P, which *is* normatively forceful. But the premise that the sky is blue will then be idle, since on these assumptions, the judgement that Socrates is a rational/human agent would already entail the judgement that he ought to make it the case that P, *without the help of any further premise* (while the judgement that the sky is blue would not). At this point, I think it will have to be agreed that the judgement that Socrates is a rational/human agent had been shown to have normative force (and *a fortiori*, normative subject-matter). For if this is denied, then it is hard to see how *any* judgement which is not overtly normative could nonetheless possibly be shown to be normative. Hence, the very idea of trying to show that intentional judgements are (or are not) normative by showing that they entail (or don't entail) overtly normative judgements presupposes a version of the "is doesn't entail ought" principle (or "IO principle"). Or in other words, it rests on the assumption (not only that no normatively forceful judgement is necessary, but also) that no (normatively contentful) judgement of the form 'for all S, if S satisfies such and such non-normative conditions, then S ought to make it the case that P' can be necessary[14]. If this assumption is untenable, then the whole dispute over intentional normativism is pointless[15].

Seen from another angle, the intrinsic/extrinsic distinction may look like a distinction between unconditional and conditional (subject-matter) normative judgements, which in turn may evoke the classical distinction between categorical and hypothetical norms. So it may be worth having a closer look at

[14] Actually, if there were necessary normatively forceful judgements, then there would have to be necessary judgements of this form; hence the ban on the former just follows from the ban on the latter.

[15] This probably is a real possibility, and it wouldn't be the first time that a philosophical dispute turned out to rest on a mistaken assumption. Yet I'm not going to defend this assumption here, as this would obviously carry us too far. It may also be possible to evade the difficulty raised in the text by simply insisting that the notion of entailment should not be construed in such a way that a necessary truth is entailed by any judgement.

these distinctions. Actually, the categorical/hypothetical distinction is quite different from the conditional/unconditional distinction, but it doesn't correspond to the intrinsic/extrinsic distinction, any more than the latter does.

First, the conditional/unconditional distinction doesn't match the intrinsic/extrinsic distinction, if only because some unconditional judgements have normative subject-matter without having normative force (i.e., without being intrinsically normative), and (as just pointed out) some conditional judgements have normative subject-matter without being extrinsically normative. But neither does it match the categorical/hypothetical distinction, since (to the best of my knowledge) the latter is generally understood as pertaining specifically to the relation between normative force and motivation.

The distinction between categorical and hypothetical "norms" can, I think, be seen as a distinction between two kinds of normatively forceful judgements, insofar as a norm can be identified with a true normatively forceful judgement. It's a distinction between two ways of being normatively forceful. If this is right, then it couldn't possibly be a distinction between conditional and unconditional judgements, since no conditional judgement is normatively forceful, or correspond to the distinction between intrinsic and extrinsic normativity, since all normatively forceful judgements are intrinsically normative.

To be sure, some conditional judgements must be involved in explaining this distinction, since a normatively forceful judgement will be categorically normative when its truth doesn't *depend* on any specific motivation that the relevant agents might contingently have, and hypothetical otherwise. Suppose that Socrates ought to wash the dishes but that it couldn't possibly be the case that he ought to do that if he didn't have some contingent motivation (such as wanting to do it, or wanting the dishes to be clean, etc.), by which I mean, if he didn't have some specific motivation over and above any motivation which might be thought to be "constitutive" of rational or human agents as such (if indeed there is any such thing). That is to say: Socrates ought to wash the dishes, and it is necessarily the case that Socrates ought to wash the dishes only if he has some appropriate motivation. The judgement that Socrates ought to wash the dishes will then have "hypothetical" normative force. Accordingly, it will have "categorical" normative force if it is not necessarily the case that Socrates ought to wash the dishes only if he has some appropriate motivation (i.e., if the fact that Socrates ought to wash the dishes doesn't entail that he has any specific motivation)[16].

[16] It is worth stressing that, although the contrast between hypothetical and categorical norms

Now, on the assumption that intentional normativism must be construed as claiming that *all* intentional judgements have normative force, it will easily be seen that it could simply not be true unless intentional judgements turned out to be *categorically* normative (i.e., unless their normative force were independent of any particular agent's contingent motivations). Thus, it would be redundant to insist that intentional normativism should be dealing with "categorical" normativity, once it had been conceded that it is concerned with normative force and not with "mere" normative subject-matter. But this is not something we have to concede. Moreover, even if I'm wrong in thinking that this would be redundant, it is hard to see why anyone should want to insist that intentional normativism must be concerned with categorical normativity, since it is hard to see why a judgement's being "hypothetically" normative should make it any *less* normatively forceful than its being "categorically" normative.

I conclude that it is sufficient (and necessary), in order to vindicate intentional normativism, to establish that (all) intentional judgements have normative subject-matter, which is equivalent to establishing that intentional concepts are normative, in the sense that they (either are or) involve normative force conferring concepts[17]. For a judgement to satisfy this condition, it is sufficient (assuming the IO principle), but *not* necessary, that it entails, possibly in conjunction with auxiliary premises involving no normative concept, some "basic" normatively forceful judgement.

This is not to say, however, that I am conceding that *no* intentional judgement is normatively forceful. At this stage, it is still an open possibility that *at least* attributions of intentional attitudes have normative force (and that the concepts of intentional attitudes are normative force conferring). Should this prove to be the case, one could then argue that, even though attributions of intentional *content* lack normative force, the very concept of content can't be explained except in terms of the attitudinal concepts, and should therefore be counted as normative[18].

---

is usually discussed in relation to obligations and duties, it is here understood as pertaining to (deontic) normativity in general. Any normatively forceful judgement could be either categorically or hypothetically normative, including, for example, the judgements that there is reason for Socrates to wash the dishes, or that Socrates is justified in washing the dishes. From this point of view, it looks like a version of Williams' distinction between internal and external reasons.

[17] I grant that it has not been made perfectly clear when a concept "involves" another. The idea is that concept A "involves" concept B when either (i) B is, at some level of embedding, a constituent of A, or (ii) no full explanation of A can dispense with B.

[18] Basically, this would be an instance of the kind of strategy deployed by Brandom (1994), discussed in Boghossian (2005), and alluded to by Bilgrami (2004).

4. Concluding Remarks

Let me conclude by briefly examining a case which illustrates some of the prejudices I have been arguing to be mistaken. Boghossian (2005: 207) writes:

> Of course, we can say that, if you mean addition by '+' and have a desire to tell the truth, then, if you are asked what the sum of [57 and 68] is, you should say '125'. But that is mere hypothetical normativity, and that is uninteresting: every fact is normative in this sense. (Compare: if it's raining, and you don't want to get wet, you should take your umbrella.)
>
> If there is to be an interesting thesis of the normativity of meaning, we ought to be able to derive a should or an ought from the mere attribution of meaning to someone and without having to rely on any auxiliary desires that that person may or may not have.

It is fairly clear from this passage that Boghossian here uses 'hypothetical normativity' to refer to what I have called 'extrinsic normativity', and that he understands the latter in something very close to the first of the two senses I have distinguished above. But this is not the point I want to bring out (it has already been made). There is a further point to make here, which is that the argument which Boghossian is putting forward clearly backfires in its own terms, at least if it is meant[19] to apply to the normativity thesis as it pertains not only to linguistic meaning, but also to attributions of intentional attitudes.

Replacing the rather dull example which Boghossian gives in his parenthetical remark with the (slightly) less boring one I have recently been using, his point seems to be that there is an obvious but uninteresting sense in which the judgement that the dishes are dirty could count as normative, since by adding the premise that Socrates wants them to be clean, we could (no doubt, only with many further assumptions) reach the (normatively forceful) conclusion that Socrates should, or ought to, wash the dishes. I don't think for a minute (and probably neither does Boghossian) that it is actually correct to reason in this way: the dishes are dirty, Socrates wants the dishes to be clean, therefore, Socrates ought to wash the dishes. But let's pretend that it is[20].

[19] As I think it is, since the paper in question does discuss the normativity of intentional attitudes/contents.

[20] Presumably, what Boghossian has in mind is something like: Socrates believes that the dishes are dirty, Socrates wants them to be clean, therefore, he has some reason to wash them. But it's not beyond question that such an inference is any good either, given how people like Broome and Dancy use the concept of a reason.

Then, if I'm right that a normatively forceful judgement couldn't follow from a set of judgments unless one of them has normative subject-matter, and given that the judgement that the dishes are dirty clearly has no normative subject-matter[21], it would follow that the (intentional) judgement that Socrates wants the dishes to be clean must have normative subject-matter, which could apparently be the case only if the concept of wanting is normative. But if the concept of wanting is normative, according to what has been said above, it must be a (or "involve" some) normative force conferring concept. Hence, examples such as these could hardly be used to undermine the claim that intentional judgements are normative.

It is worth stressing that the reply to this can hardly be to *deny* that a normatively forceful judgement couldn't follow from a set of judgements unless one of them has normative subject-matter, for denying this would mean that the judgement that Socrates wants the dishes to be clean could fail to have normative subject-matter even if it entailed a normatively forceful judgement "all by itself". To insist that this judgement would nonetheless have normative force would then require denying (something I would have thought to be obvious) that only a judgement with normative subject-matter can have normative force. Since this is unacceptable, the only remaining option would be to maintain that a judgement cannot prove to be normative in virtue of its entailing a "basic" normatively forceful judgement, i.e., that all normatively forceful judgements are "basic". This would be a very dogmatic way of bringing this debate to an end, since it would amount to a *declaration* that intentional judgements are not normative[22].

As I said, I don't think the kind of inferences we have just been considering are strictly any good. So I'm not suggesting that they actually support intentional normativism. But this is not, as Boghossian contends, because mere "hypothetical" Normativity isn't interesting. They actually point towards another kind of difficulty, which arises when we contemplate the possibility that some intentional judgements (perhaps all) may entail normatively forceful ones only with the help of further *intentional* judgements.

[21] Well, perhaps the judgement that the dishes are *dirty* could be seen as a value judgement. If this puzzles you, then simply replace it with the judgement that the window is open, and assume Socrates wants it to be closed.

[22] More precisely, it would amount to this on the assumption that the only way to show that a certain judgement is normatively contentful is by showing that it entails (possibly in conjunction with auxiliary premises involving no normative concept) some "basic" normatively forceful judgement.

Consider the following piece of reasoning, which many will take to be (approximately) correct: Socrates believes that the dishes are dirty, he wants them to be clean, therefore, there is some reason for Socrates to make it the case that the dishes are clean. In light of what has been said so far, the correctness of this reasoning could provide evidence either that the judgement that Socrates believes that the dishes are dirty is normative (if the judgement that he wants the dishes to be clean is assumed to lack normative subject-matter), or that the judgement that Socrates wants the dishes to be clean is normative (if the judgement that he believes that the dishes are dirty is assumed to lack normative subject-matter). But there seems to be no way in which it could support both conclusions! This suggests that, for the purpose of arguing that (all) intentional judgements have normative subject-matter, it would be either question-begging or self-defeating to show that they entail some normatively forceful judgements, but only with the help of further intentional judgements. If auxiliary premises are needed, they must not only lack normative subject-matter, they cannot be allowed to be intentional judgements. It would lessen this difficulty if an argument could be found, to the effect that no intentional judgement (or no intentional judgement of a certain kind) can have normative subject-matter unless they all have[23]. But even on this assumption, one would still need to find at least *one* kind of intentional judgement which entails a normatively forceful one without assuming the truth of any other intentional judgement.

I am not sure what the prospects actually are, for showing that intentional judgements are normative *by* showing that (either by themselves or with the help of non-normatively contentful and non-intentional auxiliary premises) they entail normatively forceful judgements. But if what I have been saying is correct, then, should this prove to be impossible, or too unlikely, there will still be other ways of arguing that they nonetheless are normatively contentful.

[23] Attributions of belief naturally come to mind here. Some people have indeed speculated that it suffices to show that attributions of belief are normative in order to be in a position to argue that all intentional judgements (and attributions of content in particular) are. See, e.g., Bilgrami (2004) and Boghossian (2003, 2005). Brandom (1994) could also probably be read as holding this sort of view. However, it would, *prima facie*, seem to be more likely that attributions of desire entail normatively forceful judgements without the help of any other intentional judgement, than that attributions of belief do.

5. References

Bilgrami, Akeel (2004) 'Intentionality and Norms', De Caro and Macarthur *eds.* (2004) 125-151

Boghossian, Paul (2003) 'The Normativity of Content', Sosa and Villanueva *eds*. (2003) 32-45

Boghossian, Paul A. (2005) 'Is Meaning Normative?', Nimtz and Beckermann *eds.* (2005) 205-218

Brandom, Robert B. (1994) *Making it Explicit*, Cambridge (Mass.), Harvard U. Press

Broome, John (1999) 'Normative Requirements', Dancy *ed*. (2000) 78-99

Bykvist, Krister and Anandi Hattiangadi (2007) 'Does Thought Imply Ought?', *Analysis* 67, 277-285

Dancy, Jonathan *ed*. (2000) *Normativity*, Oxford, Blackwell

De Caro, Mario and David Macarthur *eds*. (2004) *Naturalism in Question*, Cambridge (Mass.), Harvard U. Press

Dretske, Fred (2000a) 'Norms, History and the Constitution of the Mental', Dretske (2000b) 242-258

Dretske, Fred (2000b) *Perception, Knowledge and Belief*, Cambridge, Cambridge U. Press

Engel, Pascal (2000) 'Wherein Lies the Normative Dimension in Meaning and Mental Content?', *Phil. Studies* 100, 305-321

Gampel, E. H. (1997) 'The Normativity of Meaning', *Phil. Studies* 86, 221-242

Gibbard, Allan (1994) 'Meaning and Normativity', Villanueva *ed.* (1994) 95-115

Gibbard, Allan (2002) "Normative and Recognitional Concepts", *Phil. and Phenom. Research* 64, 151-167

Gibbard, Allan (2003) 'Thoughts and Norms', Sosa and Villanueva *eds*. (2003) 83-98

Gibbard, Allan (2005) 'Truth and Correct Belief', Sosa and Villanueva *eds* (2005) 338-350

Glock, Hans-Johann (2005) 'The Normativity of Meaning Made Simple', Nimtz and Beckermann *eds.* (2005) 219-241

Gluër, Kathrin (1999) 'Sense and Prescriptivity', *Acta Analytica* 14, 111-128

Glüer, Kathrin and Peter Pagin (1999) 'Rules of Meaning and Practical Reasoning', *Synthese* 117, 207-227

Hattiangadi, Anandi (2006) 'Is Meaning Normative?', *Mind and Language* 21, 220-240

Hattiangadi, Anandi (2007) *Oughts and Thoughts*, Oxford, Clarendon Press

Horwich, Paul (1998) *Meaning*, Oxford, Oxford U. Press

Horwich, Paul (2005) *Reflections on Meaning*, Oxford, Oxford U. Press

Kripke, Saul (1982) *Wittgenstein on Rules and Private Language*, Oxford, Blackwell

McLaughlin, Brian P. and Jonathan Cohen *eds.* (2007) *Contemporary Debates in Philosophy of Mind*, Oxford, Blackwell

Millar, Alan (2002) 'The Normativity of Meaning', O'Hear *ed*. (2002) 57-73

Millar, Alan (2004) *Understanding People: Normativity and Rationalizing Explanation*, Oxford, Oxford U. Press

Nimtz, Christian and Ansgar Beckermann *eds.* (2005) *Philosophy-Science-Scientific Philosophy*, Paderborn, Mentis

O'Hear, Anthony ed. (2002) *Logic, Thought and Language*, Cambridge, Cambridge U. Press

Sosa, Ernest and Enrique Villanueva *eds.* (2003) *Philosophical Issues 13: Philosophy of Mind*, Oxford, Blackwell

Sosa, Ernest and Enrique Villanueva *eds.* (2005) *Philosophical Issues 15: Normativity*, Oxford, Blackwell

Villanueva, Enrique *ed*. (1994) *Philosophical Issues 5: Truth and Rationality*, Atascadero, Ridgeview

Wallace, R. Jay *et al. eds* (2004) *Reason and Value*, Oxford, Oxford U. Press

Wedgwood, Ralph (2007a) *The Nature of Normativity*, Oxford, Oxford U. Press

Wedgwood, Ralph (2007b) 'Normativism Defended', McLaughlin and Cohen *eds.* (2007) 85-101

Whiting, Daniel (2007) 'The Normativity of Meaning Defended', *Analysis* 67, 133-140

Wikforss, Asa Maria (2001) 'Semantic Normativity', *Phil. Studies* 102, 203-226

Williams, Bernard (1981a) 'Internal and External Reasons', Williams (1981b) 101-113

Williams, Bernard (1981b) *Moral Luck*, Cambridge, Cambridge U. Press

44

How Meaning Might Be Normative *

ALAN MILLAR

1. The topic

My aim here is (i) to outline an account what it is to grasp the meaning of a predicative term, and (ii) to draw on that account in an attempt to shed light on what the normativity of meaning might amount to. Central to the account is that grasping the meaning of a predicative term is a practical matter—it is knowing how to use it correctly in a way that implicates having an ability to use it correctly. This calls for an examination of what it is to use a term correctly. Two quite different types of correctness are liable to be conflated. In sections 2 and 3 I show why they must be kept apart. In the sections 4 and 5 I consider how correctness of the second type might be conceived within a practice-theoretic framework and how that framework might make sense of the idea that meaning is essentially normative. In the concluding section I respond to an objection.[1]

*I am grateful for discussion of an earlier version of this paper at a research seminar at Stirling at which Philip Ebert, Colin Johnston, Peter Milne, Walter Pedriali, Ben Saunders, and Alexander Stathopoulos were especially helpful. Thanks also to Jonathan Dancy for written comments on an earlier version and to Walter Pedriali for written comments that prompted me to introduce much needed clarification at a late stage.

[1] Throughout I shall be building on, and I hope improving, ideas set out in Millar 2002, 2004 and 2011.

2. Meaning and correct application conditions

One type of correctness is correctness of application. To apply the term 'dog' to something is to predicate it of that thing. Thus I apply 'dog' to a thing if and only if, using that term, I say of it that it is a dog. That application will be correct if and only if what is thus said is true of the thing to which it is applied. If it is correct in this sense it is a true application. Correctness of the second type concerns use more generally: it is *use in keeping with the term's meaning*.[2] Simplifying somewhat, a use—perhaps an application—fails to be in keeping with a term's meaning if the speaker uses the term in a manner that fails adequately to respect its conditions of correct (= true) application. I say more about that in the next section. Here I consider the relation between meaning and conditions of correct (= true) application.

A meaning of a predicative term is fixed by a concept that the term can express. Conditions of correct (= true) application of a term display or exhibit a meaning that it has provided that, by employing the very concept that fixes that meaning, they spell out necessary and sufficient conditions for the term's correct application. For instance, in the sense in which it stands for a type of bird,

(1) 'goldfinch' correctly applies to a thing if and only if it is a goldfinch.[3]

This specifies the sort of thing to which the term correctly applies when used in this sense. It does this by means of the very concept that fixes its meaning if so understood. Even if true, other bi-conditionals spelling out conditions of correct application will not serve this purpose unless they do likewise. For instance,

(2) 'goldfinch' correctly applies to a thing if and only of it is a bird of the species described on p. 280 of the 1981 edition of *Field Guide to the Birds of Britain*

will not serve the purpose.

[2] See McDowell 1984 and McGinn 1984: 60 for similar expressions.

[3] Expressing a closely related idea, Michael Dummett says, 'In a case in which we are concerned to convey, or stipulate, the sense of an expression, we shall choose that means of stating what the referent is which displays the sense: we might borrow a famous pair of terms from the *Tractatus*, and say that, for Frege, we *say* what the referent of the word is, and thereby *show* what its sense is' (Dummett 1973: 227). John McDowell's (1977) treatment of proper names is, I take it, one way of developing that idea for the case of those names.

To those who share the language in which it formulated, and who know what goldfinches are, (1) is liable to seem trivial. To those who have no idea what goldfinches are it would be uninformative. This might tempt one to suppose that a better formulation of correct application conditions would provide more information as to what goldfinches are. But (1) is not meant to assist someone who did not know what goldfinches are to understand the sort of thing to which the term correctly applies. It simply exhibits the meaning of the term in that it spells out necessary and sufficient conditions for the term's correct application by means of the very concept that fixes its meaning.

People who have some grasp of the meaning of the term 'goldfinch' in the sense in which it stands for a kind of bird, thus some grasp of the concept the term expresses when so understood, might differ in their conceptions of what it is to be a goldfinch. Some might have little more than a perceptual-recognitional grasp in that they can visually recognize goldfinches as goldfinches. Others might have a rich conception of what it is to be a goldfinch. Yet others might know that goldfinches are birds but not know much else or even how to recognize goldfinches by sight. Possessing the concept the term expresses, is compatible with having any of a range of different conceptions of what goldfinches are.[4]

What about the conditions of correct application of synonyms like 'chews' and 'masticates'? Their conditions would be, respectively,

(3) '*** masticates —' correctly applies to an ordered pair if and only if the first element of the pair masticates the second element of the pair.

(4) '*** chews —' correctly applies to an ordered pair if and only if the first element of the pair chews the second element of the pair.

That 'chews' and 'masticates' are synonyms is reflected in the fact that (3) would be true if 'masticates' in its right-hand side were substituted by 'chews' and (4) would be true if 'chews' in its right-hand side were substituted by 'masticates'. For all that, one could grasp the sense of 'chew' while having no grasp of the meaning of 'masticate', and *vice versa*, which is why it can be informative to learn that to masticate is to chew. It is no surprise that a sentence like, 'No one doubts that to chew is to masticate' is false even though 'chew' is synonymous with 'masticate'. One might doubt that to chew is to masticate simply because one grasps the meaning of 'chew' while having no idea of what 'masticate' means.

[4] We could add '(the bird)' to the end of (1) without committing ourselves to supposing that *all* who grasp the sense of 'goldfinch' must know that goldfinches are birds.

3. Use in keeping with meaning

The second type of correctness is that of use in keeping with meaning. Correct use in this sense is a use of the term that respects the relevant conditions of correct (= true) application. In the simplest cases a term has a single received meaning and the relevant conditions of correct application are those that exhibit that meaning.

The key idea here is that a use of a term respects the relevant conditions of correct application only if its use on the occasion in question manifests an adequate grasp of those conditions. What I mean by 'grasp of the conditions' might more ordinarily be expressed by speaking of what a word is for. For instance, a somewhat partial grasp of the conditions of application of the term 'flu' might be expressed by saying that 'flu' is a word for a viral infection marked by fever and muscular aches.

Consider two contrasting cases of incorrect (= false) application. The first is a false application on the part of someone who knows perfectly well what the term 'dog' means, in the sense in which it picks out a species of domesticated animal, but in dim light applies it to a fox that he mistakes for a dog. Although this application is false it's in keeping with the relevant meaning of 'dog'. This person knows what the term 'dog' means, and accordingly his use manifests a grasp of the relevant conditions of correct application. The error lies simply in having mistaken a fox for a dog. The second case is an application on the part of someone who has not yet fully grasped what 'fox' means. A child might be disposed to apply the term to foxes and to dogs that look a little like foxes. Applying the term to a young Alsatian dog on some occasion the child speaks falsely but the error lies not just in the false application but in the fact that the false application derives from an inadequate grasp of the relevant conditions of correct (= true) application. The child uses the term as if it correctly applied not just to foxes but to foxes and some dogs and so fails to adequately to respect the relevant conditions of correct application. The mistake is accordingly semantic. The first subject's application of 'dog' to a fox is not.[5]

[5]Kathrin Glüer and Åsa Wikforss (2010*a*: 2.1.2) ask what motivates the introduction of the second notion of correctness. The answer to this is simply that to deny that there is this second type of correctness is to deny that examples of the sort considered point to a different dimension of evaluation from that marked by the first type of correctness. The worry might be whether the second type of correctness has anything to do with semantics. I find this hard to see since conditions of correct (= true) application surely belong to semantics and correct use in the second sense has to do with how speakers stand in relation to those conditions.

Isn't the first case one in which the term is used in a way in which it's not supposed to be used? Well, it is a misapplication—an application that is incorrect in the sense of false—but that is no reason to treat the application as incorrect in the second sense. A doctor does not fail to use the term 'flu' in keeping with its meaning if, misdiagnosing a patient, he says, 'This patent has flu' intending to say that the patient has flu. The doctor might or might not have been epistemically irresponsible in making his diagnosis but in any case his use manifests an adequate grasp of the relevant conditions of correct application and the term used is apt for saying what he intends to say. Similarly, if, lying, I say to someone, 'I have cleaned out the garage' I deliberately make a false application of the expression 'cleaned out the garage' but my use manifests an adequate understanding of the relevant conditions of correct application.

An application that is incorrect in the second sense might be correct in the first sense. If I were to apply the term 'arcane' to a ritual I might intend to convey, and mean to say, that it is ancient, not realizing that 'arcane' means *hidden or secret*. Yet the ritual might be arcane in which case my application, and what I say, would be true despite the fact that it does not manifest a grasp of the conditions of correct application for the term I use. In the envisaged circumstance I would say that the ritual is arcane and thus say something that is true, yet that the ritual is arcane in its received sense is not what I meant to convey though I uttered the words I did intentionally.

The child's use of 'fox' and my imagined use of 'arcane' fail to respect the relevant conditions of correct application because these uses do not manifest an adequate grasp of what those conditions are. These uses derive from ignorance or inadequate understandings of those conditions. It would be wrong, however, to suppose that all failures to respect the relevant conditions derive from ignorance or inadequate understanding. Slips of the tongue need not reflect ignorance or misunderstanding of the meaning of the term slipped in, yet the speaker's use fails to respect the conditions of correct application pertaining to the term used because it is not a manifestation of the speaker's grasp of those conditions. The speaker has a grasp of those conditions but his use does not stand in the right relation to that grasp.

One might be tempted to suppose that an application of a term is out of kilter with its meaning only if the subject says of the thing to which it is applied that it is one thing but intended to say that it is another.[6] Many cases of

[6]R. M. Hare (1963: 8) says that one misuses what he calls a descriptive term if one says that an object is of one kind, meaning or intending to convey that it is of another kind. He appears to

misuse, including slips of the tongue, are of this sort, yet it would be wrong to suppose that all are. When Tyler Burge's imagined patient (Burge 1979) utters the words, 'I have arthritis in my thigh' his use of 'arthritis' is incorrect in both senses. It is false because the pain has nothing to do with the patient's joints and arthritis is a condition of joints. He fails to respect the relevant conditions of correct application since his use does not manifest an adequate grasp of those conditions. Nonetheless, as the patient uses the term it stands for arthritis—the condition that doctors diagnose, that scientists research into, that in its various forms afflicts countless people. It seems right that he not only said that he had arthritis in his thigh but meant to say that he had arthritis in his thigh. Though his use of 'arthritis' was informed by a partial, albeit partially erroneous, conception of what it is for a person to have arthritis the term is apt for saying what he intended to say. By contrast, my use of 'arcane' was not informed by a conception of what it is to be arcane—I had no idea what it is to be arcane—which is why there was a complete mismatch between what I said and what I intended to say. A similar case would be a use of 'enervate' on the part of a subject who thought that 'enervate' means *energise or enliven*.

What about irony? Suppose that just after I have cleaned the kitchen floor a member of my family walks over it with muddy boots. I say, 'That was a great help' when what I mean to convey is that it was no help at all. Do I in this case respect the conditions of correct application of 'great help'? Again we need to focus on what it is to respect the conditions of correct application as that is to be understood here. In the case envisaged my ironical application of 'great help' is deliberately false, yet it manifests, and indeed is made possible by, an adequate grasp of the relevant conditions of correct application. So it satisfies our condition on respecting the relevant conditions.

The next task is to consider how the distinction between types of correctness feeds into an account of the way in which grasp of meaning is practical. After that I shall address the question of how meaning might be essentially normative.

4. The practical dimension of knowledge of meaning and the normativity of meaning

The practical dimension of grasping the meaning of a predicative term is knowing how to use it correctly, where the know-how is understood to im-

intend this to be a definition of 'misuse of a descriptive term'.

plicate an ability to use the term correctly in the second of the two senses. We are working with the idea that the measure of correct use of a predicative term in a given sense is *respect for the conditions of correct (= true) application that exhibit the term's meaning when used in that sense*. This measure is not just a standard imposed on us from the outside. If we grasp the meaning of a predicative term then in our uses of it we are sensitive to what the meaning requires of us. Moreover, in early learning we gain a sense of there being right and wrong ways of using terms through, among other things, having correct uses encouraged and misuses corrected. The conceptions we have of what terms ascribe can be refined or corrected. All this makes it natural to think that our uses of words are subject to rules in the sense of prescriptions or requirements that govern use. These are not merely norms or standards to which it is open to us to be indifferent. If we use words we incur a commitment to using them in keeping with the rules that govern their uses—commitments to which it is not open to us to be indifferent. If we are in breach of the rules then there is a sense in which *we* go wrong—we fail to discharge a commitment that *we* have incurred just by using those words. This is not to deny that we may play with words, exploiting them in ways that are not in keeping with their meanings. But such play depends for its effectiveness on there being rules that can be flouted.

Assuming there are such rules, what form do they have? Evidently there is more to using a term than applying it or denying it application. We use a term when we exploit its meaning in understanding another's use of it or when we make inferences from propositions that are articulated by the use of the term. A developed account of what is involved in correct use should accommodate the variety of ways in which predicative terms can be used. This diversity might induce despair about achieving a secure grip on what rules governing their use could look like. I think that our working idea about the measure of correct use suggests a way through the complexities.

'Goldfinch' in its most common sense correctly applies to a thing if and only if it is a goldfinch. That is not a rule, or at least not a prescriptive rule of the sort for which we are looking, but this is: when using 'goldfinch' in its most common sense respect those conditions of correct application. The activity of using 'goldfinch' in the sense exhibited by those conditions is, as I shall say, *a practice*, that is, an essentially rule-governed activity or cluster of such activities. It is essentially rule-governed in that no activity could be that activity unless it were governed by the rule prescribing respect for the

relevant conditions of correct application.[7]

The operative conception of a practice applies to many activities. An example is playing tennis. This activity is essentially rule-governed in that nothing would count as playing tennis unless it were governed by some set of rules for playing tennis. (Any differences in rules would induce differences in the activity governed even though variants would have much in common.) Obviously, those who engage in such activity are subject to the rules of the game. To be subject to rules is simply to be such that one's behaviour is liable to be evaluated in terms of accordance or lack of accordance with the rules. But people can be subject to rules because others in power subject them to those rules. Merely being subject to rules carries no obligation or commitment to obeying them. The relation between a player of tennis and the rules governing the game is more intimate than mere subjection. Should we say, then, that players are governed by the rules of the game in that they intend to conform to those rules?[8] Even if tennis players have such an intention when they play it would be absurdly naïve to suppose that necessarily players of rugby intend to conform to all of the rules of the game. While continuing to be players they might flout rules to gain advantage if they think they can do so with impunity.

There are, it seems, two dimensions to governance by a rule. One is normative; the other is psychological. I am working here with the idea that the normative dimension is best captured by the notion of a *commitment* that I briefly employed at the beginning of this section. Plausibly, players, in a game of rugby, just in virtue of being players, incur a *commitment* to following all of the rules. There is a very natural way to conceive of what this commitment amounts to. It amounts to it being the case that a player, just in virtue of being a player, ought to avoid continuing to play while not conforming to the rules of the game.[9] Why not say that the commitment amounts to it being the case

[7] I am applying here a conception of a practice that is most fully set out in Millar 2004.

[8] Glüer and Wikforss (2010*b*) point out that under an influential conception of what it is to follow a rule the answer is affirmative.

[9] In previous discussions (Millar 2004, 2011) I qualified statements to this effect so that the commitment amounts to it being the case that one ought to avoid continuing to participate while not following the rules *in the absence of some countervailing reason*. I envisaged that there might be reasons to remain within a practice, for instance, in a corrupt institution and subvert it from within. I am no longer sure that the qualification is necessary. (A suggestion to this effect was made to me at a LOGOS seminar in Barcelona in 2009, though I resisted it at the time.) A whistleblower who remains within an organization but flouts its rules to expose wrongdoing, might continue to be a member of the organization while having, in effect, abandoned some part of its practices. Playful uses of words might depart from the practices of their use while depending for their intelligibility on being departures from those very practices.

that the player, simply in virtue of being a player, ought to follow the rules of the game? The central point, I think, is that commitments are about the normative-practical implications of occupying a certain standing, not about what course to take. Believing certain things commits one to believing other things. Merely incurring such a commitment does not tell us whether to believe something we are currently committed to believing. The best response might be to give up some belief among those that incur the commitment. Intending to do something commits one to taking the means necessary to doing that thing. Merely incurring such a commitment does not tell us whether to take the means. The best response might be to abandon the intention. The general point is that incurring a commitment leaves open whether we should do that to which we are committed rather than alter the condition that incurs the commitment. This applies to the commitments incurred by participating in a practice. While participating in a practice incurs a commitment to following the rules governing the practice, the mere fact that this commitment has been incurred does not dictate that the rules should be followed. It might be that we ought to stop participating in the practice as in the case of practices of using terms of racial abuse. It makes sense that this should be so. Practices do not exist in isolation. If we are participating in a practice and the question arises whether to continue participating and, by implication continue to follow the rules, the mere fact that we are participants will yield no answer. There can be pressing reasons having to do with the impact of the practice to withdraw from it. Yet these reasons do not impugn the idea that being a participant commits one to following its rules.

Using the term 'goldfinch' in its usual sense counts as a practice because it is an activity that is essentially rule-governed in that it would not be the activity that it is but for its being governed by the rule prescribing respect for the condition of correct application that display the relevant meaning. It is in keeping with the proposed account of how participants relate to the rules of a practice that those who use 'goldfinch' in this way incur a commitment to following this rule. This commitment amounts to it being the case that one ought to avoid continuing to participate in the practice while not following this rule. It can be discharged in one of two ways—by withdrawing from the practice or by following the rule. One could withdraw by giving up the use the term or more radically by ceasing to use English.

The psychological dimension of governance by rules depends on whether the rules are formulated and explicitly treated as rules. The rules of soccer can be written down and cited to guide behaviour. It is correspondingly easy to say what would count as being governed by such rules. This would be a mat-

ter of knowing what the rules are and submitting oneself to them. (This would include preparedness to take the consequences if one flouts them.) Rules for word-use of the sort that I have posited are not generally written down. They are not rules that guide in the way that the rules of rugby guide because speakers do not routinely have them in mind. This might lead one to be sceptical that speakers are governed by rules like these. Yet there does seem be a sense in which we can follow rules, and accordingly be governed by them, even if we never have them in mind in full generality. It might be that we routinely follow the rules in question in this sense.

For the sake of argument suppose that there is a rule in the style of Grice's Cooperative Principle (Grice 1975) prescribing that we make our contributions to a conversation relevant given the accepted purpose or direction of the conversation. This rule could be *implicitly* followed even by those who never articulate it. An important part of what that would amount to is that their contributions are in general relevant to the purpose and direction of conversations in which one takes part, but that by itself would amount to their according to the rule but not to their following it. It would be crucial that they also have an ability to recognize concerning uncooperative contributions that they are *to be avoided* because they lack relevance. They need not have a general conception of what conversational cooperation amounts to. What matters is that their own contributions are modulated by their ability to recognize of irrelevant contributions that they are or would be inappropriate because irrelevant. They would have to be sensitive to irrelevance not only in that they actually avoid it (by-and-large), but also in that they have some understanding of irrelevance as to be avoided. In virtue of such understanding they would not merely accord with the rule; they would implicitly follow it. The question arises whether a similar story is plausible for the sorts of rules governing the use of predicative terms that I have envisaged.

We fail to respect the conditions of correct application of a term when our use of it does not manifest a grasp of those conditions. If the general idea in play in the discussion of the Cooperative Principle were to apply straightforwardly to the kind of rules governing the use of terms that we are considering, we would need to make sense of how those implicitly following a rule for a term can tell of uses that fail to respect the relevant conditions of correct application that they fail to respect those conditions and are of a sort to be avoided on that account. This might seem to ask for quite a lot if only because those with a grasp of the meaning of the term need not grasp in so many words what it is to respect the relevant conditions of correct application. But if they are competent users of the term they will be in command of something that is

tantamount to this. For instance, if the term is 'goldfinch' the use they make of it in their own utterances, and their reactions to uses made by others in their utterances, will be guided by a conception of the sort of thing to which it correctly applies. Should the issue arise they will think of the term as correctly applying to things of that sort. They will be able to recognize uses that clearly manifest a misconception as inappropriate because indicative of misunderstanding, and will regard such uses as inappropriate. That they implicitly follow the rule will be manifested in such ways. Of course, initiates into a practice for using words count as participants even at a stage in early learning at which they have no thoughts about words and their use, and even if they never get beyond that stage. The show keeps on the road in part because a sufficient number of participants are reflective to some degree about their use of language.

The proposal, then, is that the practical dimension of grasp of the meaning of a term amounts to knowing how to use it correctly in a sense that implicates an ability implicitly to follow the relevant rule.[10] This account accommodates the plausible thought that knowing how to use a word is not simply a matter of having various dispositions, conceived in the standard philosophical way, but implicates a sense of there being right and wrong ways to use it and an ability to tell which is which.

5. Resistance to normativity

Meaning is essentially normative if there is something about using an expression meaningfully that in and of itself makes it the case that those so using it ought to do something. On the account sketched in the previous section the normative dimension of the meaning of predicative terms is captured by the claim that just in virtue of using a term in a particular sense one incurs a com-

[10]Jennifer Hornsby, defending the view that semantic knowledge is practical, suggests that 'someone whose knowledge how to ϕ is practical is able to simply ϕ (at least so long as it is actually possible for her to ϕ' (Hornsby 2005: 115). To be able simply to ϕ is to be able to ϕ but not through doing something else. In response Jason Stanley remarks that '[i]n the case of individual words (and modes of syntactic combination), there is no … ability to do something, no ability simply to *F*' (Stanley 2005: 138). The account I am proposing suggests that Stanley is unduly pessimistic with respect individual predicative terms. When we are able to use such terms correctly, in the sense of using them in keeping with the relevant conditions of correct application, and we exercise that ability in uses that are correct in that sense, we do not do so by doing something else. And once we bear in mind that use covers so much more than application we can make sense of how the ability to use a word correctly can implicate an ability to use that word in combination with others in ways that make sense both syntactically and semantically.

mitment to following a rule for its use prescribing respect for the conditions of correct application that display the relevant meaning. The commitment arises from being a participant in a practice of using of the term in the relevant sense. To have such a commitment is a matter of it being the case that one ought to avoid continuing in the practice while flouting this rule.

Some might suggest that the commitments incurred by using terms can be explained without assuming that meaning is essentially normative. The thought might be that we have a reason to use a term in a manner that respects its conditions of correct application because otherwise we run the risk of failing to communicate. Such normativity as there is in this area is taken to relate to instrumental rationality rather than anything essential to meaning. On this view there is no need to posit the kind of practice that I am linking to the correct use of terms and to the normativity of meaning. It is true that native speakers with command of the term 'goldfinch', and likely to want to talk about goldfinches or understand the talk of other English speakers about goldfinches, have a very good reason to continue being participants in the practice, since being a participant is the means to achieve those ends. The question though is whether uses that are incorrect in the second sense can be explained as failures of instrumental rationality. I think not since a misuse would be no less a misuse if the speaker were to have decisive practical reasons to use a term in a manner that fails to respect its conditions of correct application. It is crucial that we do not conflate considerations pertaining to why one should participate in a practice with what is incumbent upon one if one is a participant. From the present perspective there is something one has reason to do just in virtue of being a participant in a practice of using a term, irrespective of any reasons there might be to participate in the practice: one has reason either to respect the conditions of correct application of the term or withdraw from the practice.

Discussion of the normativity of meaning has been seriously distorted by *the problematic assumption* that if meaning were essentially normative then its normativity would be captured by such claims as that 'red' ought to be applied to a thing only if it is red.[11] There really is no good reason to accept this assumption. Those who, by way of telling a lie, say of something that is not red that it is red might have acted wrongly but there is no reason to think that they have made some linguistic error. I take it to be a strength of the pre-

[11] This goes back to Kripke 1982. See also Gibbard 1994 and, recently, Ginsborg 2012. Some who object to the essential normativity of meaning are also guided by this assumption. See, for instance, Horwich 1998 and Hattiangadi 2007.

ceding discussion that it avoids this assumption and thus avoids objections to the view that meaning is essentially normative that rest on a conception of normativity that incorporates it. The key to felicity in this area is having due regard to the two types of correctness that I have distinguished. Where the focus is on correctness of the second type, we can happily accommodate the fact that there need be nothing linguistically incorrect about a false application of a predicative term. From such a perspective there is no incentive to link normativity to the kind of ought-statement that figures in the problematic assumption.

6. *A problem posed by occasion-sensitivity*

The position I have described is theoretically satisfying in that it connects two notions that are sometimes thought to have their natural homes in quite different theoretical frameworks. These are conditions of correct application and rules for use. But it faces what threatens to be a significant challenge. I shall describe the challenge and suggest a way to meet it that merits attention.

Jonathan Dancy objects to theories of meaning that invoke rules for use drawing upon the following conception of meaning.[12]

> ... the meaning of [a] term is what one knows when one is a competent user of that term. If the term is capable of making a range of contributions to differing contexts, this is part of what the competent user must know. To be a competent user, then, is to be in command of the *sorts of* difference that the presence of the term can make to the semantic value of the contexts in which it can appropriately be found. ... The meaning of [a] term, understood in general, is the range of differences it can make; its meaning in a given context is to be found somewhere in that range (though of course some contexts force an extension or other adaptation of that range). (Dancy 2004: 194)

In the light of this conception Dancy asserts,

> There is nothing here that could be captured in a rule. Rules, in the sense in which we are here concerned, must be articulable in principle, even if our competent speaker is incapable of articulating them in practice. But if the meaning of the term consists in

[12] I have found Whiting 2010 helpful in relation to what follows.

> an open-ended *range* of available *sorts of* semantic contribution in this way, it is essentially inarticulable. Competence with it will therefore have to consist in a kind of skill rather than a grasp of a specifiable rule … . (Dancy 2004: 196)

I am happy with the idea that linguistic competence is a kind of skill. The question is whether the view Dancy outlines poses a problem for the account of grasp of meaning that I have given here.

First we need to consider why one might think that terms have the potential to make 'an open-ended *range* of available *sorts of* semantic contribution'. Examples of a sort used by Charles Travis (for instance, in Travis 1989, 1994, 1997, 2000) are suggestive in this respect. Here is one.

> Pia's Japanese maple is full of russet leaves. Believing that green is the colour of leaves she paints them. Returning she reports, 'That's better. The leaves are green now.' She speaks the truth. A botanist friend then phones, seeking green leaves for a study of green-leaf chemistry. 'The leaves (on my tree) are green,' Pia says. 'You can have those.' But now Pia speaks falsehood. (Travis 1997: 89)

What is being suggested is not that 'green' is ambiguous in the way that 'bank' is or polysemous in the way that 'stand' is. It is that even when understood as having a particular meaning it does not make the same contribution to what is said on each occasion on which it is used.

> … the words 'is green', while speaking of being green, may make any of many semantic contributions to wholes of which they are a part, different contributions yielding different results as to what would count as being as they are said to be. (Travis 1997: 92)

One might suppose that the phenomenon is akin to the occasion-sensitivity of an indexical like 'now' which is associated with a function from times of speaking to times referred to by its use. Pursuing this line the idea, as Travis puts it, would be that

> what 'is green' means determines a set of parameters (variables in speakings), and a function from values of them onto a range of contributions 'is green' might make, such that for any argument of the function (fixed relevant values of the speaking), the value of the function is the contribution which 'is green' would make on a speaking so characterized. (1994: 174)

Travis rejects any such view and there is indeed reason to doubt that the phenomenon yields to a functional treatment because it is implausible that anyone who fully grasps the meaning of 'green' must be aware of, or even sensitive to, some fixed set of parameters of speakings that select what it contributes to what is said by its use on any occasion. A pressing question for the present discussion is whether the examples pose a problem for the claim that predicative terms are governed by rules prescribing respect for conditions of correct application. On the face of it they do if understood in Travis's way for they raise a question as to whether it can be right to suppose that there are meaning-exhibiting conditions of correct application as I have portrayed them. Travis himself rejects the view that 'green' correctly applies to a thing if and only if it is green. (Travis 2000: 213, using the case of 'blue').

Let's grant this much.

Underdetermination (in relation to the example given of uses of 'green')

> (a) What is said by 'The leaves are green' on the two occasions of use in the example given is different, and is, therefore, not wholly fixed by the words used, yet
>
> (b) this is not because the words used are ambiguous or otherwise polysemous, and
>
> (c) the difference cannot to be explained on the model of standard treatments of indexicals.

What then can account for this underdetermination? We are liable to be pulled in two different directions here.

We might think there is a sense in which that of which Pia speaks in her first utterance is different from that of which she speaks in her second utterance. Though in both cases she speaks of the leaves her first utterance speaks of them truly with respect to the colour they have after painting and her second utterance speaks of them falsely with respect to their current natural (unadulterated) colour. So one direction in which we might be pulled is to accepting (a) that what is said of them with respect to their colour after painting is exactly what is said of them with respect to their current natural colour and so (b) it is wrong to assume that what 'green' contributes to what is said by either utterance is different. (One is tempted to say here, 'If the same colour is ascribed then surely what is said on both occasions is the same—that something has that colour.) But there is a pull from a different direction. Respects in which things can be green are just ways of being green. So Pia's first

application attributes one way of being green to her leaves and her second falsely attributes another way of being green to her leaves. From this perspective, what is attributed is different. Correspondingly, what 'green' contributes to what is said by Pia's first utterance is such that the utterance ascribes one way of being green while what it contributes to what is said by her second utterance is such that this utterance ascribes a different way of being green. The contributions are different as Travis supposes. Other examples serve to suggest that this is the right way to go. A surface might be green because made of green plastic. It would look green if seen in daylight. A different surface might be green because bathed in green light. If one said of the latter that it was green in a context in which the colour of the material of which it was made was at issue one would speak falsely. So the contribution 'green' makes to an application of it to a surface can be such that what it ascribes to the surface is being green in the first respect, and it can be such that what the application ascribes to the surface is being green in the second respect. On this way of thinking what becomes of the tempting thought that since the same colour is ascribed what 'green' contributes to what is said is the same in the two cases? That thought seems compelling because the same concept is in play and accordingly there is a sense in which the meaning of the word is constant across the applications. But as Travis remarks, 'a concept by itself does not determine which ways for things to be, so which things, satisfy the concept' (1994: 181)

If this view is right how does the context of utterance contribute to fixing what is said? We are to reject the idea that for the case of 'green' there is some fixed set of parameters associated with the relevant utterances and a function from those parameters to which colour-respect is at issue. Even so, observing Pia finish her leaf painting, what she says in speaking as she does would be clear. Davidsonian considerations about interpretation kick in at this point.[13] To the extent that a person's utterance about present circumstances is intelligible it must make sense as saying something that the speaker could (perhaps ineptly) treat as being pertinent to, and reasonable in, those circumstances. What Pia says first reflects what it would make sense for her to say with the words she uses given her strange preoccupations. That it is the colour she has painted on the leaves of which she speaks would be clear *to us* because, as the situation has been described, only that understanding would make sense of her speaking as she does. Likewise in speaking to her friend the context makes it clear that she can be speaking only of the natural colour the leaves

[13] For an outline of what I take the central considerations to be see Millar 2004: ch. 1.

have. (Presumably she is being mischievous.)

Another example used by Travis (1989: 18-19) yields to a similar treatment. In one context, saying 'There is milk in the fridge' might pertain to the availability of milk for some mode of consumption—drinking, adding to tea, making a cake mixture, and the like. In another it might pertain to a puddle of spilt milk that is still there despite an attempt to clean the fridge. On Travis's view the contribution of 'milk' to what is said is different in the two cases. One might find this hard to credit because both utterances attribute the presence of milk to the contents of the fridge, but there are different ways in which milk can be present—as a stain on a garment or on a floor, in a carton for storing milk for consumption, or as a drip on a mother's breast. An utterance applying 'milk' might attribute its presence in any of various ways and so what 'milk' contributes to what is said will vary accordingly.

Travis (1997: 91) is, I think, right to dismiss the suggestion that the phenomenon in question should be considered to be a case of ellipsis if that is taken to mean that the words used are shorthand for a longer sentence that does not admit of diverse possible understandings and can therefore serve to fix what is said. The problem with this is that if underdetermination of the sort under consideration is pervasive then those further words could bear different understandings in different contexts. It's true, and instructive, that if anyone were wondering what Pia said by her words on either occasion, further words could make this clear. Pia's first utterance could be clarified by saying, 'With respect of the colour they have been painted the leaves are green'. Her second utterance could be clarified by saying, 'With respect to their natural colour the leaves are green'. This is so even if the clarifying words themselves admit of different understandings in that there are possible contexts, other than those in which the clarifying words are actually used, in which what they would say would be different.

The upshot is that a premise concerning the meaning of terms in Dancy's case against the invocation of rules for the use of terms looks to be correct. Supposing that it is, does Dancy have a good case against the account of rules that I gave—the account on which rules for use of predicative terms prescribe respect for the relevant conditions of correct (=true) application? I suggest that the account may stand if we refine our conception of meaning-exhibiting conditions of correct application to accommodate occasion-sensitivity. For a term like 'green' the form of such conditions must be something like this: in the sense in which it stands for a colour, 'green' correctly applies to a thing on an occasion of use if and only if there is a certain way for things to be coloured, the occasion is such that what is at issue is that way for things to be coloured,

and the thing is green in that way. The term 'milk' may be used to ascribe the presence of milk in a variety of forms and the words used on an occasion of use need not by themselves determine which form is at issue. Accordingly, 'milk' applies to a substance on an occasion if and only if there is a certain form in which milk can be present, the occasion is such that what is at issue is whether milk is present in that form, and the substance is milk present in that form. The form at issue might be evident because interlocutors are both attending to a stain on the floor or a puddle in a fridge or it might be evident because of conversation raising a question as to the availability of milk for consumption.

I emphasised in section 1 that those who have some grasp of the meaning of a predicative term might have diverse conceptions of what it is to fall under it. Clearly it can be part of one's conception of what it is to be green that there can be different respects in which something can be green and it can be part of one's conception of milk that milk can be present in a variety of forms. One's ability to employ the concept of milk can be refined though a developing conception of the varieties of forms in which milk can be present. Similarly one's ability to employ the concept of being green can be refined though a developing conception of the variety of respects in which something can be green. A corollary is that the degree to which one is able generally to respect the conditions of correct application of a term will vary with the level of refinement of one's conception of the range of things to which it can be applied.

The account I am offering is in keeping with something that concerns Dancy in the passages I quoted: there is much that goes into an ability to employ a term correctly, in the sense of being in keeping with its meaning, that is not brought out by formulations of rules. This is true on my account since the rules that I envisage to do not specify what it is to respect the relevant conditions of correct application. But this is a virtue of the account, not an objection to it. Since respecting the conditions of correct application is a practical ability we should not expect any rules fully to articulate what it is to have that ability. My suggestion, then, is that Dancy's remarks, for all the insight they undoubtedly contain, do not tell against the conception I have been outlining. [14]

[14]I would like to record my appreciation of the massive contribution that Pascal Engel has made to the international dissemination of clear, constructive philosophy and to wish him well on the occasion of his sixtieth birthday.

7. References

Burge, T. (1979). 'Individualism and the Mental', *Midwest Studies in Philosophy 4*, 73-121.

Dummett, M. A. E. (1973). *Frege: Philosophy of Language* (London: Duckworth).

Dancy, J. (2004). *Ethics Without Principles* (Oxford: Clarendon Press).

Gibbard, A. (1994). 'Meaning and Normativity' in E. Villanueva (ed.) *Truth and Rationality* (Atascadero, Calif.: Ridgeview).

Ginsborg. H. (2012). 'Meaning, Understanding and Normativity', *The Aristotelian Society: Supplementary Volume 86*, 128, 127-46.

Glüer, K. and Wikforss, A. (2010*a*). 'The Normativity of Meaning and Content', *The Stanford Encyclopedia (Winter 2010 Edition)*, Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2010/entries/meaning-normativity/.

Glüer, K. and Wikforss, A. (2010*b*). 'Es braucht die Regel nicht: Wittgenstein on Rules and Meaning', in Whiting (ed.), 148-66.

Grice, H. P. (1975). 'Logic and Conversation' in P. Cole and J. L. Morgan (eds.) *Syntax and Semantics, Vol. 3*, (New York: Academic Press), 41-58.

Hare, R. M. (1963). *Freedom and Reason* (Oxford: Clarendon Press).

Hattiangadi, A. (2007). *Oughts and Thoughts: Rule Following and the Normativity of Content* (Oxford: Clarendon Press).

Hornsby, J. (2005). 'Semantic Knowledge and Practical Knowledge', *The Aristotelian Society*, *Supplementary Volume 79*, 107-130.

Horwich, P. (1998) *Meaning* (Oxford: Clarendon Press).

Kripke, S. A. (1982). *Wittgenstein on Rules and Private Language* (Oxford: Blackwell).

McDowell, J. (1977). 'The Sense and Reference of a Proper Name', *Mind*, 86, 159-85.

McDowell, J. (1984). 'Wittgenstein on Following a Rule', *Synthese,* 58, 352-63.

McGinn, C. (1984). *Wittgenstein on Meaning* (Oxford: Blackwell).

Millar. A. (2002). 'The Normativity of Meaning' in A. O.Hear (ed.) *Logic, Thought and Language* (Cambridge: Cambridge University Press), 57-73.

Millar, A. (2002). *Understanding People: Normativity and Rationalizing Explanation* (Oxford: Clarendon Press).

Millar, A. (2011). 'The Epistemological Significance of Practices', *ProtoSociology: An International Journal and Interdisciplinary Project*, 213-30.

Stanley, J. (2005). 'Hornsby on the Phenomenology of Speech', *The Aristotelian Society*, *Supplementary Volume 79*, 131-45.

Travis, C. (1989). *The Uses of Sense*. (Oxford: Clarendon Press).

Travis, C. (1994). 'On Constraints on Generality', *Proceedings of the Aristotelian Society* 94, 165-88.

Travis, C. (1997). 'Pragmatics' in B. Hale and C. Wright (eds.) *A Companion to the Philosophy of Language* (Oxford, Blackwell), 1997, 87-107.

Travis, C. (2000). *Unshadowed Thought* (Cambridge, MA.: Harvard University Press).

Whiting, D. (2010). 'Particular and General: Wittgenstein, Linguistic Rules, and Context' in Whiting (ed.) 2010, 114-132.

Whiting, D. (ed.) (2010). *The Later Wittgenstein on Language* (Basingstoke, Hampshire: Palgrave Macmillan).

45

Engel on Doxastic Correctness

CONOR MCHUGH

It is a great privilege to contribute to this festschrift for Pascal Engel, and thereby pay a small tribute to his important, wide-ranging contributions to Philosophy in recent decades. I also offer this short piece as a token of gratitude. As well as lively, stimulating, and, on Pascal's side, erudite philosophical discussion, Pascal has also offered me much generous personal and professional support over the years–not least during my time as a postdoc in Paris. I even had the good fortune to spend a year living in Pascal's former office in the 14ème! Whether the atmosphere rubbed off on me sufficiently, I leave for Pascal to judge.

In addition, I think that all philosophers in the so-called analytic tradition, and all who endorse the values of clarity, humility and open-mindedness that, at its best, this tradition represents, owe Pascal thanks for his tireless efforts in its defence.

I have talked a lot with Pascal on the subject of belief and its norms, so that is the topic that I will focus on.

1. Doxastic Correctness

It is a platitude that beliefs can be correct or incorrect. It is almost as much of a platitude, and one I won't question here, that the standard of correctness for belief, or at least the primary one, is truth.[1] But what *is* correctness for belief? It would be a mistake to think that this question is answered simply by citing the standard of correctness for belief. This standard tells you what property a given belief must have in order to have the further property of being correct. It doesn't tell you the nature of the further property of correctness that the belief thereby has.

Pascal Engel, like many others, holds that doxastic correctness is normative.[2] On this view, to say that true beliefs are correct is to say that true beliefs satisfy some norm; to say that false beliefs are incorrect is to say that they violate this norm. It's not hard to see why this is a natural view. When we say that someone has an incorrect belief, we do not seem to be merely describing some feature of her, or of her belief–a feature whose presence she could sensibly remain unconcerned by. We are saying that she believes *wrongly*. This looks normative.

It is also natural to think that this normative standard of doxastic correctness not only applies necessarily, but is constitutive of the very attitude of belief. That is, it is part of what it *is* for a given attitude to count as a belief that it be subject to this standard.

I am sympathetic both to the claim that doxastic correctness is normative, and to the claim that being subject to this standard is constitutive of belief. Here I will focus primarily on the first claim. I do not want to question the truth of the claim, but rather to ask about the normative property that is involved here. What kind of property is it? Is it, for example, a deontic property? Or an evaluative one? Or something else?

These are questions with respect to which Pascal Engel has done much to advance our understanding. Here I will focus on the view that he expounds in his recent article, 'Doxastic Correctness'.[3]

[1] But see Smithies (2012).
[2] Engel (2013).
[3] Ibid..

2. E-Correctness and I-Correctness

Following Thomson,[4] Engel distinguishes between 'internal' correctness and 'external' correctness, or *i*-correctness and *e*-correctness. Consider the act of asserting some proposition. This act can be performed correctly in the sense that one utters a grammatical sentence that expresses that proposition, one pronounces all the words in the right way, one speaks sufficiently loudly, and so on. In that case, one's act of assertion is *i*-correct. We can also say that one has made a correct assertion in the sense that what one has asserted is true. In that case, one's assertion is *e*-correct. Clearly, these two kinds of correctness can come apart. One can perform in an exemplary way the act of asserting a proposition that is in fact false, and one can do a very bad job of asserting a proposition that is true.

The truth-standard of doxastic correctness seems to correspond to *e*-correctness. But this, for Engel, raises a worry about whether doxastic correctness is really normative. He writes:

> is it clear that [*e*-]correctness is a normative property? The standard for a tune is fixed by a set of notes, the standard for a map is fixed by the similarity between the map and the territory represented, the correct spelling is fixed by a certain pronunciation of the word. These are descriptive properties, not normative ones. ... The normative concept of correctness is distinct from this descriptive one. It concerns the way, or the operation which, an agent has to perform in order to meet the descriptive condition (Engel 2013, 200).

The worry here seems to be that *e*-correctness for various kinds consists simply in the possession of a certain descriptive property, such as containing certain notes in a certain order. Note that this seems to be true in particular for the kind *belief*, since truth appears to be a purely descriptive property.[5] On the other hand, *i*-correctness seems to be a matter of doing what one does more or less well–something that *does* look normative.

The moral of the story, Engel concludes, is that

[4] Thomson (2008).

[5] More precisely, the relevant property for belief is that of having a true propositional content. Some argue that content is normative. This would complicate things in ways that I leave aside here.

> for a kind *K* to be correct it has to meet *both* the *e*- and the *i*-correctness conditions. It would be wrong to reduce correctness to either one of these two dimensions (ibid., 201).

The suggestion, then, seems to be that correctness, and doxastic correctness in particular, comprises both *e*- and *i*-correctness, and that it is only because it includes *i*-correctness that it is normative.

It's worth noting that this isn't exactly Thomson's view of things. She does not talk of some overarching correctness property consisting of the conjunction of *e*-correctness and *i*-correctness. Rather, she simply holds that *e*-correctness and *i*-correctness are distinct. Of course, one could use 'correctness' in a stipulative way to refer to the property that consists in being both *e*-correct and *i*-correct. But the question would remain whether this property has any theoretical interest.

On the face of it, the notion of *i*-correctness does not obviously apply to belief at all. Pauline can assert that *p* better than Pierre asserts it - because of her superior grammar or diction, say. But, as Thomson points out, it would be odd to talk of one person believing that *p* better than another person. Pauline might believe that *p* more strongly than Pierre does, or with more certainty than him. She might understand better what is involved in its being the case that *p*. But none of this seems to amount to believing *better*; believing as such doesn't seem to be the sort of thing one can do a better or worse job of. After all, as Engel says, believing is "not a performance" (ibid.). This suggests that, in so far as the distinction between *e*-correctness and *i*-correctness is in good standing, doxastic correctness is just a species of *e*-correctness.

In response to this sort of worry, Engel suggests that we understand *i*-correctness for belief as follows:

> "Believing for bad reasons, or on the basis of insufficient evidence, is poor believing and thus *i*-incorrect" (ibid.).

Thus, doxastic *i*-correctness is a matter of basing one's beliefs on appropriate grounds. This is an ingenious idea, taking advantage of the point that, while believing *per se* is not something one can do more or less well, *basing* one's beliefs arguably is.

Nonetheless, I am sceptical that this is really a kind of doxastic correctness. Or, at any rate, I am sceptical that this property, together with doxastic *e*-correctness, forms part of an interesting, overarching correctness property possessed by beliefs. This is for several reasons.

First, appropriate basing is, as Engel acknowledges, not so much a property of the belief, as of some broader performance that culminates in one's holding the belief. Second, it seems odd to think of it as a form of correctness, as opposed to a form of rationality, reasonability, justification, or something of that sort. Third, it seems to me that what *counts* as appropriate doxastic basing *depends* in part on the standard of correctness for belief: it is appropriate to base one's beliefs on evidence precisely because evidence is connected to truth, and true beliefs are (*e*-)correct. If that's right, then the truth-standard of doxastic correctness is more fundamental than the standard for appropriate basing of beliefs. It would be odd, then, to think that the truth-standard, and the standard for appropriate basing, somehow come together to form an overarching property of correctness.

I am suggesting that the notion of *i*-correctness does not apply to belief. Are we then forced to concede that doxastic correctness is not normative after all? As we saw, being *e*-correct seems to be simply a matter of possessing some descriptive property - whatever property is fixed as the standard of *e*-correctness for the relevant kind.

I don't think we should concede so easily that *e*-correctness is not normative. Recall the distinction between the property in virtue of which something counts as correct, and the property of correctness itself (having argued that the notion of *i*-correctness has no role in this context, henceforth I will drop the '*e*' prefix). It may be that the former property is non-normative, but the latter property is normative. That is, it may be that the standard of correctness for a certain kind can be specified in wholly non-normative terms–for example, the standard might be having a true propositional content–but satisfying that standard gives a thing of that kind the normative property of being correct. This point can be obscured by formulations to the effect that correctness for belief *is*, or *consists in*, truth.

In general, it is not obviously problematic to say that something's having a non-normative property gives it a distinct normative property. In ethics, utilitarians may say that what makes an action right is that it maximises pleasure. This is a non-normative property. Utilitarians need not claim that rightness itself is therefore non-normative. They can say that pleasure-maximisation *makes* actions right, but deny that the property of rightness is *identical* to the non-normative property of pleasure-maximisation. Defenders of other ethical theories, like Kantianism and contractualism, can make similar moves. Perhaps none of these views are sustainable, but they are not obviously confused from the start.

It might seem mysterious that the seemingly non-normative property of

truth could give beliefs some further, normative property. How could it perform this alchemy? The answer, I think, is that it is not just the property of truth that does the work. The nature of belief does some of it too. The attitude of belief is, in a certain way, made for truth. Here, the claim that the standard of correctness for belief is constitutive of belief becomes important.

Correctness, Ideals of Reason, and Fittingness

Among those who hold that doxastic correctness is normative, the most common view is that this norm is deontic or prescriptive in character.[6] That is, it says that we ought not believe what's false, and that we ought (or may) believe what's true. Here, 'ought' and 'may' are understood in their normal senses, as enjoining, prescribing, permitting or forbidding certain pieces conduct. In this case they would apply to our doxastic conduct.

There are well known objections to this sort of view, to the effect that no such prescriptive norm could plausibly be what guides, or what ought to guide, our doxastic conduct.[7] Engel agrees that doxastic correctness is not prescriptive, and proposes an alternative view. The truth norm, he writes

> "does not give us any prescriptive—or even permissive—guidance... It is an *ideal of reason*, in the sense that it tells you what you ought to ideally believe, namely the truth, and thus it belongs to the category of the *ought-to-be* rather than to the category of the *ought-to-do*" (ibid., 208).

This is an intriguing suggestion for retaining a normative view of doxastic correctness, without getting into the difficulties associated with the deontic version of this view–a project with which I am very sympathetic. For my part, however, I find the suggestion somewhat elusive. I will set out a few initial worries about it, before raising some questions about how Engel's proposal relates to other views that take doxastic correctness to be normative in a broad sense, even if not deontic or prescriptive.

If believing the truth is an ideal of reason, then failing to do so would presumably constitute a failure of reason. But false beliefs do not seem always to constitute such a failure. In a situation of misleading evidence, believing what is in fact false might be perfectly reasonable and rational. It seems that,

[6] E.g. Shah (2003).

[7] Bykvist and Hattiangadi (2007); Glüer and Wikforss (2009); McHugh (2012).

far from true belief being an ideal of reason, what reason (or perhaps Reason with a capital R) tells you to do is to believe according to your evidence. What's more, Reason does this precisely *because* true belief is correct.

I am also unsure about the suggestion that believing the truth is an ideal, if this means *merely* an ideal. The term 'ideal' has a whiff of the supererogatory. Ideals seem to be things that we should aspire to, but failure to satisfy which is not normally grounds for serious criticism or blame. While falsity in a belief can be blameless and beyond criticism, it nonetheless seems like the sort of thing we should be worried about, and rectify when we become aware of it. To believe falsely is not merely to manifest one's inevitable human fallibility and imperfection. It is to go wrong. This seems to me a rather more urgent problem.

Ideals also seem typically to be things that we can be more or less close to satisfying. While omniscience is like that, doxastic correctness as such doesn't seem to be.[8]

For these reasons, I'm not sure that I have a grip on how doxastic correctness could be fundamentally either an ideal or a standard of reason.

One might also worry here about Engel's abandonment of the idea that the truth norm *guides* believers. When we reason theoretically, for example, we certainly seem to be guided by truth in some way; *modulo* self-deception and other such funny business, we typically form through reasoning only beliefs that we take our evidence to support. This fact requires explanation. It would be surprising if the explanation had nothing to do with the fact that true belief is correct belief, and that believers are in some way sensitive to this. Indeed, facts about the kinds of considerations to which believers are sensitive in their theoretical reasoning are, for some theorists, the primary motivation for positing a normative standard of correctness for belief in the first place.[9]

This is not to say that we should return to the idea that the truth norm guides by prescribing true beliefs, and forbidding false ones. Our notion of guidance, and of normativity, need not be so narrow. Indeed, I think that Engel's work has done a great deal to illuminate this point, and to show that normativists' options are not nearly as restricted as some of their critics suppose. In closing, I want to turn briefly to some of these other options.

The normative, in the broad sense in which it is opposed to the descriptive, includes evaluative properties as well as deontic ones. To say that something

[8] Engel rejects Wedgwood's claim that doxastic correctness comes in degrees (Wedgwood 2013). I agree with Engel on this point.

[9] Shah (2003).

is good, or that it is good of its kind, is a normative claim in this sense. One might claim, then, that doxastic correctness is an evaluative property.

It is natural to read Engel's proposal in this way. Ideals seem to be standards for evaluation. Failure to satisfy an ideal is not necessarily a big problem, but the closer one gets to satisfying it, the better. Engel also talks of 'ought-to-be'. This looks like an evaluative notion. For example, the claim that the world ought to contain less suffering seems to be an evaluative claim, to the effect that it would be better if the world contained less suffering.

I think that the evaluative view of doxastic correctness is attractive. Some philosophers have expressed doubt or puzzlement about the claim that true belief is good *simpliciter*. But the value of true beliefs can be understood as a form of attributive value–roughly, beliefs are good *qua* beliefs when true.[10] Believers may plausibly be guided by such a value. For example, the value of true belief may make it the case that evidence for a proposition constitutes a reason to believe that proposition–which reason can then guide thinkers in their theoretical deliberation. At the same time, this value would not generate implausibly strong or impossible-to-follow prescriptions.

Nonetheless, the evaluative view faces some objections. To my mind, the most compelling one is the rather simple thought that correctness and attributive goodness are just very different sorts of properties. Correctness seems to be a matter of there being a certain sort of 'fit' or 'match' between two things. Attributive goodness is different. For something to be a good knife, or a good heart, is not for it to fit some other thing, but rather for it to have properties that enable it to perform its function appropriately in normal circumstances. We are not at all tempted to say that good knives and good hearts are thereby *correct* knives or *correct* hearts. This suggests that attributive goodness and correctness are different properties.

It might be said that this is compatible with the evaluative view. For the evaluativist may claim that being correct *makes* a belief good *qua* belief, rather than that doxastic correctness *is* an evaluative property. For present purposes let me just note that this is not the sort of view that I am trying to explore, nor, as I understand it, the sort of view that Engel means to endorse (although I am open to correction here). On this view, correctness is no more normative than truth; it is goodness that does the real normative work. Indeed, this view raises the question why we bother talking about correctness at all. Why not just say that, when beliefs are true, this gives them the further property of being good beliefs?

[10] I defend such a view in McHugh (2012).

If doxastic correctness is neither a deontic property nor an evaluative one, does it follow that it is not normative at all? Not necessarily. Correctness may be a normative property in its own right, neither deontic nor evaluative. This sort of view, which Thomson endorses, becomes plausible when we reflect on the connection between doxastic correctness and correctness for other, non-doxastic attitudes. Attitudes like admiring a person, desiring ice-cream, fearing torture and intending to keep a promise can all be correct or incorrect–or, as we might say, *fitting* or *unfitting*. This point is familiar from fitting-attitudes accounts of value, to which Engel makes reference (ibid., 203). As he says: "Correctness is determined by the way our attitudes *fit* a certain feature."

Some proponents of fitting-attitudes accounts of value understand fittingness in terms of reasons: for an attitude to be fitting is for there to be reasons for it. But it's not at all obvious that fittingness should be understood in this way. This is particularly clear for the case of belief. What is fitting–correct–to believe is not always what one has most or any reason to believe. My point here is that it is also far from obvious that fittingness must be understood in either deontic or evaluative terms. The domain of the normative may, as Thomson argues, be richer and more differentiated than philosophers have traditionally realised.

To what extent is the sort of view I am suggesting here a rival to Engel's? I am not sure. But I do want to insist that it is distinct from the evaluative view that some of his remarks may lead us towards.

While fittingness need not be understood in terms of reasons, it may nonetheless be the case that, in so far as fittingness *guides*, it does so through reasons. When we engage in theoretical reasoning, what we think about, and what determines our conclusions, is reasons for believing this or that. How could this count as being guided by fittingness? Just as the evaluativist may claim that reasons are generated by values, so the defender of the view presently under consideration can hold that there is a systematic or constitutive connection between reasons and fittingness. A full explanation here may therefore require an account of reasons in terms of fittingness, rather than the converse.

Whether the sort of view sketched here is sustainable or not, I look forward to reading many more of Pascal Engel's contributions to this debate.

3. References

Bykvist, K., and Hattiangadi, A. (2007). Does Thought Imply Ought? Analysis 67, 277–85.

Engel, P. (2013). Doxastic Correctness. Proceedings of the Aristotelian Society Supplementary Volume 87, 199-216.

Glüer, K., and Wikforss, Å. (2009). Against Content Normativity. Mind 118, 31–70.

McHugh, C. (2012). The Truth Norm of Belief. Pacific Philosophical Quarterly 93, 8–30.

Shah, N. (2003). How Truth Governs Belief. Philosophical Review 112, 447–82.

Smithies, D. (2012). The Normative Role of Knowledge. Noûs 46,265–88.

Thomson, J. J. (2008). Normativity. La Salle, IL: Open Court.

Wedgwood, R. (2013). Doxastic Correctness. Proceedings of the Aristotelian Society Supplementary Volume 87, 217-34.

46

Norms for emotions: intrinsic or extrinsic?

STÉPHANE LEMAIRE

Abstract It is often suggested that emotions are intrinsically normative or that they have conditions of correctness that are intrinsic. In order to assess this thesis, I consider whether the main argument in favor of the normativity of belief can be transposed to emotions. In the case of belief, the argument is that when we wonder whether to believe that p, we acknowledge that we must abide by some norms. This is understood as showing that these norms are intrinsic to the concept of belief. In contrast, it appears that no similar constraint applies when we deliberate about emotions. Indeed, I argue that extrinsic norms are sufficient to understand thoroughly the normativity of emotions. Therefore, the postulation of intrinsic norms or correctness conditions seems unmotivated. Worse, if emotions had intrinsic norms or correctness conditions, they would manifest themselves when we wonder whether an emotion is appropriate in this or that context. But they don't.

1. Introduction

We make numerous judgments about the appropriateness of emotions or about the emotions that we should or should not experience. Nobody disputes that some of these judgments are grounded on norms that are extrinsic to emotions but that apply to emotions as they apply to action. For instance, prudence seems often to suggest that we should be humble in order to avoid others' jealousy. Maybe, one is even morally required to restrain the joy caused by a success in the presence of less fortunates. Whether it is my own interest or another's interest that is—or engenders—a reason to restrain my pride, in both cases it is clear that the norm[1] that applies to my excitement is not grounded in the very nature of excitement. However, it is often considered obvious that beyond these extrinsic norms, emotions are also subject to intrinsic norms, or at least that there is a sense in which they are correct or incorrect independently of any extrinsic norm.[2] The aim of this paper is to object to these views.

A possible strategy in order to justify the intrinsic normative or correctness of emotions is to claim that emotions have evaluative contents. Indeed, if this is accepted, then it follows that an emotion in response to an object is correct iff the object has the value represented in the emotion's content. For instance, if one holds that fear represents its object as dangerous or fearsome, then it follows that fear is correct iff its object is indeed dangerous or fearsome. Emotions would thus have intrinsic correctness conditions in virtue of the correctness of their evaluative content. However, it is not my aim in this paper to discuss the nature of emotional content and whether emotions represent or present evaluative properties.[3] There are two reasons to take another route. First, it seems to me that the phenomenology of emotions and the nature of their content is not easy to grasp very firmly. Therefore, any objection to the

[1] In this paper, I use only (with few necessary exceptions) the word 'norm' for normative considerations which might be expressed otherwise in terms of reason, value, normative rules, requirements, etc. and which may be different in nature and in strength since one may argue in particular that prudential, moral, personal and aesthetical norms have very little in common. There are two reasons for that. First, I want to stick to the usage in the literature on the aim of beliefs, which is mostly couched in terms of norms, and not to rephrase it. Second, I do not want to commit myself on the nature of the normative vocabulary or normative properties, objects, or whatever, that are considered. Hence, by talking of norms, my aim is mainly to avoid long disjunctions. For what I have to say, this quite abstract way of talking will be precise enough.

[2] See for instance D'Arms & Jacobson (2000) and authors such as Skorupski (2007) and Danielson & Olson (2007) who attempt through this claim to defend a fitting attitude analysis of value against the so-called wrong kind of reason problem.

[3] For a thorough discussion of the presentational version of this view, see our Dokic & Lemaire (2013).

view that emotions have an evaluative content may lack the kind of immediate appeal that we would like to have. Second, even if we were convinced that the content of emotions is not evaluative, there are other ways to flesh out the idea that emotions are intrinsically normative or have intrinsic correctness conditions.[4] A central one is to claim that emotions are not assessed as correct or incorrect only through an assessment of their content but that this assessment relies partly on the types of emotions we consider and partly on its content. For example, one would claim that an emotion of fear is correct if and only if its object has the property that makes the attitude appropriate–here, danger–although it is not part of the content that the object is dangerous. Given this second avenue to the view that emotions have correctness conditions, it would be nice to have an argument that objects directly to the idea that emotions are intrinsically normative or have intrinsic correctness conditions. The goal of this paper is to provide such an argument. My strategy to reach this goal is to transpose and assess the main argument that has been offered in favour of the intrinsic normativity or correctness of beliefs. This argument, that I will call the doxastic deliberation argument, starts from what is presented as a fact about deliberation. Then, it claims that the best explanation of this fact is that beliefs are intrinsically normative. Section 2 of the paper is thus very straightforward : I recall the doxastic deliberation argument and I show that its transposition to emotions pleads against the intrinsic normativity of emotions. In section 3, I turn to a more modest strategy to which the defender of the intrinsic normativity or correctness of emotions could retreat. The starting point of the argument would now grant that beliefs and emotions are only subject to extrinsic norms while insisting that the application of these norms relies on an aspect of beliefs and emotions that is independent of the extrinsic norms. The hope would then be that this element of independence is able to ground at least the intrinsic correctness of beliefs and emotions, even if this correctness is understood non-normatively. But again, it will appear that emotions are different from beliefs on this count and that there is no ground to attribute to emotions even non-normative correctness conditions. In Section 4, I consider several objections to my argument and Section 5 draws its main conclusions.

[4] Mulligan (2007) is certainly a leading proponent of this view. Interestingly, he defends the intrinsic correctness thesis while denying that emotions have evaluative content, or in his words, « disclose values » (p. 222). Deonna & Teroni (2012) have more recently adopted the above combination of views.

2. Doxastic deliberation and deliberation about emotions

The doxastic deliberation argument to the effect that beliefs are intrinsically normative starts from what is taken to be a fact : when one wonders whether to believe that p, one excludes prudential and other normative considerations in favour of believing that p and one focuses exclusively on considerations that are relevant to the truth of p. In other words, the question as to whether one should believe that p is answered by the question as to whether p. As Shah makes very explicit, « the phenomenology of deliberation 5 [...] is that evidence is the only kind of consideration that can provide a reason for belief » (2003 : 464). This phenomenon is called in the literature[5] the *transparency* of doxastic deliberation. Once this much is accepted, one may wonder why doxastic deliberation is regulated by an exclusive concern for truth? The response offered by the normativist about beliefs is that it is because our concept of belief encompasses a norm to the effect that one should believe that p if and only if p.[6] Starting with the phenomenon of transparency, we conclude that beliefs are intrinsically normative.

Can we transpose this doxastic deliberation argument to emotions ? In order to respond to this question, we need first to clarify what would be the transposition of the argument to emotions. It would run like this : First, we deliberate about the kind of emotions which we should have in response to this or that object. Second, this deliberation about emotions sets aside some considerations and focuses exclusively on others. Therefore the concept of each type of emotion encompasses a norm that tells precisely which considerations are acceptable when we deliberate whether one should have an emotion of that type.

Do we have the elements to make the argument go through ? Firstly, do we deliberate about the emotions that we should have in this or that context ? We do. Moreover this deliberation is not completely inefficient. If I become convinced that my rage in a given context is counterproductive, it may help me to control my emotion and thus to have a different emotion. It may even change my emotional dispositions. For sure, we do not have emotions at will but the same remark is true for beliefs. Even if the deliberations about which beliefs or which emotions we should have do not directly and immediately change our beliefs and emotions, they nevertheless inform and affect the for-

[5] This literature has its roots in Evans (1982) and Moran (2001).

[6] Several objections have been raised about this formulation of the norm. For the sake of the argument, I will make as if there is a formulation of this norm that avoids the objections that have been raised. For a very recent defense of the possibility of such a formulation, see Engel (2013).

mation of our beliefs and emotions. The first element is thus present : we deliberate about emotions and it influences our emotional dispositions. Let us then see if we have the second element : Does deliberation about emotions exclude some considerations at the benefit of others ? To begin with, there is an important dissymmetry between doxastic deliberation and deliberation about emotions insofar as various practical norms are considered relevant in the latter case. Prudential considerations are central and seem able to explain thoroughly why and when we consider emotions such as fear, disgust and jealousy as appropriate. Moral considerations seem to underlie the appropriateness of shame and guilt. They seem to contribute also to some of our judgments about the appropriateness of sadness and admiration. In particular, it is certainly for moral reasons that we believe that it is proper to—or that we should—admire moral behaviours. Consider even the case of sadness. Although not obvious at first sight, it seems that sadness is sometimes required for moral reasons. If I am not sad enough over the death of my friend, I know that some will consider that I am somehow unfaithful to our friendship or even that I was only pretending to consider him as a friend of mine. Beyond prudential and moral considerations, aesthetic considerations seem to bear on the appropriateness of laughter, amusement and admiration, among others. Finally, if we look at the various positive and negative emotions that result from the fulfillment or frustration of our desires (e.g. joy and disappointment), it seems that our deliberation focuses on the question whether the desires themselves, as the underlying causes of these emotions, were rational : we wonder whether having the desire was in the first place a good idea, whether we had the means to fulfill our desire, whether its object would really make us happy, good, etc. In short, if we deliberate about emotions, this deliberation is not exclusively focused on considerations that could be interpreted in terms of truth conditions or in term of *sui generis* norms that would be specific to emotions[7] whereas it seems that doxastic deliberation excludes all the considerations that are not related to the truth of the content of the belief considered.

The foregoing remarks may lead the defender of intrinsic normativity to claim that all that has been shown is that the intrinsic norms of emotions must be understood in terms of norms that are prudential, moral, aesthetic, etc. Why not suggest that the intrinsic norms of each type of emotions are in terms

[7] Indeed, it has been argued by several authors (Skorupski, 2007 ; Danielson and Olson, 2007; and more recently Chappell, 2012) that there is something like a concept of evaluative reason, of correctness or of fittingness which is relevant when one considers the appropriateness of emotions and which is a primitive normative notion.

of these norms ? Indeed, there is after all no reason to consider that intrinsic norms must be *sui generis* norms. Why not suggest then that each type of emotion has an intrinsic link to some norms, although the existence of these norms does not rely on emotions. For instance, it would be part of the concept of fear that its appropriateness should be understood exclusively in terms of prudential considerations. We would thus have again for each type of emotion a divide between a set of intrinsic norms and the extrinsic norms excluded from the first set.

Although interesting, I believe that this proposal is much less plausible than the alternative one to the effect that all the norms just considered apply without discrimination to all emotions. For instance, prudential norms apply to almost all types of emotions : to fear and disgust very obviously, but to many others, maybe to all others. For instance, even if there are moral reasons to experience guilt, the experience of guilt should not be so important as to prevent agents from acting with a certain degree of spontaneity at least in some circumstances. Even if guilt is morally justified insofar as it indicates that one is taking responsibility for one's deed, guilt must end at some point. Why ? For prudential reasons : because guilt diminishes our well-being and diminishes our ability to act spontaneously, it should not be too important. Similarly, anger is certainly required for prudential reasons as a response to acts of aggression, and especially to illegitimate ones, but there is also a point at which too much anger seems to diminish our well-being more than it helps us to confront situations of conflict. The argument can be generalized to all negative emotions. Even if negative emotions are not considered as required by prudential reasons as fear and disgust are, there is a degree at which and objects for which the loss of well-being produced by these negative emotions outweighs their moral or prudential benefits. This point can even be extended to positive emotions. On the one hand, we certainly have prudential reasons to experience positive emotions insofar as some of them are good to experience. But, on the other hand, there are also prudential limits since it is prudent to be aware of and to worry to some extent about the dangers and all the ways in which our actions and life might go wrong. In other words, norms of prudence explain why we should favour positive emotions and limit them. Therefore, I claim that the upshot of all these considerations is that emotions are not paired with specific norms to the exclusion of others in virtue of their own nature. Rather, the appropriateness of emotions is better understood as the result of general and extrinsic norms that are not specific to any type of emotion. They only apply differently depending on the nature of these emotions and their objects.

In summary, deliberation about emotions does not exclude extrinsic norms such as prudential, moral or aesthetic norms. Moreover, it cannot even be argued that within these considerations we can discern those which are intrinsic to each type of emotion and those which are extrinsic. To this extent, deliberation about emotions differs from doxastic deliberation because only the latter is able to distinguish intrinsic norms and to exclude prudential, moral and aesthetic norms as extrinsic.

Now, the defender of the intrinsic normativity of emotions can still argue that even if extrinsic norms apply to emotions very broadly, it does not prove that emotions have no intrinsic norm that operates by default. She may acknowledge that practical considerations sometimes overwhelm the intrinsic norm of emotions but that nevertheless each type of emotion has its proper intrinsic norms. The problem with that response is that prudential, moral and aesthetic considerations, which are all extrinsic norms, seem *sufficient* to explain thoroughly our intuitions about the norms of emotions. This point has already been very clear with fear and disgust, for which prudential norms provide all we need to explain the appropriateness of fear. Similarly, we have seen that the norms that apply to shame can be completely understood in terms of moral and prudential norms. Finally, emotions such as amusement seem to be explainable as we have seen above in terms of prudential norms and in terms of aesthetic norms : some fun but not too much insofar as many situations seems to require other responses. And when one wonders at which joke one should laugh, the answer is certainly : at the jokes that are superior in terms of one or another aesthetic property. For sure, to determine precisely for all emotions the extrinsic norms that could explain their normativity is a task that has yet to be accomplished. But there is no reason from the cases already considered to doubt that it can be achieved. Therefore, one may wonder : why should we hypothesize intrinsic norms if we do not need them to explain our judgments about the appropriateness of emotions ? Obviously, the burden of proof falls on those who claim that emotions have intrinsic norms.

Another response that I am going to develop in the next section is to grant that emotions have no intrinsic norms while insisting that they nevertheless have non-normative correctness conditions. A good reason to follow this path is that one may argue that this is in fact true of beliefs. Hence, if this strategy proves successful for beliefs, why could we not apply it to emotions ?

3. Categorisation schemes and force-makers

To begin with, this more modest strategy raises a doubt about the starting point of the doxastic deliberation argument. As we have already seen, this starting point is the supposedly obvious fact that doxastic deliberation excludes all non-evidential considerations. But is this truly a fact ? Consider a woman who reflects on the best strategy to win a 100-meter race. Let us suppose that she has come across empirical studies which show that believing that one is going to win increases the probability of winning. Given her goal, it is perfectly clear that, in order to assess whether she should believe that she is going to win, she can and should take into account that it will improve her probability of winning. In other words, the question as to whether to believe that p does not reduce to the question as to whether p.

The normativist about beliefs is not defeated by such an objection. She may acknowledge that practical considerations sometimes overwhelm the intrinsic norm of belief but that nevertheless the concept of belief implies that one should believe the truth by default. As we have already seen with emotions, the problem with that response is that it can be argued, as Papineau (2013) has indeed, that the only norms that require us to believe truths are extrinsic, that is, prudential, moral or maybe personal.[8] Why should one be interested in truth if it does not further any of our interests—among which we may count our interest for truth—or any of the interests that we should pursue ? Therefore, the burden of proof falls again on those who wish to claim that belief encompasses an intrinsic norm of truth.

In what follows, I do not intend to show that such an argument cannot be given in the case of belief. Rather, I will suggest that if there may be grounds for such an argument in the case of belief, we lack similar grounds in the case of emotions. More precisely, I will show that even if we assume, for the sake of the argument, that beliefs have only extrinsic norms, beliefs still have a relation to truth that plays a role in relation to these norms. In contrast, it will appear that emotions are not similar on this count.

Let us then assume for the sake of the argument that no ought applies to beliefs in virtue of their nature. If truth is to be pursued, it is only in virtue of extrinsic norms such as prudential or moral norms as Papineau suggests. However, even if the norms that demand of us to have true beliefs are extrinsic, the distinction between true and false beliefs is not itself extrinsic to

[8] I do not take any stance on the view suggested by Papineau and according to which there are personal norms. This is certainly not required by the argument that I develop here.

beliefs. The latter exists independently of any extrinsic norm and is relevant for the norms that apply to beliefs. As Tim Schroeder has rightfully shown, two elements need to be distinguished when considering a norm. As an initial step, norms « may be thought of as dividing up domains into mutually exclusive and jointly exhaustive categories. Thus, etiquette divides actions into those which are polite, those which are impolite and those which are neither. » (2003 : 2). However, the existence of such a *categorisation scheme*, as Schroeder calls it, is insufficient in and of itself to constitute a norm. For a norm to exist, something more is needed that Schroeder calls the *force-maker*. On Schroeder's view, this is « what takes one category and makes it true that it is the good, to be preferred, correct, or otherwise normatively positive category. » (2003 : 3). Relying on this conceptual apparatus, the distinction between true and false beliefs appears as a categorical scheme that is in need of a force-maker. Papineau's view can thus be specified by saying that although the force-makers of the norms that apply to belief are extrinsic, these extrinsic force-makers nevertheless give force to a categorisation scheme that is independent of these norms. In a nutshell, the categorisation scheme that distinguishes true from false beliefs receives extrinsic force-makers in the form of extrinsic norms such as the rules of prudence, morality or even the pursuit of personal goals.

Moreover, it must be emphasized that it is *in virtue* of the externality of the force-makers that considerations that are not evidence should be excluded in doxastic deliberation. It is because I want to fulfill my desires or because I have duties that I need to have good information about the world. Without correct information about the world, I would risk failing in both aims. Thus, it appears paradoxically that it is precisely because we pursue practical ends that our beliefs must track the truth. No doubt, we have practical reasons to believe what may not be true in specific circumstances such as the 100 meters competition considered above. However, in most cases, it is precisely because the force-makers of the norm that apply to beliefs are extrinsic that doxastic deliberation must exclude the considerations that are irrelevant to the truth of the belief considered. It is because we want to succeed in our action that we need true beliefs. Indeed, the more important our aims are, the more we must abide by the truth norm. In other words, the claim that the force-makers of the norms that apply to belief are extrinsic allows us to understand why these force-makers apply to a categorisation scheme that is independent of them.

In summary, even though the force-makers of the norms that apply to beliefs are extrinsic, it seems that these extrinsic norms apply to a categorisation scheme that distinguishes true from false beliefs and which is independent of these extrinsic norms. The normativist about beliefs may not be satisfied by

these concessions but at least they mark what his opponent must concede, and on which the former may hope to argue in favor of the intrinsic correctness or even normativity of beliefs. However, my aim here is not to see whether we can construct such an argument for belief. Rather, I will show that there is no reason to attribute to emotions a proper and similar independent categorisation scheme. It will follow that the case for an intrinsic correctness or normativity for emotions is even worse than the parallel case for beliefs insofar as emotions lack an independent categorisation scheme.

Let us then return to deliberation about emotions. We have already seen that extrinsic considerations are relevant to deliberation about emotions. To this extent, there is a point in arguing that the norms that apply to both emotions and beliefs are extrinsic. However, we have seen that the categorisation scheme in the case of belief is not reducible to the extrinsic norms that apply to beliefs. If prudence is the force-maker of belief norms, the categorisation scheme to which the force-maker applies distinguishes true from false beliefs. Does our deliberation about emotions show a similar dissociation ? It does not. Consider deliberation about fear. The only relevant consideration is obviously prudential : If the object is dangerous, fear is appropriate and if it is not, fear is inappropriate ; there is nothing to add to that. The norm of prudence is thus the force-maker of the norm insofar as it is in virtue of prudential norms that it is appropriate to be afraid of dangerous objects. But, in addition—and this is the crucial point—the norm of prudence determines also the categorisation scheme of the norm since this categorisation scheme divides fears into those that are prudent and those that are not prudent. Indeed, this seems to be just another way to say that fear is justified if and only if it is prudentially justified in response to danger. Thus, the extrinsic norm of prudence does all the work insofar as it is responsible for the categorisation scheme that distinguishes fears as prudent or not, and insofar as it also provides the force-maker—prudential norms— for which we should choose the first class of fears—those that are prudent. Bringing these two elements together, we obtain the following trivial result : it is in virtue of prudential norms that we should have prudential fears ! Though trivial, it must be constrasted with the corresponding *motto* for beliefs : it is in virtue of prudential norms that we should have true beliefs.

If we now add the results of the previous section to those of the present one, it seems that the comparison between doxastic deliberation and deliberation about emotions reveals the kind of norms that apply to emotions. The main difference between these norms is that extrinsic norms are all we need to assess the appropriateness of emotions. Not only do extrinsic norms ap-

ply to fear but they provide us with the categorisation scheme with regard to which the extrinsic norms are force-makers. We do not need anything else in order to apply these extrinsic norms. In contrast, the norms that apply to beliefs, even if extrinsic, rely on or make use of an independent categorisation scheme which may allow us to say that beliefs have intrinsic correctness conditions in a non-normative sense. This latter fact may even be a starting point to argue that beliefs are intrinsically normative, although it seems to me that if the transparency of doxastic deliberation is not a fact, the normativist about belief still has to provide us with a good argument to convince us of her view.

4. Objections

Among the possible objections that may be raised against the above arguments, I will discuss three of them. The first one is to point out that emotions have a biological function as beliefs do. Hence, the argument goes, there is a sense in which emotions have intrinsic correctness conditions and maybe even normative correctness conditions.

This argument faces two problems. First, biological functions are not normative. Admittedly, the biological function of, say, beliefs is to track the truth and to this extent, one may say that beliefs accomplish correctly their function when they are true. In that sense, beliefs are correct if and only if they are true. However, to ascribe a biological function to a system is not to ascribe it a normative property and as such, biological functions are not force-makers. This is because to say that a system or an organ has a biological function is just to recall that it has been selected for certain effects in the past. It says nothing about the effects that it should have. That is why there is a crucial difference between a system that has been built with the intention of accomplishing a certain function, in which case a norm applies to the system in virtue of the intention that has been conducive to its existence, and the case of natural evolution where no such intention is present. Thus, although beliefs have correctness conditions as products of evolution, these correctness conditions are not normative. This argument applies to all mental states and especially to emotions. Hence, that emotions have been selected because they were efficient responses for our survival and reproduction allows us to say that they have correctness conditions but not that they have normative correctness conditions.

At this juncture, the defender of intrinsic correctness might think however that she has all she wants : have we not just granted that emotions have

non-normative intrinsic correctness conditions ? Yes, but the problem of these correctness conditions is that they cannot play the role of the categorisation scheme for which prudential or other extrinsic norms would provide their force-makers. The reason is that neither prudential nor moral nor any other extrinsic norm may favour emotions that are correct from an evolutionary point of view. Consider the case of envy. It may be, and at least it could be, that whatever person you consider, it is neither moral nor prudential to be envious of her. Hence, not only would the correctness conditions not be themselves normatives but they will not receive the support of extrinsic norms. In other words, while extrinsic norms and intrinsic conditions of correctness can be added to yield norms that apply to beliefs, this is not true for emotions because they are not suited for one another. The extrinsic norms that are the force-makers apply only to the categorisation scheme that divides emotions as, for instance, prudent or not. They do not, except by accident, apply to the emotions that are correct from an evolutionary point of view.

It might be replied that the crucial point is the existence of these intrinsic correctness conditions, and not that there are no extrinsic norms that enforce the intrinsic correctness conditions. The lack of extrinsic norms to have the emotions that are intrinsically correct may even be considered as a further reason to see the correctness of emotions as independent from extrinsic considerations.

But then one might wonder why we should be concerned with these correctness conditions. Consider again the case of envy or, at least, of a possible emotion close to envy. Let us assume that nature has endowed us with a disposition to experience this emotion in order to reproduce ourselves as much as possible even if at a very high price. To this extent, envy has correctness conditions from an evolutionary point of view. But why should we bother to be correctly envious? For sure, one might reply that there may be no response to this question and that one is ill-advised to ask for a justification of correctness conditions. One should simply acknowledge the fact that these intrinsic correctness conditions exist.

However, this answer cannot be offered by those who take evolutionary considerations as their starting point in order to explain the correctness conditions of emotions. For if biological functions set correctness conditions, it is in virtue of their effects. Therefore, it is incoherent to claim on the one hand that no explanation is needed in order to show that some correctness conditions are relevant to us, and on the other hand to build the correctness conditions for emotions on considerations that explain their existence in terms of their effects.

Let us then turn to a second objection. Against the argument presented in the previous section, one might be tempted to insist that emotions are not really different from beliefs since they cannot be appropriate if their content is not true. Hence, it might be argued that emotions can also be categorised as having a true or false content. Isn't this not showing that emotions are similar to belief insofar as they both have an independent categorisation scheme ?

Unfortunately, this objection relies on a conflation, for we must distinguish the norms that apply to the cognitive base of emotions and the norms that apply to emotions themselves. Certainly, if one experiences fear in response to the illusory perception of a dog, the emotion is inappropriate, but this is only because the perceptual experience is itself incorrect. The appropriateness of emotions relies on these other norms when they exist but they are not thereby intrinsic to emotions. Indeed, it is worth noticing that an emotion in response to an imagined fact may be appropriate although the imagined fact need not be true. This shows, first, that the appropriateness of an emotion does not systematically rely on the truth of its content or of its cognitive base. Perceptions and beliefs are not the only possible cognitive bases of emotions. Secondly, it shows that the appropriateness of emotions in its wider sense has two components, a component that applies to its cognitive base and another component that applies to emotions once its cognitive base is correct. Since the former cannot be seen as intrinsic to emotions insofar as it concerns primarily the normativity of the cognitive bases of emotions and only derivatively the normatively of emotions, it cannot ground an argument for the intrinsic normativity of emotions.

A final objection might go in the opposite direction : it would claim that there is no difference between the categorisation schemes of the norms that apply to beliefs and emotions. After all, we can categorize beliefs by distinguishing those that are prudential from those that are not. Therefore, the objection goes, they both distinguish prudent from imprudent states, whether these states are emotions or beliefs. For sure, prudential beliefs are in most cases true beliefs, but one can insist that this is not sufficient to distinguish beliefs from emotions. Therefore, I would not have shown that the categorisation scheme of the norms that apply to emotions is relevantly different from the categorisation scheme of the norms that apply to beliefs.

The main problem with this objection is that this way of closing the gap between emotions and beliefs goes in the wrong direction. Instead of showing that emotions are similar to beliefs in having an independent categorisation scheme, it shows that beliefs are similar to emotions in having no categorization scheme independent of extrinsic norms. Therefore, this objection is

nothing other than a further concession to those who claim that beliefs have no intrinsic correctness conditions. In any case, it is worth noticing that even if we grant that the normativity of belief is closer to that of the emotions, it is still true that the prudential assessment of a belief needs always to take into account the value increase that would result from the success of actions guided by a true rather than a false belief. Furthermore, insofar as we can conjecture that nearly any belief may contribute to the guidance of some action the success of which has some importance, the question as to whether we should believe p most frequently boils down to the question as to whether p is true. The cases in which the prudential norms to adopt a belief that p do not rely on the guiding role of p in action are certainly rare. This shows that even if it may be granted that the distinction between prudential and non-prudential beliefs cannot be reduced to the distinction between true and false beliefs, in most cases, the prudential beliefs will identify with the true ones. At the end of the day, this is a fact that does not apply to emotions since the appropriateness of emotions need not appeal to the notion of truth at any stage if we exclude the norms that apply to its cognitive base.

5. Conclusion

In this paper, I have tried to consider arguments from the ongoing debate on the aim of belief in order to assess whether emotions are intrinsically normative or have intrinsic correctness conditions without making any hypothesis on the evaluative content or nature of emotions. The central argument in this debate starts from the phenomenon of transparency that is taken to be manifest in doxastic deliberation. However, it appears that deliberation about emotions is *not* governed by a similar constraint : extrinsic norms such as prudential norms are relevant when we assess whether we should, say, fear the barking dog. In order to avoid this objection, the defender of the intrinsic normativity or correctness of emotions may adopt a more modest strategy. She may grant that there is no phenomenon of transparency and that only extrinsic norms regulate the question as to whether one should believe something. Nevertheless, she may insist that the categorisation scheme that distinguishes true beliefs from false ones is independent from the force-makers that apply to it and that this may, at least, ground that belief has intrinsic non-normative correctness conditions. She would finally contend that a similar argument can be constructed in the case of emotions. But, as we have seen, this is not the case. Hence, the comparison between doxastic deliberation and deliberation

about emotions teaches us several things. It shows that we do appeal to extrinsic considerations to assess our emotions and moreover that these extrinsic norms are sufficient to understand everything we need concerning the normativity of emotions. Even the categorisation schemes to which extrinsic norms apply are derived from these extrinsic norms. In my view, this provides a serious objection against the idea that emotions have intrinsic normativity or correctness conditions ; if emotions were intrinsically normative or have intrinsic correctness conditions, then these intrinsic features of emotions would manifest themselves when we deliberate about them.

6. References

Chappell R. Y. 2012. « Fittingness : The sole normative primitive ». *The Philosophical Quarterly* 62, pp. 684-704.

Danielson S. & Olson J. 2007. « Brentano and the Buck-Passers ». *Mind* 116, pp. 511-22.

D'Arms J. & Jacobson D. 2000. « The moralistic fallacy : On the 'Appropriateness' of Emotions ». *Philosophy and Phenomenological Research* 61.1 (2000): 65-90.

Deonna J. & Teroni F. 2012. *The Emotions : A Philosophical Introduction*. New-York : Routledge.

Dokic J. & Lemaire S. 2013. « Are emotions perceptions of value ? », *Canadian Journal of Philosophy* 43.2, pp. 227-247.

Engel P. 2013. « In defense of normativism about the aim of belief », in T. Chan (ed.) *The Aim of Belief*. Oxford : Oxford University Press.

Evans G. 1982. *The varieties of reference*. Oxford : Oxford University Press.

Moran R. 2001. *Authority and Estrangement*. Princeton : Princeton University Press.

Mulligan K. 2007. « Intentionality, knowledge and formal objects », *Disputatio* 2.23, pp. 205-228.

Papineau D. 2013. « There are no norms of belief » in T. Chan (ed.) *The Aim of Belief*. Oxford : Oxford University Press.

Schroeder T. 2003. « Donald Davidson's theory of mind is non-normative », *Philosopher's Imprint* 3.1, pp. 1-14.

Skorupski J. 2007. « Buck-Passing about Goodness », in D. Egonsson, J. Josefsson, B. Petersson, and T. Rønnow-Rasmussen (eds.) *Hommage à Wlodek*. online resource : http://www.fil.lu.se/hommageawlodek/index.htm

Shah N. 2003. « How truth governs beliefs », *The Philosophical Review* 112.4, pp. 447-482.

47

Truthful Liars *

GIOVANNI TUZET

> La vérité est si obscurcie en ce temps, et le mensonge si établi, qu'à moins que d'aimer la vérité, on ne saurait la connaître.
>
> Pascal

*A first draft of this paper was presented in a workshop in Montreux on September 2007. I wish to thank two persons in particular: Michael Esfeld, who hosted me with a scholarship at Lausanne University in 2003-2004 and also invited me to participate in that workshop, and Pascal Engel, who participated in that workshop as well and has always been for me a source of philosophical inspiration and an example of intellectual integrity.

1. *Is Truth a Norm?*

In what sense, and of what, is truth a norm? Is it a norm of inquiry? Of belief? Of judgment? Or of assertion? In this paper I will claim that truth *is not* a norm of belief and assertion in the sense that having a belief and making an assertion commits you to the truth of what you believe and assert. At the same time, in a different sense of normativity, I will claim that truth *is* a norm of belief and assertion in the sense that belief aims at truth and the practice of making assertions aims at truth-transmission.

In particular, the thesis that truth is a norm of assertion, in the first sense just sketched, has been held explicitly or not by many important philosophers in the analytic tradition and in the pragmatist as well. The basic idea is that asserting that p commits you to the truth of p. Sometimes the thesis is presented, for instance by Peirce, in terms of responsibility: by virtue of some social norm, asserting that p makes me responsible of the truth of the propositions I assert[1]. Sometimes, in particular by Frege, the thesis is taken to depend on the very nature of the act of assertion: since to assert that p is to assert that p is true, if one asserts that p he is committed to the truth of p[2]. The same is held by Searle in the contemporary field of speech act theory[3].

Now, I take that thesis to be wrong and I will try to show that making an assertion commits one to *sincerity*, not to truth. This will explain how it is possible to be *truthful liars*. But this does not throw the concept of truth out of the picture: since asserting that p is asserting that one believes that p, and believing that p is believing that p is true, when we make an assertion we commit ourselves to believe that what we say is true. Plainly this is correct if believing that p is believing that p is true. So I don't want to present an opposition, say the truth norm vs. the sincerity norm, but rather see *in what different senses of normativity* truth and sincerity are norms of assertion on the one hand, and truth and justification are norms of belief on the other. What in the end must be made clear, in effect, is the sense in which truth is a norm of belief and assertion even though believing and asserting rather commit one to justification and sincerity. That sense is, in my view, the *teleological* one in which belief aims at truth and assertion aims at truth-transmission.

To discuss these topics, I will present in §2 Peirce's theory of assertion, which will be critically examined in §3, where the case of the Truthful Liar

[1] Cf. e.g. CP 2.314-315, 2.252, 5.30, 8.313.

[2] See Frege (1918-1919: 294 Eng. trans.).

[3] "When I say something and mean it, I am committed to the truth of what I say. And this is so whether I am sincere or insincere" (Searle 1999: 144).

and other similar cases (the Reliable Falsity-teller and the Unreliable Truth-teller) are presented. Finally I will consider in §4 the sense of normativity in which, still, truth is a norm of belief and assertion[4].

2. *Peirce's Theory of Assertion*

Peirce did not write a specific work on assertion, but in his papers one can find several considerations on it and its relations to proposition, belief, and judgment.

In a paper written around 1895, he defines assertion as the act of the speaker communicating to the listener that he has a certain belief, namely that in certain circumstances a certain idea is for him "definitively compulsory".

> The assertion consists in the furnishing of evidence by the speaker to the listener that the speaker believes something, that is, finds a certain idea to be definitively compulsory on a certain occasion (CP 2.335).

In this and in other passages Peirce remarks the *pragmatic dimension of asserting*: every assertion is an act communicating a belief. On this basis he also claims that asserting makes one responsible for what is asserted (CP 5.546-548, c. 1908; cf. CP 2.315).

Of course assertion is not to be confused with proposition: the same proposition can be articulated to various propositional attitudes, giving place to different pragmatic relations. A proposition can be doubted, asked, judged, asserted, ordered.

> I may state it to myself and worry as to whether I shall embrace it or reject it, being dissatisfied with the idea of doing either. In that case, I doubt the proposition. I may state the proposition to you and endeavor to stimulate you to advise me whether to accept or reject it: in which I put it interrogatively. I may state it to myself; and be deliberately satisfied to base my action on it whenever occasion may arise: in which case I judge it. I may state it to you: and assume a responsibility for it: in which case I assert it. I may impose the responsibility of its agreeing with the truth upon you:

[4] This paper elaborates on some previous work like Tuzet (2006) and Canale and Tuzet (2006), where more details are given about Peirce, Frege, Searle and Brandom.

> in which case I command it. All these are different moods in which the same proposition may be stated (NEM 4: 39).

An interesting point made by Peirce concerns the difference between asserting and judging. In his *Syllabus*, a work dating (presumably) from 1902, he refers to "judgments" as acts of mental acceptance of propositions (CP 2.309). In another passage of the same work – even if there is no explicit reference to the act of judging – the distinction is put forward between the act of *asserting*, which implies some responsibility to *other subjects*, and the act of *assenting*, which implies some consequence for the *own conduct* of the assenting subject:

> an act of assertion supposes that, a proposition being formulated, a person performs an act which renders him liable to the penalties of the social law (or, at any rate, those of the moral law) in case it should not be true, unless he has a definite and sufficient excuse; and an act of assent is an act of the mind by which one endeavors to impress the meanings of the proposition upon his disposition, so that it shall govern his conduct, this habit being ready to be broken in case reasons should appear for breaking it (CP 2.315).

Peirce refers a bit vaguely to the social or moral law inflicting some sanctions on those who make a false assertion. What is worth noting is the defeasible character of this ascription of responsibility: he who makes a false assertion is liable to some penalties "unless he has a definite and sufficient excuse"[5]. On this basis Peirce remarks the difference between an act concerning the agent's own conduct – the act of *assenting* and undertaking the practical consequences of a certain proposition believed – and the act of *asserting*, namely the act of declaring to others the truth of a certain proposition (cf. CP 8.115, c. 1900).

In 1904 Peirce insists that *assertion* is not an act of pure signification, but a "public" act implying some penalties as possible consequences in case the assertion is false (CP 8.337). On the contrary *judgment* remains a "private" act, "the self-recognition of a belief"[6]. In Peirce's terms to *judge* is to *assent*, not to *assert*.

[5] On the ascription of responsibility and rights see Hart (1949).

[6] "According to my present view (I may see more light in future) the act of assertion is not a pure act of signification. It is an act of exhibition of the fact that one subjects oneself to the penalties visited on a liar if the proposition asserted is not true. An act of judgment is the self-recognition of a belief; and a belief consists in the deliberate acceptance of a proposition as a basis of conduct" (CP 8.337). But this view is disputable; cf. Brandom (1994: 158): "The judgment is the internalization of a public process of assertion". On acceptance cf. Burge (1993), Engel (1999) and (2000).

Such a distinction is basically maintained in a subsequent fragment (presumably of 1908) entitled *Judgment and Assertion* (CP 5.546-548)[7]. The fragment starts from the analysis of assertion. As already said, asserting implies an undertaking of responsibility. Whereas in 1902 (CP 2.315) Peirce referred to the social and moral law, now he refers more concretely to the legal practice.

> If a man desires to assert anything very solemnly, he takes such steps as will enable him to go before a magistrate or notary and take a binding oath to it. Taking an oath is not mainly an event of the nature of a setting forth, *Vorstellung*, or representing. It is not mere saying, but is *doing*. The law, I believe, calls it an "act". At any rate, it would be followed by very real effects, in case the substance of what is asserted should be proved untrue. This ingredient, the assuming of responsibility, which is so prominent in solemn assertion, must be present in every genuine assertion (CP 5.546).

This passage sketches a sort of speech acts theory *ante litteram*[8]. According to it, an *assertion* differs from other speech acts in virtue of the consequences it typically has. Asserting implies an assuming of responsibility: he who makes an assertion exposes himself to the consequences of it, namely to some penalties or sanctions in case the assertion is false (CP 5.546)[9]. This rightly happens because, according to the theory, asserting commits to the truth of the proposition asserted.

Then, supposing this is correct, what is the *ratio* of the norm imposing penalties or sanctions on those who make false assertions? Presumably the norm has a complex *ratio* but a central aspect of it is the desire to avoid the bad consequences faced by those who rely on false assertions. More precisely, when we assert a proposition we are responsible to those who will eventually orient their conduct on our assertion. So our liability for a false assertion is grounded in the fact that some negative consequences can be the case for those

[7] With the difference that in this fragment Peirce introduces the idea that judgment is something which *ripens* in the mind. See Tuzet (2006). On the normativity of judgment in a naturalist picture, see Papineau (2003: chap. 1).

[8] Cf. obviously Austin (1955) and Searle (1969).

[9] For a comparison of Peirce and Searle on these topics, see Brock (1981). Note that, according to Peirce (but using a later terminology), every "illocutionary act" has a perlocutionary aspect which enters its definition; from this point of view, Searle's separation of the illocutionary (primary) from the perlocutionary (secondary) is inadequate (cf. Searle 1969). What interests Peirce are the "real consequences" of assertions and judgments (CP 5.546-547).

who act on it, or in particular for those who suffer harm from a decision based on it – as is the case, for instance, of someone convicted on the basis of a false testimony (the legal example is made by Peirce himself in NEM 4: 249).

Conceiving of assertion as an act implying a responsibility is in tune with a pragmatist conception of meaning[10]. Any assertion has certain *effects*, which can be predicted, tested and assessed taking into account not only the ethical and practical but also the legal and institutional features of the situation[11]. But a point should be made clear, in my opinion: in our discursive practices we are not directly committed to the truth of our assertions, but rather to their *sincerity*. To put it differently: what is directly relevant for the ascription of responsibility is not the relation between what is asserted and what is *true*, but the one between what is asserted and what is *believed*.

3. *A Critique of Peirce's Theory*

According to Peirce we have to say that, first, the speaker is responsible of the truth of his assertions, and, second, in virtue of this very responsibility a false assertion is to be sanctioned. I take this view to be wrong[12]. The responsibility of assertion does not directly depend on truth, but rather on *belief*. We are not directly committed to the truth of our assertions: what matters is (a) what we *believe* to be true or false and (b) whether we assert what we believe.

Peirce claims that "one subjects oneself to the penalties visited on a liar if the proposition asserted is not true" (CP 8.337). However, I shall remark, a lie is not a false statement, but a statement that contradicts the actual belief of the speaker. Someone lies when he says that *p* and believes that *not-p*, even if it is true that *p*; and vice versa he lies when he says that *not-p* and believes that *p*, even if it is true that *not-p*. In this sense we should say that *sincerity* rather

[10] Cf. Tiercelin (1993: 303): "l'acte d'assertion proprement dit met en cause la vérité ou la fausseté de l'énoncé, et implique un engagement ou la responsabilité de celui qui l'effectue. L'acte d'assertion suppose donc d'une part, une analyse des conditions auxquelles l'assertion doit obéir pour être susceptible de rencontrer le vrai mais d'autre part aussi, des effets de toute nature, en raison de la multiplicité possible des interprétants, qu'elle peut avoir dans le contexte de la communication. L'acte d'assertion ne comporte donc pas seulement des *dimensions* pragmatistes: il est pragmatiste de part en part". Cf. Pape (2002).

[11] Cf. CP 8.313; Tiercelin (1993: 304).

[12] Unfortunately even some prominent scholars do not remark it is wrong. E.g. Hilpinen (2004: 156): "In an assertive speech act, the utterer of a proposition 'assumes responsibility' for its truth and is assumed to suffer some untoward consequences if the sentence turns out to be false, and the hearer or the interpreter will suffer the negative effects of the acceptance of false proposition unless he detects its falsity".

than truth is the norm of assertion. Note that this is not my own concept of a lie constructed for the occasion, but the concept that emerges from an analysis of our discursive practices and commitments. Let me clarify this point with the following examples.

I will start from the case of the *Truthful Liar* (or perhaps Wishful Liar). Theodore believes that the person who killed Basil is not Anastasia. (He has some evidence that it is Sophia). But he has some personal motive for wanting Anastasia to be convicted. So, when he gives his testimony at the trial for Basil's murder, he asserts that he saw Anastasia kill Basil. So, he believes that *not-p* but asserts that *p*. Now, suppose that it is in fact true that Anastasia killed Basil. Is Theodore lying? This is an analysis of his case:

The Truthful Liar

(1) believes that *not-p*;

(2) asserts that *p*;

(3) it is true that *p*.

So, is he lying? According to Peirce's account, he is not, since his assertion is true. However, according to mine, he is, for his assertion contradicts his actual belief. I have no empirical data to display, but my intuition is that many of us are willing to qualify that assertion as a lie.

Of course to settle the matter one should understand *what to count as a lie*. What are the conditions of a lie? Pascal's quote at the beginning of this paper implies that a lie is incompatible with truth; I suspect it is not so. Anscombe says that "a lie is an utterance contrary to one's mind"[13]. Austin says that insincerity is "an essential element in lying as distinct from merely saying what is in fact false."[14] But things are more complex. What is the correct concept of a lie if it is neither saying what is false nor saying something contrary to one's mind? Lynch claims[15] that to say what you don't believe is too strong, because, for example, actors on stage speak what they believe to be false, or because of the "mental reservation" doctrine (saying a crucial qualification to

[13] Anscombe (1957: 4). I must add that I cannot understand why after a couple of lines she says that "that a lie is an utterance contrary to one's mind does not mean that it is a false report of the contents of one's mind."

[14] Austin (1955: 41). Cf. Austin (1979: 99): "If I say 'S is P' when I don't even believe it, I am lying: if I say it when I believe it but am not sure of it, I may be misleading but I am not exactly lying."

[15] Lynch (2004: 147-148). I leave aside the issue of "alethic functionalism" that he raises in his later work.

yourself that makes your thought true) or the "equivocation" doctrine (saying something which is ambiguous knowing that the audience will take it in a different meaning). "So lying isn't simply saying one thing and believing another. To lie, rather, is to assert something you believe to be false with the *intention of misleading or deceiving*."[16] So the intentional aspect is crucial. But things are even more complex. We should consider the following:

> (i) asserting that *p* but believing that *not-p*, or vice versa asserting that *not-p* but believing that *p*, with the intention of misleading or deceiving;
>
> (ii) asserting something false.

Now, we could take an assertion to be a lie *either* when (i) obtains *or* when the conjunction of (i) and (ii) obtains. If we take (i) as a sufficient condition of a lie, Theodore is a truthful liar; if we think that the conjunction of (i) and (ii) is needed, he is rather a wishful liar: he tried to lie but didn't succeed[17].

Consider now some further situations. Imagine the case of the figure we might call the *Reliable Falsity-teller*: Theodore believes that the person who killed Basil is Anastasia and he has some good evidence that it is she. Then he asserts that it is she. But suppose that in fact it is Sophia. Is he lying? This is what happens here:

> *The Reliable Falsity-teller*
>
> (1) believes that *p* and is justified in believing that *p*;
>
> (2) asserts that *p*;
>
> (3) it is false that *p*.

Is Theodore lying here? If making an assertion commits to sincerity, of course he is not. If it commits to truth, he is, unless – being the commitment not strict (more on this below) – he has a "definite and sufficient excuse" (in our example, he has some good evidence that it is Anastasia the person who killed Basil).

Finally, consider the case of the *Unreliable Truth-teller*: imagine that Theodore believes it is Anastasia who killed Basil, but he is not justified in so believing

[16] Lynch (2004: 148), who also wonders when a lie is justified. A possibility is whether it passes the "publicity test" (149): "Lies are by nature secrets, but a justified lie is one that – probably – would pass muster were it exposed to the light of day and subjected to the examination of reasonable people."

[17] I am indebted to Susan Haack for a suggestion on this last point.

(he has no evidence at all). Suppose also that he asserts so and that it is in fact true that it is she who killed Basil. In this case:

The Unreliable Truth-teller

(1) believes that *p* but is not justified in believing that *p*;

(2) asserts that *p*;

(3) it is true that *p*.

Now is Theodore lying? According to Peirce's account, he is not, for his assertion is true. According to mine, neither, for his assertion is not only true but also sincere. However, we are disappointed by the fact that the Unreliable Truth-teller asserts what he has no justification to believe.

So, if I were to sum up these considerations, I would say that asserting that *p* commits to:

(a) being sincere;

(b) giving on demand a justification of the belief that *p*;

(c) accepting the consequences of the belief that *p*[18].

Therefore, if I am right, lying is determined by the relation between what is believed and what is asserted, in a way which is *relatively independent* from the truth. Why do I say "relatively independent"? Because that independence is relative to the cases in which a false belief may be nevertheless justified (and the asserting subject is capable of providing such a justification). The situation is different when an error is not justified: when, according to a given norm or criterion, one is expected to have a *true* belief, not a merely justified one. In this case, even if one were saying what really corresponded to his actual belief (strictly speaking, he were not lying), he would be responsible of the failure in forming accurately his own belief. (To talk law, in such cases the belief-forming subject wouldn't have a means- but an end-obligation, and he wouldn't be liable for fraud nor for negligence[19], but for the simple fact of failing to satisfy that end-obligation, namely to form a true belief). Thus, to be responsible of an assertion is not to be responsible of its truth, rather of its conformity to actual belief, on condition that the latter can be justified and the

[18] In this paper I don't deal with this last requirement, some considerations on which – elaborated from Brandom (1994) and (2000) – can be found in Canale and Tuzet (2005), (2006) and (2007).

[19] Cf. e.g. Holmes (1881: chaps. 3-4); Hart and Honoré (1959: part 2).

justification is acceptable because the requirement on the formation of belief is not a strict one that makes justification irrelevant. In this sense, *justification* rather than truth is a norm of belief and *sincerity* rather than truth is a norm of assertion[20]. Remember what Peirce specified in 1902: the asserting subject is not responsible if he has "a definite and sufficient excuse" (CP 2.315). Peirce got the point but perhaps didn't put a sufficient emphasis on it.

However, as I said above, this does not throw the concept of truth out of the picture: insofar as asserting that p is asserting that one believes that p, and believing that p is believing that p is true, we shall give an account of the normativity of truth for belief and at the same time we shall ask whether truth, in this sense of normativity, is also a norm of assertion.

4. *What Sense of Normativity?*

Pascal Engel is among the philosophers who have given an account of truth as being a norm of belief[21]. The starting question is whether *truth* is the goal of our epistemic practices and beliefs or whether this role is played by *justification*. Engel has claimed that truth is a norm of belief in the sense that it is *constitutive* of belief that "belief aims at truth"[22]. Justification is not enough, since a justified false belief ought to be abandoned or revised. This does not mean that if something is true then one ought to believe it, for this would commit us to believe even trivial and practically irrelevant truths. It rather means, according to Engel, that one ought to believe only what is true[23].

Engel rejects principle (A): For any p, if it is true that p, one ought to believe that p; instead he subscribes to principle (B): For any p, one ought to believe that p only if p (is true)[24]. He believes that the latter expresses a constitutive *norm of belief* and that truth is normative only insofar as there are norms of belief formation[25]. These norms provide criteria of justification and not the other

[20] Notice that this is not incompatible with the idea that truth is a norm of belief in a different sense: when we assert that p we are supposed to believe that p is true. On the conceptual relations between belief, assertion and truth, cf. Engel's remarks in Engel and Rorty (2005: 31 ff.).

[21] See e.g. Engel (2001), (2002), (2007). On truth's normativity and a pragmatist conception of truth (explaining it in terms of its role in discursive practice) cf. Price (2003). See also Esfeld (2005).

[22] Engel (2001: 43). On constitutive rules cf. Searle (1995).

[23] Engel (2001: 47). See also Engel (2002: chap. 5).

[24] Or rather, in a formulation that takes into account our standards of *knowledge*: For any p, believe that p only if, for all you know, p (is true). See Engel (2002: 128-129). Cf. Williamson (1996).

[25] Engel (2002: 129-130).

way round. Concerning the belief justification, then, Engel remarks there are internalist and externalist conceptions of it: for the former, one can have a justification for believing that *p* even though it is false that *p*; for the latter, if it is false that *p* one cannot have a justification for believing it[26]. An externalist conception would clearly rule out my above considerations on justified false beliefs. But I have the impression that a conception of that sort is not the one which is embedded in our cognitive and discursive practices. The case of the Reliable Falsity-teller shows there are justified false beliefs, and the case of the Unreliable Truth-teller shows there are unjustified true beliefs. We are disappointed in both cases because we care for both truth and justification. So an externalist conception of justification might be welcome for certain purposes and in certain contexts at least, but in my view it does not give an account of our actual practices. It might be a good revisionary conception, in sum, but it is not an explanatory one as far as we are concerned. However it is true that evidence[27] and justification standards are truth-oriented, in that we care about justification because we care about truth. Therefore it is fine to say that belief and justification aim at truth, and that truth is a norm in this sense.

Now we may think something similar about assertion. It is implausible to accept principle (A[2032?]): For any *p*, if it is true that *p*, one ought to assert that *p*; but it is not implausible to accept principle (B[2032?]): For any *p*, one ought to assert that *p* only if *p* (is true). Truth is a necessary, not a sufficient, condition of a correct assertion. In this sense, for those who follow this norm, "asserting something is asserting something that one takes to be true"[28]. Varying on this theme, one could say that when you utter that *p* you mean that you *know* that *p* (not only that you believe it)[29]. Hence truth, and possibly knowledge, is a norm of assertion. For assertoric practice aims at transmitting truth and possibly knowledge.

The problem is, however, that we have no absolute guarantee that what we believe to be true or assert to be true is in fact true. So, even if truth is

[26] Engel (2007: 32-34).

[27] On different stripes of evidentialism, cf. Engel (2007: 114-115). Going back to a past discussion with Engel (see Tuzet 2008), I wish to point out again that practical considerations play a role in the justification or in the acceptance of a belief, as Carnap remarked time ago (1936: 426): "Suppose a sentence S is given, some test-observations for it have been made, and S is confirmed by them in a certain degree. Then it is a matter of practical decision whether we will consider that degree as high enough for our acceptance of S, or as low enough for our rejection of S, or as intermediate between these so that we neither accept nor reject S until further evidence will be available."

[28] Engel (2001: 43).

[29] Cf. Engel (2007: 102 ff.).

a constitutive norm of belief and assertion, it is not a norm in the sense that having a belief or making an assertion commits us to the truth of such belief or assertion. From a constitutive point of view, when we believe that *p* we believe that *p* is true, and when we assert that *p* we assert that (we believe that) *p* is true. But the ascription of responsibility does not (directly) depend on truth, but on justification with regard to belief and on sincerity with regard to assertion[30]. As to assertion, this was showed in the case of the Truthful Liar. As to belief, think of the cases of the Reliable Falsity-teller and the Unreliable Truth-teller: what determines the fact that the latter but not the former deserves some social criticism or sanction is the fact that the latter has no justification for his (true) belief, while the former has a justification even though his belief turns out to be false. If believing would commit us to truth, the Unreliable Truth-teller would satisfy the requirement and deserve no sanction, while the Reliable Falsity-teller would deserve a sanction regardless of the fact that his (false) belief is justified.

To put it differently, truth is a correctness condition for belief and assertion[31], but is not a requirement of them in the sense that false beliefs and assertions deserve as such a form of social criticism, a kind of blame or even a stronger sanction. We can say that the norm of truth is basic, fundamental; the norm of sincerity wouldn't exist without it, since the latter prescribes to say what one believes, that is what he takes as true. We can also contend that truthfulness is the default position and the liar in general, not only the truthful one, takes advantage of that[32]. Yet, this is not to say that truth is a requirement in the absence of which beliefs and assertions deserve as such a form of social criticism or sanction.

One of the troubles in the discussion about this is the ambiguity of "norm"[33]. Among the several meanings attributed to this word (including rule, standard, criterion, constitutive condition, prescription, custom, aim, goal, etc.) it is helpful to select here the idea of norms as *prescriptions* and that of norms as *aims*. Given the foregoing argument I strongly disagree with the claim that truth is a norm of belief and assertion in the prescriptive sense that we should have true beliefs and make true assertions on pain of sanctions if our beliefs

[30] Not "directly" because it depends on some norm or criterion whether an error might be justified or not.

[31] But see Price (1998) for a more articulated picture.

[32] "Lying works because truthfulness is the default position, and the good liar takes advantage of just that fact. [...] By lying, the liar increases her own power and decreases ours" (Lynch 2004: 152).

[33] On that ambiguity see e.g. von Wright (1963: chap. 1).

and assertions turn out to be false. On the other hand I strongly agree with the claim that truth is a norm of those attitudes in the teleological sense that belief aims at truth and assertion aims at truth-transmission. (To talk law again, if there is an obligation here it is a means-obligation, not an end-obligation: we are supposed to do our best to get true beliefs and spread them, and if we don't achieve that goal we are not to blame when we have a justification, with the exception of the cases in which the truth requirement is strict and the justification we might have for our failure is irrelevant). I take this teleological sense to be the root of what Engel calls the "constitutive" normativity of beliefs and assertions. It is constitutive of belief that "belief aims at truth" (and similarly for assertion) and the fact that we consider it constitutive depends in my view on the fact that such features of aiming at truth and aiming at truth-transmission shape our concepts of belief and assertion.

Let me also note that my argument has, if I am right, some important theoretical consequences. The first is to make less ambiguous the use of "lie", distinguishing between (1) saying what is false and (2) saying what one believes to be false with the intention of misleading or deceiving[34]. The second important consequence is the upholding of different senses and kinds of normativity, taking into account in particular the distinction between norms as prescriptions and norms as aims. The third consequence it that we have the possibility to maintain a non-epistemic conception of truth, as the one which is implicit in the present account, and combine it with an inferential semantics about the vocabulary of responsibility (something I have not done here but consider not just feasible but also recommended given the fact that responsibility ascriptions depend upon social norms and practices).

To conclude. Truth is a norm of belief and assertion in the constitutive sense of what it means to have a belief and to make an assertion, and it is a norm of these attitudes in the fundamental teleological sense in which belief aims at truth and assertion aims at truth-transmission. However, it is not a norm in the sense that a false belief or assertion is to be negatively sanctioned as such: as far as our practices are concerned, the ascription of responsibility rather depends on justification and sincerity. The case of the Truthful Liar

[34] So Lynch (2004: 153): "to be sincere [...] is to be disposed to say what you believe, with the intention not to mislead." And also (155): "Other things being equal, a sincere person says what she thinks is true, on any particular subject that arises, *because she thinks it is true*. In sum, sincerity is good because true beliefs are good, but sincerity requires caring about the truth as such for its own sake. As with intellectual integrity, caring about truth is a necessary part of being sincere. Someone who couldn't care less about the truth may end up telling the truth about this or that when it suits him, but he won't be a sincere person. He'll just be honest when it pays."

and the other cases presented above support this conclusion, for one cannot explain what happens in those circumstances unless he recognizes those different commitments of belief and assertion.

5. References

CP *Collected Papers* of C.S. Peirce, 8 vols., ed. by C. Harthshorne, P. Wiess (vols. 1–6), and A. Burks (vols. 7–8), Harvard University Press, 1931–1958. For example: CP 5.189: volume 5, paragraph 189.

NEM *The New Elements of Mathematics* by Charles S. Peirce, ed. by C. Eisele, Mouton, The Hague, 1976. For example, NEM 3:187: volume 3, page 187.

Anscombe, G.E.M. (1957), *Intention*, Basil Blackwell, Oxford, 1972.

Austin, J. (1955), *How to do Things with Words*, ed. by J.O. Urmson and M. Sbisà, Harvard University Press, Cambridge (Mass.), 1975.

— (1979), *Philosophical Papers*, ed. by J.O. Urmson and G.J. Warnock, Oxford University Press, Oxford.

Brandom, R.B. (1994), *Making It Explicit. Reasoning, Representing, and Discursive Commitment*, Harvard University Press, Cambridge (Mass.) and London.

— (2000), *Articulating Reasons. An Introduction to Inferentialism*, Harvard University Press, Cambridge (Mass.) and London.

Brock, J. (1981), "Peirce and Searle on Assertion", in K.L. Ketner, J.M. Ransdell, C. Eisele, M.H. Fisch, and C.S. Hardwick (eds.), *Proceedings of the C.S. Peirce Bicentennial International Congress*, Texas Tech Press, Lubbock (Texas), pp. 281-287.

Burge, T. (1993), "Content Preservation", *The Philosophical Review*, vol. 102: 457-488.

Canale, D. and Tuzet, G. (2005), "Interpretive Scorekeeping", in P. Comanducci and R. Guastini (eds.), *Analisi e diritto 2005*, Giappichelli, Torino, pp. 81-97.

— (2006), "L'impegno assertivo", in R.M. Calcaterra (ed.), *Pragmatismo e filosofia analitica*, Quodlibet, Macerata, pp. 159-172.

— (2007), "On Legal Inferentialism. Toward a Pragmatics of Semantic Content in Legal Interpretation?", *Ratio Juris*, vol. 20: 32-44.

Carnap, R. (1936), "Testability and Meaning", *Philosophy of Science*, vol. 3: 419-471.

Engel, P. (1999), "Dispositional Belief, Assent, and Acceptance", *Dialectica*, vol. 53: 211-226.
— (ed.) (2000), *Believing and Accepting*, Kluwer, Dordrecht.
— (2001), "Is Truth a Norm?", in P. Kotako, P. Pagin, and G. Segal (eds.), *Interpreting Davidson*, CSLI Publications, Stanford, pp. 37-51.
— (2002), *Truth*, Acumen, Chesham.
— (2007), *Va savoir!*, Hermann, Paris.
Engel, P. and Rorty, R. (2005), *À quoi bon la vérité*, Grasset, Paris.
Esfeld, M. (2005), "Le pragmatisme en sémantique et en épistémologie contemporaines", *Philosophia Scientiae*, vol. 9: 31-48.
Frege, G. (1918-1919), "The Thought: A Logical Inquiry", *Mind*, vol. 65 (1956): 289-311.
Hart, H.L.A. (1949), "The Ascription of Responsibility and Rights", *Proceedings of the Aristotelian Society*, vol. 49: 171 ff.
Hart, H.L.A. and Honoré, T. (1959), *Causation in the Law*, sec. ed., Clarendon Press, Oxford, 1985.
Hilpinen, R. (2004), "On a Pragmatic Theory of Meaning and Knowledge", *Cognitio*, vol. 5: 150-167.
Holmes, O.W. (1881), *The Common Law*, Little, Brown & Company, Boston, 1923.
Lynch, M.P. (2004), *True to Life. Why Truth Matters*, The MIT Press, Cambridge.
Pape, H. (2002), "Pragmatism and the Normativity of Assertion", *Transactions of the Charles S. Peirce Society*, vol. 38: 521-542.
Papineau, D. (2003), *The Roots of Reason*, Clarendon Press, Oxford.
Price, H. (1998), "Three Norms of Assertibility, or How the Moa became Extinct", *Noûs*, vol. 32 (supplement): 241-254.
— (2003), "Truth as Convenient Friction", *The Journal of Philosophy*, vol. C: 167-190.
Searle, J.R. (1969), *Speech Acts*, Cambridge University Press, Cambridge.
— (1995), *The Construction of Social Reality*, Free Press, New York.
— (1999), *Mind, Language and Society*, Weidenfeld & Nicolson, London.
Tiercelin, C. (1993), *La pensée-signe. Études sur C.S. Peirce*, Éditions Jacqueline Chambon, Nîmes.
Tuzet, G. (2006), "Responsible for Truth? Peirce on Judgment and Assertion", *Cognitio*, vol. 7: 317-336.
— (2008), "La justification pragmatique des croyances", *Revue philosophique*, n. 133: 465-476.
von Wright, G.H. 1963, *Norm and Action. A Logical Enquiry*, Routledge & Kegan Paul, London.

Williamson, T. (1996), "Knowing and Asserting", *The Philosophical Review*, vol. 105: 489-523.

48

Moral Minimalism in the Political Realm *

STELIOS VIRVIDAKIS

There are various diverging answers to the traditional questions concerning the correct assessment of the relations between morality and politics. From Plato and Aristotle to Macchiavelli, Hobbes and Kant, philosophers have elaborated different conceptions of these relations which could be interpreted as involving a form of subordination of politics to morality, or, on the contrary, of morality to politics. Contemporary liberal thinkers are usually suspicious of any talk about the need for a "moralization" of political life, to the extent that it may hide an objectionable commitment to the promotion of some substantive ideal of the good as a collective *political goal*. However, they often admit that they do respect and sustain a kind of political morality conforming to the values of liberal democracies[1]. The political morality they are ready to defend is sometimes associated with what is characterized as a *minimalist* approach to moral issues. The aim of this paper is to cast light on some aspects and versions of this approach, the interest of which goes beyond the concerns of liberal political philosophers, and to try to cast light on the more or less "thin" moral concepts which constitute its core. Minimalism here implies a substantial restriction or attenuation of the demands of morality and not a negative

*Earlier versions of this paper were presented to different audiences in Herakleion, Tokyo, Nanjing and Athens. I am grateful to many friends and colleagues for their questions and suggestions and more particularly, to Dionyssis Anapolitanos, Georgia Apostolopoulou, Moon Such Byeon, Myrto Dragona-Monachou, Wolfgang Ertl, Anthony Hatzimoysis, Takashi Iida, Vasso Kindi, Patricia Kitcher, Philip Kitcher, Chrys Mantzavinos, Filimon Peonidis, Stathis Psillos, Pavlos Sourlas, Yannis Stephanou, and Gu Su.

[1] I am not interested in dwelling on arguments supporting political liberalism. The issues that I intend to emphasize are related to discussions probably concerning metaethics and moral philosophy more than political philosophy or politics.

stance of indifference or rejection of moral values or principles, which would amount to some form of thorough-going amoralism.[2] It will be argued that the normative model to be adopted should include both deontological and consequentialist components, that we may want to ascribe a priority to the former, and that its minimalist character will depend mostly on the construal of its central principles and on the way they are supposed to be implemented.

Let us begin with a few introductory remarks regarding the interpretation of the concepts of ethics and politics on which we intend to concentrate. In fact, there are alternative construals of the notions of the *moral* and of the *ethical*, on the one hand, and of the *political*, on the other, which one should eventually take into account. Here, we will begin our discussion by seeking a preliminary specification of their content allowing us to get a first picture of their complex relations. Thus, morality could be conceived as consisting of a set of norms for the assessment and the guidance of one's actions, insofar as their outcomes affect not only oneself but also the lives of other persons and sentient creatures. It should be noted that, although the terms "morality" and "ethics" are often taken to be coextensive, the word "ethical" is used by many philosophers to refer to broader issues regarding the good life and the values that constitute it, or are conducive to it, while the word "moral" is employed in the more narrow sense of what conforms to a set of abstract principles regulating one's conduct.[3] Morality, as we understand it in the modern era, comprises norms entailing duties and obligations, while ethics is interpreted as involving a richer set of concrete evaluative properties, includ-

[2] For a conception of forms of ethical minimalism which involve egoism or even nihilism, see Shelly Kagan, *The Limits of Morality*, Oxford: Clarendon Press, 1989, 5-6. Of course, the term "minimalism" is widely used in many areas. The notion of minimalism is usually associated with styles of modernist visual art and music, but the idea has also become fashionable in philosophy, especially in the philosophy of language and in relation with a certain conception of truth, discussed by Pascal Engel in his *Truth*, Chesham: Acumen, 2002, 65-98.

[3] This distinction between the "ethical"(ethisch) and the "moral" (moralisch) is elaborated in the writings of Jürgen Habermas. The notion of the moral is supposed to capture the proper, other-regarding goals of right action. See the discussion in Rainer Forst, "Ethik und Moral", in Lutz Wingert & Klaus Günther (Hrg.), *Die Öffentlichkeit der Vernunft und die Vernunft der Öffentlichkeit,* Frankfurt: Suhrkamp, 2001. See also Stephen Darwall, *Philosophical Ethics*, Boulder, Colorado: Westview Press, 1998 and Kieran Setiya, *Reasons without Rationalism*, Princeton: Princeton University Press, 2007, 2. More recently, Ronald Dworkin has proposed to use the terms "ethics" and "morality", to refer respectively to "the study of how to live well" and to "the study of how to treat other people". See his *Justice for Hedgehogs*, Cambridge MA: Harvard University Press, 2011, 19, 191 and *passim*.

ing character traits, that is virtues and vices. This distinction is often summarized in the contrast between "thin" and "thick" concepts, such as, on the one hand, good, just or right, and on the other, generous, honest, courageous, magnanimous, jealous, cruel, etc.[4] There is a clear analogy with the opposition put forth by Hegel, between *Moralität* -a system of principles adopted by the moral agent-, and *Sittlichkeit* -morality embodied in social institutions-, although the two distinctions are not equivalent in meaning.

In any case, we are not going to proceed by taking for granted the details of the distinction between the meanings of the terms "morality" and "ethics", to which we may return in our concluding reflections. However, even if one is occasionally willing to speak more loosely, and use the two terms interchangeably, one should not fail to take into consideration the deep going differences between modern and ancient philosophical conceptions of morality, the paradigms of which are, respectively, Kant's deontology and utilitarianism, and Aristotle's ethics.[5] It is generally agreed that Greek philosophers, who lay emphasis on the *thick* ethical dimension, clearly espouse the priority of the "good" over the "right", and that the opposite is true in the case of modern thinkers who employ mostly *thin* notions. Actually, according to the analysis that has prevailed in contemporary moral theory, the priority of the good, construed in an abstract, thin sense, is also attributed to teleological and consequentialist accounts of moral norms, put forth in the modern era, and the clear priority of the right over the good is thought to characterize only those who defend deontological views. The rightness of an action, or of a rule of action, depends on the amount of non moral value ("goodness") realized in the states of affairs brought about or aimed at by this action, or the rule to which it conforms.[6]

---

[4] See the discussion in Bernard Williams, *Ethics and the Limits of Philosophy*, Cambridge MA: Harvard University Press, 1985, 129, 140, 143-145, 162, 193, 200 and in Michael Walzer, *Thick and Thin: Moral Argument at Home and Abroad*, Notre Dame: Notre Dame University Press, 1994, *passim*.

[5] See Elizabeth Anscombe "Modern Moral Philosophy", *Philosophy* 33, reprinted in *Collected Philosophical Papers,* vol. III : *Ethics, Religion and Politics,* Minneapolis: University of Minnesota Press, 1981, Alasdair MacIntyre, *After Virtue*, 2nd ed., Notre Dame: Notre Dame University Press, 1984, Williams, *op.cit*.

[6] Charles Larmore criticizes philosophers who speak of a priority of the good over the right in modern teleological and consequentialist theories in his monograph *The Morals of Modernity*, Cambridge: Cambridge University Press, 1996, 19-40, 22. A characteristic target of his criticism is William Frankena, (*Ethics*, 2nd ed. Englewood Cliffs: Prentice Hall, 1973, 14-17) and extends to John Rawls', *A Theory of Justice*, Cambridge MA: Harvard University Press, 1971, 30-33. Although I agree with the main idea behind's Larmore's argument to the effect that such a priority is primarily found in ancient ethics, I think that there is a clear sense in which we may acknowledge

Now, politics can be taken to refer to a set of practices aiming at the effective satisfaction of needs and at the prevention and the eventual adjudication of conflicts among the members of a society, or, at a higher level, among different societies. Moreover, we often describe politics as an art, rather than as a science, of governing people and of managing the central institutions of organized, complex communities. Thus, it is generally accepted that, at least in Western liberal societies, politics is conceived as "primarily concerned with public order and safety and the protection of freedom".[7] Even in different cultures and in distant historical periods, which present us with ambitious and far reaching political ideals, deriving from religious, metaphysical and ethical accounts of social life –what Rawls describes as "comprehensive" doctrines or conceptions– the fundamental function of political activity in securing the peaceful coexistence of citizens seems to come first. Of course, such basic political activity may be complemented by much richer and more ambitious "policies" aspiring to the realization of different conceptions of the good. In fact, one may be interested in the moral appraisal, both of political activity in its more general form and of the particular policies designed and implemented by governments and political parties.

It is, I think, evident from the above, more or less uncontroversial conceptions of the moral and of the political dimensions of human life how they can and do come into conflict. On the one hand, we often acknowledge the attraction of the realm of ethical ideals, the demands of the *deon* or of the moral *telos* of actions, and of the quest for perfection. On the other, we are obliged to live in the actual world and we must be ready for compromise, limiting and adjusting our moral aspirations. In other words, we have to display "realism" in dealing with the political context in which we find ourselves. In fact, it may be inevitable that we violate some moral principle and we "dirty our hands".[8] Thus, as we have already observed, we come across philosophical models and real life circumstances in which it could be said that morality is subordinated to politics, or politics to morality.[9]

---

that rightness of action, or of rules of action, is given full priority only in modern *deontological* theories. See Stelios Virvidakis, *La robustesse du bien*, Nîmes: Éditions Jacqueline Chambon, 1996, 209n.

[7] See Edward Kainz, *Ethics in Context*, Houndmills, Basingstoke and London: Macmillan, 1988, 125.

[8] Here, I shall not dwell on particular moral dilemmas confronting politicians which are often described as instances of the problem of "dirty hands". On this, see Bernard Williams', "Politics and Moral Character", in Stuart Hampshire (ed.), *Public and Private Morality*, Cambridge: Cambridge University Press, 1978, 55-74. ere

[9] Indeed, it could be argued that Hobbes' *Leviathan* and Macchiaveli's *Il Principe*, properly

The problem is that both the political and the moral approaches to our practical concerns express important and apparently irreducible dimensions of the life of members of any human society– and we believe we should avoid subordinating either one to the other. We would like to retain the relative autonomy of both. Hence, we may appeal to the conception of moral minimalism which would do justice to our intuitions concerning the moral justification of actions, but would not aspire to determine the central goals of political activity. We want neither an excessively *moralized politic*s -leading to dangerous utopianism, or to the imposition of moralistic controls on the function of democratic institutions-, nor a clearly *politicized morality* -an attitude which amounts to loosening or jettisoning altogether ordinary moral standards and betrays skepticism about any moral constraints on political conduct, thorough-going relativism, or cynicism and nihilism. Thus, we need to determine the central components of such a minimalism and the extent to which it could help us resolve the tensions between morality and politics.

The idea of a "minimal morality" is introduced by Michael Walzer in his critical study of concepts and principles which could provide a moral framework for liberal democracies.[10] Walzer refers to the existence of minimal moral *senses* of terms such as "justice" or "truth", which seem to be easily understood by most people belonging to different societies and cultures. The possibility of a common, elementary construal of moral discourse provides a kind of "moral Esperanto" allowing them to communicate at a basic level and, more importantly, to reach a point of view from which they can also criticize the "thick" notions employed in their actual practices. However, Walzer clearly rejects the ambitious project, embraced by some liberal thinkers, to build a robust universalist ethics on such a minimal basis. He believes that at the end of the day one cannot avoid appealing to the "maximalist", "thick" and plural, partly particularist interpretations of the common moral vocabulary adopted in a variety of contexts.

Here, I shall not try to analyze Walzer's subtle arguments, which involve a careful balancing of liberal and communitarian insights and sustain the defense of his notion of "complex equality", supposedly pursued in distinct

interpreted, constitute examples of the first form of subordination, while Plato's *Republic* and Marx' *Critique of the Gotha Program*, instances of the second. For specific, real life examples of the subordination of morality to politics and of politics to morality, see Kainz, *op.cit.*, 125-130.

[10] See Michael Walzer, *op.cit.*

"spheres of justice .[11] Indeed, one may endorse his remarks concerning the importance of the "thick" dimension of ethical concepts in sociopolitical debates and still remain interested in the prospects of some form of moral minimalism, for the purpose of casting light on the main elements of an adequate *political* morality - a morality mostly appropriate for the public domain and for the regulation of political activity at different levels. Of course, it must be doubted whether such a minimalist attitude can lead to a satisfactory, comprehensive account of all aspects of public and private morality.[12]

However that may be, the discussion that follows will rely on a variety of criteria which can be invoked in order to isolate a more or less determinate essential core of moral minimalism and provide a basis for further assessment of its different construals. Thus, I am going to dwell on common *platitudes* concerning the nature of morality, on methodological issues pertaining to the construction of moral theory and to the specification of its aims, on *principles* and the norms or values constituting the moral reasons that they are supposed to express, and finally on the question of the scope and the authority of such moral reasons. I shall argue that determining the components of the minimal morality we think we need depends to an important extent on the conception of the content of normative principles that we will eventually decide to accept and to the interpretation of their role and scope of application. I will conclude my analysis by returning briefly to the issue of the suitability of the alternative minimalist options, which could be thus isolated, for moral agents in the political realm.

1. *Recognition of platitudes*

To begin with, we should take into account certain generally accepted platitudes concerning our moral concepts and the moral judgments in which they are employed. In fact, among the platitudes appealed to by Michael Smith in the course of his investigation of metaethical issues in *The Moral Problem*, we cannot ignore the role in our thinking of those regarding the *practicality*,

[11] See also Michael Walzer's earlier, *Spheres of Justice*: *A defense of Pluralism and Equality* New York: Basic Books, 1983.

[12] For a general defense of a minimalist ethics in all areas and at all levels, see Ruwen Ogien, *L'éthique aujourd'hui*, *Maximalistes et minimalistes*, Paris: Gallimard, 2007. For doubts concerning the liberal distinction between the private and the public domain, see Raymond Geuss, *Public Goods, Private Goods*, Princeton: Princeton University Press, 2001.

the *objectivity* and the *substance* or *content* of moral judgments[13]. As Smith points out, the idea that moral judgments have a *practical* significance entails that "if someone judges her f-ing to be right, then, other things being equal, she will be disposed to f"; acknowledging their objectivity amounts to maintaining that "when A says that f-ing is right and B says that f-ing is not right, then at most one of A and B is correct"; the acceptance of a more or less definite conception of their *substance* means that one is ready to endorse certain limitations on what may count as a moral requirement, as opposed to non-moral requirements, and to recognize the importance of the promotion of specific values such as human flourishing, or equal concern and respect for other persons.[14]

To be sure, it is obvious that focusing on such platitudes by itself doesn't entail the adoption of a *minimalist* or a *maximalist* approach. Most of the platitudes we mentioned could be construed either in a maximalist or in a minimalist spirit, although some seem to be more suitable for a minimalist stance. Moreover, there are intricate metaethical and normative issues that have to be settled by the moral theory or theories which will be eventually selected before one is able to uphold the commitment to a satisfactory conception of minimal morality. Nonetheless, we may determine the direction that will have to be followed in the quest for the essential core of such a conception, precisely on the basis of the platitudes that we consider to be a plausible, more or less pre-theoretical starting point.

Thus, we could perhaps agree on the following suggestions: a) We don't have to espouse a strong *internalist* position concerning the relation between moral judgments and the will or the disposition to act, in order to acknowl-

[13] See Michael Smith, *The Moral Prob*lem, Oxford: Blackwell, 1994, 39-41. Smith's list, which is not presented as exhaustive, also includes assumptions regarding the *supervenience* of the moral on the natural and the *procedure* by which we seek rational agreement on moral issues. Universalizability of moral judgments, defined as the requirement to treat exactly similar cases in the same way, could also be regarded as a platitude connected to the idea of supervenience. One may object that some of the items on Smith's list are controversial and wouldn't be accepted as platitudes by everybody, but we may provisionally agree on the importance of most of them.

[14] *Ibid*. Smith appeals to supposedly platitudinous ideas about the substance of morality elaborated by philosophers including James Dreier, Philippa Foot, Ronald Dworkin and Will Kymlicka. Here, one could rely on a historical or genealogical account. Thus, according to Philip Kitcher's evolutionary genealogy of morals, the ethical project has a primary original function or "remedying altruism failures" and a derivative one of "enhancing human possibilities", thus contributing to human flourishing. See his *The Ethical Project*, Cambridge, MA and London: Harvard University Press, 2011. The primary function seems to point to a substantial element of the minimalist core of morality.

edge their practicality.[15] All we have to accept is the *status* of moral reasons as *prima facie* reasons for action, which do not leave us indifferent. We will eventually have to examine issues concerning their authority and their strength. b) Similarly, a minimalist interpretation of the objectivity of moral judgments may allow for important limitations or qualifications of its grounds. One doesn't have to search for a realist moral *ontology*, which might provoke the objections of various philosophers and especially of certain liberals.[16] A certain form or degree of moderate relativism could be regarded as compatible with the ideal of objectivity guaranteeing the rationality of moral debates and could even go together with a weak form of moral realism.[17] Not even cognitivism is indispensable, provided one can develop a plausible quasi-realist supplement to expressivist models of moral thought, such as the intricate account elaborated by Simon Blackburn[18]. However, what may eventually prove necessary is the commitment to a conception of moral truth which would be *minimally realistic* in the sense defined by Pascal Engel, that is, our assertions in the moral domain may have to display truth-aptness for relevant debates to be possible.[19] c) Finally, concerning substance and content, it can be argued that the core we are looking for should combine deontological and consequentialist elements that cannot be neglected. Their particular form and the way in which they have to be combined depend in part on the political values informing our minimalist goals.

---

[15] On the contemporary debates between internalists and externalists, see Smith's discussion, *op.cit*., *passim*.

[16] See Dworkin's objections to the pursuit of a moral ontology and his more general objections to metaethical investigations disconnected from first-order normative inquiry in his "Objectivity and Truth: You'd Better Believe It", *Philosophy and Public Affairs* 25 (1996):87-139 and *Justice for Hedgehogs*, *op.cit*.

[17] See Stélios Virvidakis, "Stratégies de modération du réalisme moral", in Ruwen Ogien (dir.), *Le réalisme moral,* Paris, Presses Universitaires de France, 1999, 420-456.

[18] On quasi-realism, see Simon Blackburn, *Essays in Quasi-Realism*, New York: Oxford University Press, 1994 and *Ruling Passions*, Oxford: Clarendon Press, 1998.

[19] See Pascal Engel, *op.cit*. To be sure, Engel doesn't openly endorse commitment to *moral* realism or even cognitivism, although he is clearly enclined to opt for cognitivist views in most areas of philosophical inquiry. See also the positions put forth by Michael Lynch, in his *True to Life*. *Why Truth Matters*, Cambridge MA: The MIT Press, 2004. On the importance of a commitment to truth in liberal, democratic politics, see Joshua Cohen, "Truth and Public Reason", in his *Philosophy, Politics and Democracy*, Cambridge MA and London: Harvard University Press, 2009, 348-386. See also Bernard Williams, *Truth and Truthfulness*, *An Essay in Genealogy*, Princeton: Princeton University Press, 2002, 205-232 and Engel's debate with Richard Rorty concerning the normativity of truth in Richard Rorty and Pascal Engel, ed. by P. Savidan and tr. by W. McCuaig, *What's the Use of Truth?*, New York: Columbia University Press, 2007.

2. *Further methodological and epistemological observations*

Now, before proceeding to any study of the necessary components of the content of a minimal morality appropriate for political activity, we should pause to reflect upon certain methodological and epistemological issues concerning our investigation. To the extent that minimalism entails the rejection of strong foundational claims and we wish to pursue the justification of our judgments without seeking to ground them in some form of infallible basis, we will probably opt for a coherentist model of justification. Moreover, the ideal of a *reflective equilibrium* among well considered moral judgments or intuitions and general principles, which was recently elaborated by John Rawls and could be traced back to Aristotle's dialectical approach to ethics, may be regarded as the expression of the most popular and dominant coherentist conception of justification and even truth in moral philosophy.[20] Indeed, despite well known objections to coherentism, and, more particularly to the method of reflective equilibrium, minimalists, recognizing the need for moderation in their cognitive aspirations, would probably prefer it to alternative accounts of justification. Thus, political considerations which will presumably help us decide about the proper construal of moral concepts and principles shall be an integral part of the ideally coherent set of theoretical and practical beliefs constituting reasons of action.

However, it is not clear whether the minimalist model we want to arrive at entails a preference for either a *particularist* or *generalist* paradigm of moral thinking. On the one hand, the coherentism that we believe we should favour goes along with a holistic account and holism about reasons provides key premisses for some of the strongest arguments in favor of particularism. On the other, many particularists tend to espouse strong versions of moral intuitionism and realism, often associated with virtue ethics at the normative level,

[20] In fact, Michael Smith includes the coherentist conception of moral reasoning among the platitudes about morality that he takes as a basis for his investigation. See Smith, *op.cit*, 40 and above note 14. However, I am not so sure that coherentism can be considered to be the most *evident* and *natural* model of justification that most people would presuppose when engaging in moral thinking and arguing, unless, of course, one isolates the platitudes in question in the discourse of philosophers. For the compatibility of coherentism and moral realism, see, among other, David Brink, *Moral Realism and the Foundations of Ethics*, Cambridge: Cambridge University Press, 1989 and Stélios Virvidakis, *La robustesse du bien*, *op.cit.*, 113-115. In fact, I do propose to include the adoption of a coherentist approach among the strategies of moderation of moral realism, in Virvidakis, "Stratégies de modération du réalisme moral", *op.cit.*, 440-451. For a summary of recent discussions regarding the notion of reflective equilibrium see my article "Reflective Equilibium", forthcoming in James Wright (ed.), *International Encyclopedia of Behavioral and Social Sciences*, Oxford: Elsevier.

which do not seem to conform to the requirements of the minimalism explicitly or implicitly defended by most of the contemporary thinkers reflecting on the proper understanding of the relations between ethics and politics.[21] Moreover, anti-theoretically minded particularists do not recognize the need for even some general principles, which could be employed by moral and political thinkers in order to systematize our central intuitions and serve as *prima facie* guidelines for action.

At this point, we can perhaps bypass the particularist challenge and submit that the minimalist is entitled to follow traditional principled approaches. Of course, this doesn't mean that we may completely ignore the lessons to be derived from moderate versions of particularism.[22] In any case, we have to acknowledge the epistemological peculiarity of the domain of human action, first emphasized by Aristotle in his *Nicomachean Ethics*. It would be foolish to aspire to the construction of a moral theory conforming to the standards of a scientific or logical theory. Nor could we propose the adoption of strict, exceptionless principles, or the use of some form of algorithmic decision procedure for the application of moral rules in real life.[23]

3. *Which principles? - deontological and consequentialist reasons*

Philosophers who pursue the central aims of a normative theory of conduct often try to come up with an "economical" set of moral principles, which would serve as norms helping us assess and eventually guide action. In other words, they formulate very few general principles, presumably embodying the criterion -or criteria- of moral rightness and enabling us to justify the derivation and employment of more particular moral rules. Kant's *categorical imperative* and Mill's *principle of utility* are well-known traditional examples of such over-arching principles that are supposed to capture the essential

[21] For such arguments for an extreme form of particularism, see Jonathan Dancy, *Moral Reasons*, Oxford: Blackwell,1992 and *Ethics without Principles*, Oxford: Oxford University Press, 2004. For a more balanced account of the debate between particularists and generalists, see Brad Hooker & Margaret Little (eds.), *Moral Particularism*, Oxford: Clarendon Press, 2000.

[22] A plausible and sophisticated generalist stance is elaborated in Sean McKeever & Michael Ridge (eds.), *Principled Ethics. Generalism as a Regulative Ideal*, Oxford: Clarendon Press, 2006.

[23] Concerning these issues, see George Anagnostopoulos, *Aristotle on the Goals and Exactness of Ethics*, Berkeley and Los Angeles: University of California Press, 1994 and Onora O'Neill, "Abstraction, Idealization and Ideology in Ethics", in J. D. G. Evans (ed.), *Modern Philosophy and Contemporary Problems*, Cambridge: Cambridge University Press, 1989, 55-70.

core of moral thinking and acting. There are various descendants of such basic norms, elaborated by contemporary philosophers who recognize the need to take into account more than one criteria and to go beyond monistic conceptions, in attempting to combine different elements from the platitudes that we highlighted in the preceding discussion. In fact, as we also remarked after the initial recognition of the importance of a few platitudes, economy in the selection of principles or criteria, which is thought to ensure simplicity and efficiency[24], doesn't necessarily imply a commitment to minimalism.[25] What matters is not their number but their content, their interpretation and their implementation.

Indeed, we realize that we shall have to deal with the crucial issues of the authority, the scope and the force attributed to the principles in question and to the central moral reasons that they are supposed to express and codify. At this point, we can refer briefly to the nature or content of such principles and of the corresponding reasons appealed to in their articulation, which we believe should constitute an integral part of minimal morality.

Some of the recent models of normative theory that we characterized as economical, include a small number of principles, which are taken to cover both deontological and teleological or consequentialist dimensions of moral thought. Among the examples that may provide us with the more or less essential components of the minimal core we are interested in, one could mention drafts of theories offering a mixed package, made up of materials that are elsewhere encountered as basic constituents of principles of Kantian and of utilitarian inspiration.

For instance, William Frankena, in his classical introduction to moral philosophy, proposes a "mixed deontological theory of obligation", consisting of a *principle of (distributive) justice* and of a *principle of beneficence*.[26] The former is supposed to express deontological constraints without which we would fail to conform to some of our most basic intuitions about what counts as moral thinking and acting. Comparative treatment of individuals involves not only *meritarian* but also *egalitarian* criteria. In fact, the notions of impartiality and of equal concern and respect for each individual, underlying

[24] See J.P.De Marco, R.M.Fox, (eds.) *Moral Reasoning* : *A Philosophical Approach to Applied Ethics*, Fort Worth, Chicago : Holt, Rinehart and Winston, 1990, 173-174.

[25] Here, one could think of a very strict divine command theory, containing just a single central principle such as "Obey whatever moral rules have been dictated by God in the Holy Book or imposed by the authority of the Church", which it would be wrong to describe as minimalist, in so far as it would impose an austere and thorough-going regulation of all aspects of our conduct.

[26] Frankena, *op.cit.*, 45-52f .

the idea of fair treatment,[27] seem to pertain to the *form* of morality, and to be related to the concept of the *universalizability* of moral judgments.[28] The latter makes it possible to endow moral action with *content*, by specifying its goals and/or consequences as involving the realization and promotion of non-moral value. Frankena acknowledges that some regard the principle of beneficence as entailing not a real duty or obligation, but just *supererogatory*, praiseworthy, though not morally required acts. He distinguishes between its stronger negative parts, namely, "avoiding to inflict, preventing and removing evil or harm" and the weaker requirement to "do or promote good". The latter could perhaps be complemented by a version of the *principle of utility*, (conceived as an additional fifth part of the principle of beneficence), if we could manage to measure value in a reliable way that would enable us to seek the greatest balance of good over evil.[29]

One comes across similar examples of hybrid theories, proposing analogous principles, also including a general norm of respect of freedom or autonomy,[30] or, on the contrary, combining all indispensable criteria of rightness in one dense principle. It is sometimes argued, by thinkers drawing on the great religious traditions, such as the German theologian Hans Küng, that the complementary principles expressing the most basic approaches to the value of humanity can all be derived from the "Golden Rule", properly interpreted and elaborated.[31]

There are various interesting proposals for a synthesis of consequentialist and deontological considerations, such as the central norm of James Rachels' theory of "morality without hybris": "We ought to act so as to promote impartially the interests of everyone alike, except when individuals deserve particular responses as a result of their own past behavior".[32] A preference for con-

[27] It is precisely this idea which informs Rawls' conception of distributive justice. Of course, there are different ways one can construe and defend fairness, including Rawls'own thought experiment of the *original position*.

[28] See above, note 14.

[29] Frankena, *op.cit.*, 46-48. Here, it is worth comparing Frankena's suggestions, which are not presented as minimalist, to Ogien's recent defense of the two basic principles of "no harm to others" and of "equal consideration of everybody", complemented by a third principle of "moral indifference towards oneself" (justified by an alleged moral asymmetry between the relations to others and the relations to ourselves). See Ogien, *op.cit.*,153-159.

[30] See De Marco & Fox, *op.cit.*, 176-187.

[31] See, Hans Küng, *A Global Ethic for Global Politics and Economics*, tr. by John Bowden, New York and Oxford: Oxford University Press, 1998, 97-98ff. For a critical discussion of Küng's project for a "Global ethic", see Aleksi Kuokkanen, *Constructing Ethical Patterns in Times of Globalization: Hans Küng's Global Ethic Project and Beyond*, Leiden and Boston: Brill, 2012.

[32] James Rachels, *The Elements of Moral Philosophy*, 2nd ed., New York: McGraw Hill, 1993,

tractualist approaches may make us focus on the central directive of Thomas Scanlon's account of moral wrongness: "An act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behavior which no one could reasonably reject as a basis for informed, unforced general agreement";[33] or to the similar principle 'U', put forth by Habermas in his theory of communicative action: " A moral norm is valid just in case the foreseeable consequences and side effects of its general observance for the interests of each individual could be jointly accepted by all".[34] The quest for a convergence of Kantian, consequentialist and contractualist conceptions of morality could lead to the formulation of Derek Parfit's ambitious "Triple theory": "An act is wrong just when such acts are disallowed by the principles that are optimific, uniquely universally willable, and not reasonably rejectable."[35]

Now, most of the above attempts at the construction of a moral theory employing one or very few complementary principles, may qualify as versions of minimalism, regardless of the original aspirations of their authors. However, one still has to assess their real purport and the modalities of implementation of the normative guidance they seem to provide. In any case, we should note some of their salient features, which can be regarded as tokens of a minimalist orientation:

a) The clear absence of any commitment to a particular substantive and comprehensive conception of the good to be promoted, which could be detected in the principles referred to, guarantees the neutrality of the State towards diverging ideals. There is room for the peaceful coexistence of a

185. In the most recent edition of Rachels' book, the central principle to be adopted is presented as expressing "a multiple-strategies utilitarianism", since "perhaps the single moral standard is human welfare". See James Rachels and Stuart Rachels, *The Elements of Moral Philosophy*, 6th ed., New York: McGraw Hill, 2012, 196-197.

[33] See Thomas Scanlon, "Contractualism and Utilitarianism", in Amartya Sen, Bernard Williams, (eds.), *Utilitarianism and Beyond*, 103-28, 110 and his *What We Owe to Each Other*, Cambridge MA: Harvard University Press, 1998, 4.

[34] See Jürgen Habermas, *Moral Consciousness and Communicative Action*, tr. by C.Lenhardt, S.W. Nicholsen, Cambridge MA: The MIT Press, 1990, 65.

[35] See Derek Parfit, *On What Matters*, vol.I, Oxford and New York: Oxford University Press, 2011, 25, 404-419. Parfit defends a position that he describes as "Kantian consequentialism" and which he presents as pointing in the same direction as contractualism. A systematic study of previous large-scale efforts to combine deontological and consequentialist components of moral thinking should include the works of H. Sidgwick and R.M. Hare.

plurality of such ideals, freely chosen and pursued by individual members of contemporary liberal societies, who are not supposed to endorse a further *telos* of communal life.[36]

b) Many of the above formulations indicate *negative* duties which begin with the avoidance and prevention of evil rather than the promotion of the good. It is implied that it is easier to agree on what is experienced as *evil* than on what counts as good. The rightness of an action or of a rule may be harder to discern than its *wrongness*.[37] Here, one also brings to mind the liberal emphasis on negative rather than on positive rights, which would be more difficult to isolate and defend. Similarly, as we saw, according to Scanlon's contractualist approach, the proposed justification of moral rules would support directly those that it wouldn't be reasonable *to reject*, rather than those that it would be reasonable to *accept*.

c) The deontological dimension that principles of justice render prominent, as a basic component of any theory of obligation, is usually interpreted in a way which lays emphasis on more or less formal characteristics and not on substantive conceptions, involving, for example, the aspiration to a thorough-going, revolutionary restructuring of society for the promotion of equality in property or resources, as in more thoroughly and substantially egalitarian, socialist models. Equality would be construed mainly as fairness, as equal concern for the protection of rights and liberties and eventually for the promotion of interests, which may allow for differential treatment, presumably according to norms established without coercion, through what is regarded as reasonable agreement or rather as "non-rejectability".

There are still many questions concerning the correct understanding of the basic core of moral thinking revealed by such mixed accounts of rightness, attributing particular importance to deontological elements. If one wants to conform to the requirements of minimalism of the strictest and most austere

[36] Concerning the alleged neutrality of liberalism and its limits, see Nancy Rosenblum (ed.), *Liberalism and the Moral Life*, Cambridge MA: Harvard University Press, 1989. See also, Charles Larmore, "The Moral Basis of Political Liberalism", *The Journal of Philosophy* 96 (1999): 599-625

[37] See De Marco & Fox, *op.cit.*, 176-187. ("Do no harm", "Do not be unfair", "Do not violate another's freedom") Such emphasis on negative formulations can be compared to the differentiation between perfect and imperfect duties, specified through the employment of Kant's categorical imperative.

kind, one should perhaps follow Stuart Hampshire in interpreting justice as a purely *procedural* notion which doesn't go beyond a "minimum fairness in established procedures of settling conflicts". As Hamsphire puts it, "decent fairness ... is a value independent of any conception of the good... rooted in the fact that human beings have to some degree the habit of balancing contrary arguments and of drawing conclusions from them. Minimal justice is the elaborate application of this habit to interpersonal relations, entailing fair rules of procedure."[38] Moreover, consequentialist, and more particularly utilitarian norms, appeal to which is to a certain extent unavoidable in the pursuit of political goals, should be employed only in ways compatible with respect for fundamental deontological constraints imposed by basic principles of procedural justice and of negative freedom and by their corollaries. Hence, deontological reasons imposing the protection of rights and liberties would retain their priority except in cases of a serious threat to the survival or well-being of a society.

4. *Limits of the authority of moral reasons*

Our inquiry into the characteristics of the minimalist model constituting the background of a liberal political morality cannot be completed without a brief assessment of its scope and strength. It must be ensured that the construal and application of principles such as the ones that we have just examined doesn't betray their minimalist intent. Indeed, I want to highlight the fact that minimalism requires significant limitations of the authority of moral claims at different levels and in different senses, Here, it should be asked whether and to what extent we ought to regard moral reasons as *pervasive*, *overriding* and *stringent*.[39]

Now, pervasiveness implies unlimited scope, in the sense that "no voluntary human action is in principle resistant to moral assessment"; overridingness means that moral claims always "defeat" the authority of other reasons, so that "it is never rational knowingly to do what morality forbids";[40] while

[38] See Stuart Hampshire, *Innocence and Experience*, Cambridge MA: Harvard University Press, 1989, 169.

[39] In what follows, I draw upon the distinctions and the penetrating analysis of Samuel Scheffler in his *Human Morality*, New York: Oxford University Press, 1992, although my construal and final positions may differ from his at various points.

[40] For an interesting, summary presentation of alternative assessments of the overridingness of moral reasons in the history of philosophy, see Thomas Nagel, *The View from Nowhere*, Oxford: Oxford and New York, 1986. According to Nagel's analysis, in ancient thought any real conflict

stringency expresses the "demanding-ness" of these claims "in whatever domain they apply", and regardless of whether they may be rationally defined by appealing to other, equally, or more important considerations."[41]

Therefore, when we wonder about the correct implementation of moral principles in the political domain, we understand why an appropriate interpretation of their content seems to involve a certain degree of moderation or weakening in most of these senses. We have already hinted at the fact that principles focusing on the goals or consequences of actions should not be taken to entail a missionary commitment to the attainment of the greatest balance of good over evil for the greatest number. Such a missionary mentality would be clearly regarded as supererogatory and could not be imposed as a political ideal at the expense of the free pursuit of a variety of conceptions of goodness by different individuals.[42] Liberal, democratic societies embrace less demanding forms of utilitarian or non-maximizing teleological principles of beneficence, which are taken to dictate *prima facie* duties to avoid evil or harm and promote good, and are not supposed to contravene deontological obligations to respect rights and liberties. Only the minimal components of such principles of beneficence, and especially of principles of justice and au-

between the moral life and the good life would be impossible, because of the internal relations of the two notions. In the modern era, the tension between them is obvious in the thought of philosophers such as Kant and Nietzsche, who, in cases of conflict, respectively defend the overridingness of the moral life and that of the good life, while one may hold a middle position, arguing that "neither the moral life nor the good life consistently overrides the other". Nagel expresses his inclination to accept a Kantian view, although he recognizes the force of the "middle position". (195-200). For a novel approach to these issues one should turn to Dworkin's unitary account of value in *Justice for Hedgehogs*, which introduces an interesting distinction between an ethically and morally laden ideal of "living well" and a neutral conception of "having a good life". Dworkin sees a continuity between the *ethical* and the *moral* and acknowledges ethical responsibilities to oneself. His original and controversial proposals which try to integrate Aristotelian notions and a strong Kantian emphasis on the value of human dignity indicate a clear distance from the aspirations of moral minimalism, although he doesn't renounce his commitment to political liberalism. See Dworkin, *op.cit.*, 191-210 and *Religion without God*, Cambridge MA: Harvard University Press, 2013. Here, it could be noted that in our discussion we have excluded virtue ethics from all the versions of minimalism the main aspects of which we have attempted to sketch, in so far as the appeal to notions of virtue is usually related to substantive ethical considerations.

[41] Scheffler, *op.cit.* , 25-26ff.

[42] For such strong and maximalist views, which one would be justified in regarding as betraying a "missionary" attitude associated with some forms of consequentialism and more particularly utilitarianism, see Kagan, *op.cit.*, and the works of Peter Singer. See, a.o. Peter Singer, *One World: The Ethics of Globalization*, 2nd ed., New Haven and London: Yale University Press, 2002. For a critical perspective on these and related issues, see Liam B. Murphy, *Moral Demands in Nonideal Theory*, Oxford and New York: Oxford University Press, 2000.

tonomy, interpreted as primarily negative and proceduralist notions, seem to entail stringent and overriding moral reasons, insofar as such reasons sustain the normative framework of a well-ordered communal life. The core of this normative framework is constituted by the idea of a mutual acknowledgment, by citizens functioning as moral agents, of one or more basic principles dictating "what they owe to each other", to use the apt expression proposed by Scanlon. Hence, only rules derived through universal application of the latter principle(s), recognized as holding everywhere and for everybody, regardless of social conditions prevailing in particular cultural and historical contexts, would embody considerations which could override all other prudential, political or aesthetic reasons. Such considerations could also be described as *prima facie* pervasive since they couldn't be ignored by rational human subjects acting voluntarily in any domain.[43] Still, they are not fully and thoroughly pervasive, always overriding or stringent in the sense that they would entail duties or obligations to oneself and would affect personal values endorsed privately by each individual.[44]

Of course, this minimalist construal of moral notions doesn't suffice if we are seeking an understanding of the whole of morality, and more generally, ethics, to return to the distinction we highlighted at the beginning of our discussion.[45] There are various, more or less widely shared values, virtues and conceptions of the good life, which may be of paramount normative significance for individuals and for social groups, whose existence would be seriously impoverished without them. However, they do not and should not concern political activity, at least directly and in ways that would threaten its contractual democratic framework and its implicit norms. In fact, it is by acknowledging the latter that one realizes why it would be wrong to present political life as totally dissociated from morality. One should not think that politicians, or any citizens of a liberal society, who stress the need for the respect of fundamental principles of beneficence, justice and autonomy, and

[43] Scanlon, *What We Owe to Each Other*, *op.cit*., 348-349.

[44] See Ogien, *op.cit*., 33-57. See also Thomas Scanlon's forthcoming John Locke lectures, *Being Realistic about Reasons*, Oxford and New York: Oxford University Press, 2014, 105-123. Here, I would like to thank Professor Scanlon for permission to use the text of his unpublished Locke lectures and for our discussions of relevant metaethical issues.

[45] Here, one may be persuaded by Dworkin's arguments to which we have already referred regarding the organic connections, if not the continuity or unity of ethical and moral values. See above, note 41.

for the promotion of the relevant values and virtues of truthfulness, honesty and fairness, are always either disingenuous or necessarily guilty of pernicious moralism.[46] Something like the minimal morality that we have tried to describe should not be absent from the political realm, provided, of course, politics aims at handling problems of social interaction in a decent way.

[46] On the important issue of moralism, see, a.o. Julia Driver, "Moralism", *Journal of Applied Philosophy* 22/2 (2005): 137-152, Robin Fullinwider, "On Moralism", *Journal of Applied Philosophy* 22 (2005): 105-120 and Craig Taylor, *Moralism: A Study of a Vice*, Durham: Acumen, 2012. For an interesting discussion of varieties of moralistic attitudes in international relations, see C.A.J. Coady, "The Moral Reality in Realism", *Journal of Applied Philosophy* 22 (2005): 121-136. At this point, one should go on to concentrate on the problems of *legal moralism* and of *moral paternalism* and on issues regarding the eventual need to invoke moral norms for the regulation of markets, pertinent to the further study of the relations among morality, the economy, law and politics, but such a task would require another paper.

49

Modes et modalités dans le système de droit naturel de Samuel Pufendorf *

DANIEL SCHULTHESS

1. Introduction

Samuel Pufendorf (1632-1694) est pris en considération, de nos jours, surtout par les historiens du droit et de la philosophie politique[1]. A ma connaissance il n'a pas été discuté dans la perspective logique (en un sens large) qui m'intéresse ici. De fait, la polémique menée par Leibniz contre Pufendorf – dans la lignée d'autres polémiques anticartésiennes – a sans doute affaibli la stature de ce dernier aux yeux des philosophes pourvus d'un intérêt pour les

*Une version antérieure de cet exposé a été présentée au Colloque « Normes, Vertu et Autonomie dans la Philosophie des XVII^e et XVIII^e Siècles » organisé par Richard Glauser à l'Université de Neuchâtel les 27-29 octobre 2011. Je remercie les participants au Colloque pour leurs commentaires avisés.

[1] Voir l'excellente esquisse de Simone Goyard-Fabre, « Pufendorf et Grotius : Deux faux amis, ou : La bifurcation philosophique des théories du droit naturel », dans Vanda Fiorillo (dir.), *Samuel Pufendorf Filosofo del Diritto et della Politica, Atti del Convegno Internazionale di Milano, 11-12 novembre 1994, Naples*, La Città del Sole, 1996, p. 171-207.

questions logiques[2]. Cependant, Pufendorf mérite de redevenir un auteur de référence même pour ces derniers, comme il apparaîtra dans ce qui suit.

Pour commencer, voici quelques mots d'explication sur l'intitulé de mon article. Je parle de « modes » dans le sens de l'opposition catégorielle de style cartésien entre la substance et le mode (j'ajouterais « ontiques » pour le contraste avec « modalités déontiques »[3]). Les modes se rangent chez Pufendorf sous deux grandes rubriques, le physique et le moral. Il ne s'agit donc pas de l'opposition des modes de la *res extensa* et des modes de la *res cogitans*, mais d'une tout autre opposition. Les modes moraux (Pufendorf parle des *entia moralia*, des êtres moraux) sont – pour faire court – des modes de la chose composée (tant étendue que pensante), mais des modes distincts des modes spécifiques de l'étendue et de la pensée[4]. La question des modes moraux attachés aux choses exclusivement étendues (donnés par exemple avec la notion de propriété au sens juridique), fait l'objet chez Pufendorf d'un traitement distinctif[5].

A parler strictement, les *entia moralia* sont des modes et non des substances (ils ont besoin de substances naturelles pour se greffer sur elles). Par extension, cependant, l'expression *ens morale* vient à s'appliquer aussi aux substances, quand celles-ci soutiennent un mode moral : comme quand Pufendorf écrit : « Les êtres moraux qui sont conçus par analogie avec les substances, sont appelés les *personnes* morales, c'est-à-dire les hommes pris individuellement, ou bien liés en un système par un lien moral [...][6]. » On note ici un usage de l'expression *persona moralis* qui inclut le cas où notre usage contemporain demande « personne physique ».

Je passe à l'autre mot, plus brièvement : je parle de « modalités déontiques » pour indiquer les modalités de l'obligatoire, de l'interdit et du permis. Pu-

[2] Cf. « Monita quaedam ad Samuelis Puffendorfii Principia » (1706), traduits dans : *Leibniz : Le droit de la raison*, édition de René Sève, Paris, Vrin, 1994, p. 19-35 ; cf. l'introduction de R. Sève, p. 11-19, qui prolonge les analyses magistrales de cet auteur dans *Leibniz et l'Ecole moderne du droit naturel*, Paris, PUF, 1989.

[3] Cf. Jean-Louis Gardies, *Essai sur la logique des modalités*, Paris, PUF, 1979.

[4] Samuel Pufendorf, *De Jure Naturae et Gentium*, éd. Frank Böhling, Berlin, Akademie Verlag, 1998 (=*Gesammelte Werke*, Bd. 4), Livre I ; *Le droit de la nature et des gens*, trad. Jean Barbeyrac, 2e éd., Amsterdam, Pierre De Coup, 1712. Titre abrégé ci-après en DJNG. La traduction des passages est de l'auteur de cet article.

[5] D'ailleurs, pour Pufendorf, « être la propriété de quelqu'un » n'est pas à proprement parler un mode, mais une dénomination extrinsèque (DJNG I.i.16, p. 22). Ce contraste est important, mais je ne puis le discuter ici.

[6] « Entia moralia, quae ad analogiam substantiarum concipiuntur, dicuntur *personae* morales, quae sunt homines singuli, aut per vinculum morale in unum systema connexi (...). » DJNG I.i.12, p. 19, l. 5-6.

fendorf n'utilise pas ce mot, mais bel et bien une notion qui lui correspond, exprimée par *qualitas moralis operativa*[7].

Ce qui m'intéresse, c'est donc le rapport entre modes moraux et modalités déontiques ; sous l'angle de l'absence d'un côté : tant qu'il n'y a pas *d'entia moralia*, il n'y a pas de modalités déontiques ; sous l'angle de la présence de l'autre : une fois qu'il y a des *entia moralia*, il y a des modalités déontiques. Pufendorf se montre rigoureux sur cette coïncidence.

2. Le nouveau concept de nature

Par son approche ontologique de la thématique juridique, Pufendorf se situe de façon originale dans l'histoire du droit naturel : il est l'auteur qui prend en compte de la façon la plus cohérente le nouveau concept de nature, le concept mécaniciste, exhaustivement décrit par des quantités[8]. Face à ce nouveau concept de nature[9], qui ne soutient plus comme l'ancien concept de nature la mise en place de modalités déontiques, Pufendorf a dû innover. Cela se marque dans la structure même du grand ouvrage *De Jure Naturae et Gentium* (1e éd. 1672, 2e éd. 1684), puisque le premier chapitre du premier livre, « Praecognita », porte le titre « De Origine et Varietate Entium Moralium » (« L'origine et la variété des êtres moraux »), et que la suite met longuement en place le concept de *ens morale*. Pufendorf explique dans l'introduction du premier chapitre du premier livre (§1) qu'une nouvelle discipline doit être constituée, qui traiterait des *entia moralia*, à côté de la discipline déjà reconnue qui traite des choses naturelles (*res naturales*). Il met en place un plan systématique certes simple, mais très représentatif de l'ambition qui l'anime : parmi les disciplines, il en est une qui traite des *res naturales* (sans doute faut-il supposer que la *res cogitans* en fait partie), l'autre qui traite des *entia moralia*. L'une

[7] DJNG I.i.19, p. 23, l. 37.

[8] Sa démarche peut être comparée, de nos jours, avec celle que mène John Searle, tout particulièrement dans *The Construction of Social Reality*, New York, Free Press, 1995 ; *La construction de la réalité sociale*, trad. Claudine Tiercelin, Paris, Gallimard, coll. « NRF Essais », 1998. Pour une brève comparaison de ces auteurs, cf. mon article « L'ontologie du monde social chez Samuel Pufendorf et John R. Searle », dans A. Chenoufi, T. Cherif, S. Mosbah (éds), *L'Universel et le devenir de l'humain, Actes du XXXIIe Congrès de l'Association des Sociétés de philosophie de langue française (ASPLF)*, Tunis-Carthage, 28-1er septembre 2008, Tunis, Association Tunisienne des Etudes Philosophiques, 2010, p. 171-175.

[9] Cf. l'ouvrage classique de Robert Lenoble, *Histoire de l'idée de nature*, Paris, Albin Michel, 1990.

de ces disciplines est constituée, l'autre reste à construire[10], ce à quoi son traité va se consacrer : ce sera le droit naturel[11].

Dans ce schéma, les modalités déontiques se rattacheront aux seuls *entia moralia*. En effet, en, Pufendorf donne la définition suivante des *entia moralia* :

> « [Les êtres moraux] sont certains modes surajoutés par des êtres intelligents ou à des choses (*res*) ou à des mouvements physiques (*motus*), en vue principalement de diriger et de tempérer la liberté des actes volontaires de l'homme, ainsi que d'apporter quelque ordre et beauté à la vie humaine »[12].

Les fonctions de diriger et de tempérer qui reviennent aux *entia moralia* passent par des obligations, des interdictions, des permissions.

3. La thématique d'arrière-fond

La mise en place des *entia moralia* répond à une préoccupation ontologique, dont les aspects logiques ne sont pas tout de suite manifestes. Cependant, ma thématique d'arrière-fond, au moment d'aborder ces matériaux, est celui d'une interrogation de caractère logique au sens large ; c'est celle qui porte sur la coupure logique entre ce qui est de fait et ce qui est de droit ; et sur la manière d'envisager le domaine de ce qui est de droit au regard de cette coupure. Le philosophe qui a pris position de la façon la plus éclatante sur ce sujet est David Hume, au début du Livre III du *Traité de la nature humaine* – livre III (1740) qui est à sa façon aussi un traité de droit naturel[13] –, construit précisément sur une revendication de la coupure entre fait et droit, entre les propositions portant un *is* et les propositions portant un *ought*, comme le dit

[10] Sur les ambitions scientifiques de ce programme, cf. Alfred Dufour, « Le paradigme scientifique dans la pensée juridique moderne », dans *L'Histoire du droit entre philosophie et histoire des idées*, Bruxelles, Bruylant, Zurich, Schulthess, 2003, p. 472-492.

[11] Adam Smith présente des remarques incisives sur ce programme dans sa *Theory of Moral Sentiments* (1e édition 1759), ed. D.D. Raphael et A.L. Macfie, Oxford, Clarendon, 1976 ; *Théorie des sentiments moraux*, trad. M. Biziou, C. Gautier et J.-F. Pradeau, éd. révisée, Paris, PUF, Collection Quadrige, 2011, VII.iv, spécialement depuis le §7.

[12] « [Entia moralia sunt] modi quidam, rebus aut motibus physicis superadditi ab entibus intelligentibus, ad dirigendam potissimum & temperendam libertatem actuum hominis volontariorum, & ad ordinem aliquem ac decorem vitae humanae conciliandum », DJNG I.i.3, p. 14, l. 19-22.

[13] Cf. par exemple *Treatise* III.ii.2 : « L'origine de la justice et de la propriété ».

Hume (c'est-à-dire la « loi de Hume », ou la dénonciation du sophisme naturaliste) ; construit aussi sur la manière d'envisager le domaine de ce qui est de droit au regard de cette coupure revendiquée[14].

4. Un droit naturel compatible avec la notion d'imposition

Je présente une première trace du problème, une dénégation fort remarquable.

Le droit naturel de Pufendorf, ce sera un droit *naturel* quelque peu paradoxal, puisqu'il se revendique comme droit naturel (avec l'opposition standard du droit naturel et du droit historiquement réalisé), mais doit cependant intégrer la notion de l'imposition, qui en principe jusque là caractérisait le droit positif-conventionnel :

> « Pour que la discipline du droit naturel puisse satisfaire la mesure d'une vraie science, il n'est pas du tout nécessaire que nous disions avec certains [N.d.A : à savoir Grotius et les siens], que certaines choses soient honnêtes ou déshonnêtes sans aucune imposition : et que ce soient celles-là [N.d.A : certaines choses honnêtes ou déshonnêtes sans aucune imposition] qui fassent l'objet du droit naturel et perpétuel. »[15]

L'imposition dont il est question ici reste compatible avec la notion de droit naturel :

> « Nous parlons d'un état de nature de l'homme, non du fait qu'il découlerait des principes physiques de l'essence de l'homme hors de toute imposition ; mais parce qu'il procède de l'imposition divine, qui accompagne l'homme depuis sa naissance. »[16]

Pufendorf opère donc une redistribution des positions en matière de droit naturel, que je résume par le tableau suivant :

---

[14] Sur la compatibilité chez Hume de cette coupure avec l'établissement d'une éthique normative, voir notre étude « Hume and Searle : the 'Is-Ought' Gap versus Speech Act Theory », dans *The Institute for Advanced Studies in the Humanities (IASHE, The University of Edinburgh) Occasional Papers*, Nr. 17, 2011, 26 p. (Series "Dialogues with Hume"). Sur la problématique générale de la « loi de Hume », voir spécialement Jean-Louis Gardies, *L'erreur de Hume*, Paris, PUF, 1987.

[15] « [U]t disciplina juris naturae (…) verae scientiae mensuram implere possit, hautquidquam necessarium arbitramur cum nonnullis statuere, quaedam per se citra omnem impositionem esse honesta aut turpia : & haec facere objectum juris naturalis et perpetui, » DJNG I.ii.6, p. 29, l. 34-37.

[16] « Naturalem hominis statum vocamus, non quod is citra omnem impositionem ex physicis principiis essentiae humanae fluat ; sed quod est impositione Numinis, non ex arbitrio hominum, hominem statim ab ipsa nativitate comitetur. » DJNG I.vii, p. 16, l. 12-15.

	Valide sans qu'il y ait imposition ?	Valide du fait d'une imposition ?
Droit naturel	Oui : Grotius, mais aussi Hobbes (pour eux, c'est une question de définition) ; Non : Pufendorf	Non : Grotius, Hobbes ; Oui : Pufendorf (une imposition divine, par opposition à une imposition humaine)
Droit positif	Non, ce serait absurde : réponse générale	Oui : réponse générale (mais pour Pufendorf ce n'est pas ce qui définit le droit positif)

5. Des modes aux modalités

Comment Pufendorf fait-il face, au moment de construire le droit naturel, à l'exigence de coupure entre fait et droit ? Comment passe-t-il de la substance physique aux déterminations déontiques (comme l'obligation) ? Cela se fait au travers des étapes suivantes :

- la substance – pour autant quelle s'y prête – se voit imposer un mode moral qui est l'état (*status*)
- l'état en tant que tel fait connaître des règles de conduites (adaptées aux différentes conditions telles que celle de père et de fils, d'époux et d'épouse, etc.)
- ces règles apportent ensuite aux actions possibles les modalités déontiques, celles-ci étant des modes moraux spécifiques, appropriés aux actions, les qualités morales opératives, telles que *potestas*, *jus* et *obligatio*[17].

La nature humaine prise au physique possède des caractéristiques (amour de soi, *amor sui*, faiblesse, *imbecillitas*, pauvreté, *naturalis indigentia*, dérèglement, *pravitas animi*[18]) qui rappellent Hobbes plutôt que Grotius. A ce titre, la nature humaine comporte en creux une exigence de sociabilité, *socialitas*, dès lors que la vie en société est la condition d'une existence humaine. Du coup, la loi possible la plus basique – résumant en quelque sorte les autres – est la loi de la socialité. Celle-ci va être déployée dans le reste de l'ouvrage.

[17] DJNG I.i.19-21.
[18] Cf. DJNG II.i.

6. Le traitement du *status*

Pufendorf veut corréler les *res extensae* et les personnes, et créer, à partir de la notion d'espace (*spatium*) un parallèle avec quelque chose qu'il appelle le *status* (un déterminable), pour ainsi dire un espace pour les personnes :

> « De même en effet que les substances physiques supposent un espace, dans lequel elles puissent poser l'existence naturelle qui est la leur, et déployer leurs mouvements physiques ; ainsi de manière analogue on dit et on comprend que les personnes morales se trouvent dans un état (*status*), qui de même est présupposé par elles ou est sous-jacent à elles, de sorte qu'en lui elles produisent leurs actions et leurs effets. Ainsi la nature de l'état peut correctement être décrite en disant qu'il est un être moral suppositif, du fait de l'analogie avec l'espace. »[19]

Nous obtenons donc le tableau suivant :

	Etres physiques	Etres moraux
Substance	Res extensa	Res, c'est-à-dire personnes morales (DJNG I.i, les §§ 12-15)
Support de modes	Spatium	Status (les §§ 6-11)
Modes *bona fide*	Quantités, qualités	Modes moraux (les §§ 17-22)

7. Le thème de l'indifférence de l'ordre physique en tant que tel

Il existe un régime de description des « animaux » – c'est la description proprement physique – qui rend cette description tout à fait dénuée d'implications morales, déontiques, etc. Soutenir que ce régime de description est aussi

[19] « Verum quemadmodum substantiae physicae velut supponunt spatium, in quo suam quam habent naturalem existentiam ponunt &motus suos physicos exercent : ita ad harum analogiam etiam potissimum personae morales dicuntur, &intelliguntur esse in *Statu* ; qui itidem iis velut supponitur aut substernitur, ut in eo actiones atque effectus exserant. Inde natura status non incongrue exprimi potest, quod sit ens morale suppositivum, ob analogiam, quam habet cum spatio. », DJNG I.i.6, p. 15, l. 31-36.

le régime de description qui correspond à la physique, c'est l'occasion pour Pufendorf de se séparer d'autres notions de la nature (aristotélicienne, stoïcienne) :

> « Ainsi tous les mouvements et toutes les actions de l'homme, toute loi ou divine ou humaine étant enlevée, sont indifférentes ; et ces choses qui sont dites naturellement honnêtes ou déshonnêtes, ce sont celles que la condition de nature, que Dieu a librement attribuée à l'homme, requiert le plus d'accomplir ou d'omettre. »[20]

8. L'autonomie

Pufendorf discute ce thème de « l'indifférence »[21] non seulement en DJNG I.ii, mais aussi dans *Eris scandica*, une collection de réponses apportées à différents contradicteurs après la publication de son traité, collection publiée en 1686[22]. *Eris scandica* met ce thème en relation avec le thème de l'autonomie. Cela se fait de la manière suivante. D'abord Pufendorf retient le partage du physique et du moral :

> « Il apparaît à l'investigateur attentif que dans l'action morale, il y a quelque chose de physique, et quelque chose de moral, qui procède de l'imposition, de la détermination et de la définition due à des êtres intelligents. »[23]

Une fois cette imposition effectuée, les implications morales sont données :

> « Pour qu'il y ait une action morale pleine, il faut qu'au mouvement physique vienne se surajouter (*supervenit et accedit*) une cer-

[20] « Sic ut revera omnes motus & actiones hominis, remota omni lege tam divina quam humana, sint indifferentes ; earum autem aliquae ideo tantum naturaliter honestae aut turpes dicantur, quod eas fieri aut omitti quam maxime requirat conditio naturae, quam Creator homini libere attribuit. » DJNG I.ii.6, p. 30, l. 8-10.

[21] Ce thème – largement développé dans *Eris Scandica*, cf. les notes suivantes – présente une affinité avec celui du caractère « incolore » des mouvements corporels, coutumier dans la philosophie analytique de l'action. Cf. Marc Neuberg, « L'intention définit-elle l'action ? », dans R. Glauser (dir.), *Philosophie de l'action*, *Revue de théologie et de philosophie* 124 (1992), p. 217.

[22] Samuel Pufendorf, *Eris Scandica, und andere polemische Schriften über das Naturrecht*, éd. Fiammetta Palladini, Berlin, Akademie Verlag, 2002 (=*Gesammelte Werke*, Bd. 5). Titre abrégé ci-après en ES. La traduction des passages est de l'auteur de cet article.

[23] « [A]ctionem moralem penitius introspicienti patet, aliquid in illis esse naturale seu physicum, aliud morale, quod ab impositione, determinatione, et definitione entium intelligentium promanat. » (ES, p. 165, l. 32-34)

> taine qualité ou affection morale, par l'adjonction de laquelle l'action est dite bonne ou mauvaise dans l'ordre moral. »[24]

Les implications morales quant à elles viennent de la confrontation avec une loi :

> « Cette affection résulte de la congruence ou de la disconvenance de cette action avec une norme morale ou une loi. »[25]

Cette confrontation n'est donnée que par la médiation de l'intelligence et de la volonté de l'agent :

> « Et [...] cette convenance ou disconvenance ne s'établit pas sans réflexion et par hasard, mais par une libre application et direction de l'homme ; d'un homme qui sait et veut que son action se conforme (*adplicuerit*) à la loi, ou veut qu'elle ne s'y conforme pas. »[26]

Dès lors un partage se fait entre la part de l'action qui relève de la nature et la part qui la dépasse :

> « Cette considération distincte fait apparaître, ce qu'il y a dans l'action humaine, à quoi Dieu concourt aussi comme auteur et conservateur de la nature, et ce qui est propre à l'homme. »[27]

Le thème du *concursus dei* fait ainsi apparaître, par un partage assez sommaire, une région propre de l'homme, une région spécifiquement morale (qui n'est pas littéralement une autonomie puisque l'homme ne se donne pas à lui-même sa loi) :

---

[24] « [U]t plena actio moralis fiat, huic motui physico supervenit et accedit qualitas quaepiam seu affectio moralis, juxta quam actio illa bona et mala in genere morum dicitur[.] » ES, p. 165, l. 38-40.

[25] « [Q]uae affectio resultat ex congruentia aut disconvenientia ejus actionis cum norma morali seu lege [.] » ES, p. 165, l. 40-41.

[26] « [E]t quidem ut ea convenientia aut disconvenientia non temere (sans réflexion) et quasi fortuito, sed per ultroneam adplicationem et directionem hominis provenerit, i.e. ut homo sciens volensque actionem suam ad normam istam adplicuerit, aut adplicare noluerit. » ES, p. 165, 41-p. 166, l. 2.

[27] « Ex hac autem distincta consideratione adparet, quid sit illud in actione humana, ad quod Deus quoque tanquam autor et conservator naturae concurrit, et quid hominis sit proprium. » ES, p. 166, l. 5-7.

> « Ce qui est propre et particulier à l'homme, c'est ce qui est moral dans l'action ; c'est-à-dire d'appliquer les facultés et les forces naturelles, données et soutenues par Dieu, et de les diriger ou non vers la norme ; et cette application et direction, dans la mesure ou elle se fait bien ou non, rend l'action bonne ou mauvaise. »[28]

9. Le débat de Pufendorf avec Hobbes

Le débat de Pufendorf avec Hobbes sur l'état de nature est alors particulièrement éclairant pour notre sujet : les modes ontique dans leur rapport avec les modalités déontiques. (Mais il faut noter évidemment que Pufendorf ne l'aborde pas avec l'instrument logique de Hume.) Je décrirai l'opposition de Pufendorf à Hobbes par deux volets concomitants :

- Pour Pufendorf, Hobbes a le tort de vouloir définir l'état de nature à partir des seuls « modes physiques » (compris à la Pufendorf), et nullement des « modes moraux » (compris à la Pufendorf) ; Hobbes se trompe donc sur la définition de l'état de nature.
- Pour Pufendorf, Hobbes a le tort de prêter à l'état de nature tel qu'il le conçoit des conséquences déontiques ; la bonne identification des bases du droit naturel, qui procède d'une imposition, fait défaut à Hobbes.

Pour structurer la discussion, je rappelle quelques éléments du Chap. XIII du *Léviathan*[29]. L'état de nature hobbesien se construit à partir d'une situation naturelle ou les hommes sont égaux en force. A partir de là, différents « domaines de lutte » se définissent :

- Chacun désire diverses choses, et entre en conflit avec les autres à leur sujet (compétition).
- Chacun veut se défendre contre les incursions d'autrui (et ce qui sert à se défendre est *permis*). Cette défense est menée aussi préventivement.
- Chacun veut être estimé des autres, et veut se venger du refus de l'estime

[28] « Hominis autem proprie et peculiare est, quod in actione ista est morale ; nempe facultates et vires naturales, à Deo concessas et sustentatas applicare, et ad normam dirigere vel non dirigere ; quae adplicatio et directio prout recte aut secus se habet, actionem bonam aut malam reddit. » ES, p. 166, l. 10-13.

[29] Thomas Hobbes, *Léviathan*, trad. F. Tricaud et M. Pécharman, Paris, Vrin et Dalloz, 2004.

De ces trois ressorts procède une guerre de tous contre tous.

Toutefois, les ressorts de cette guerre de tous contre tous ne sont pas seulement descriptifs de la nature humaine. Chez Hobbes ils font l'objet d'une interprétation déontique. Cela ressort tout particulièrement au chap. XIV du *Léviathan* : les deux premiers ressorts ont tout particulièrement une interprétation déontique : il est *permis* de se défendre ; et par ailleurs, assure Hobbes : « dans la condition naturelle des hommes il existe un droit de tous sur toutes choses ».[30]

10. La critique décisive

Aussi faut-il dire que chez Hobbes, la guerre de tous contre tous ne suit pas seulement de données factuelles concernant les hommes ; elle suit aussi de données déontiques. Pufendorf s'inscrit en faux contre cette construction : « Ce droit Hobbesien sur toutes choses n'existe pas. »[31] Plus explicitement :

> « Ainsi nous ne reconnaissons pas non plus la conséquence que tire Hobbes : ce droit (=ce droit sur toutes choses) que l'homme aurait, ou pourrait avoir, et qui aurait quelque effet par rapport aux autres hommes. En effet nous montrons ci-dessus qu'en aucune façon cet état Hobbesien n'est naturel à l'homme, destiné qu'il est à la vie sociale. »[32]

La critique porte sur le moment déontique :

> « Ce n'est pas que n'importe quelle faculté naturelle de faire quelque chose soit proprement un droit, mais celle-là seulement, qui enveloppe un certain effet moral auprès des autres, qui sont de même nature que moi. »[33]

La pratique de l'animal non humain est prise à témoin par Pufendorf :

---

[30] Op. cit., p. 111.

[31] « Jus in omnia Hobbesianum non datur », DJNG, III.v.3 p. 263. A noter que Pufendorf dirige une critique similaire contre Spinoza, DJNG II.ii.3.

[32] « Enimvero quemadmodum supra ostendimus (=nous montrons), statum illum Hobbesianum hautquidquam esse naturalem homini, ad socialem vitam destinato ; ita neque consequens istius agnoscimus, ejusmodi jus, quod homo habuerit, aut habere potuit ad omnia, &quidem, quod effectum aliquem in ordine ad alios homines obtineat. » DJNG, III.v.3, p. 263, l. 18-21.

[33] « Non quamlibet facultatem naturalem aliquid agendi propie ius esse, sed illam demum, quae effectum aliquem moralem involvit apud alios, qui ejusdem mecum sunt naturae. » DJNG, III.v.3, p. 263, l. 21-23. Cf. aussi DJNG I.vii.13 ; II.ii.3.

> « Ainsi, le cheval de la fable avait une faculté naturelle de paître au pré, et le cerf l'avait aussi ; ni l'un ni l'autre cependant n'avait de droit, parce que cette faculté de l'un n'affectait pas la faculté de l'autre. Il en va ainsi de l'homme, quand il emploie des choses dénuées de raison, ou bien des animaux, en se bornant à exercer sa seule faculté naturelle ; tant que celle-ci est considérée précisément dans l'ordre qu'elle a à des choses et à des êtres animés selon un usage qu'elle en a, abstraction étant faite de la relation à d'autres hommes. »[34]

Et en renversant du côté positif, en ce qui concerne la pratique humaine :

> « Mais [la faculté] n'acquiert la nature du droit proprement dit, que quand chez les autres hommes se produit cet effet moral, qu'ils ne doivent pas empêcher l'homme (*eum*), ni s'efforcer, sans son consentement, de s'accaparer ces mêmes choses. Il est donc insensé de vouloir distinguer cette faculté par le nom de droit, alors que tous les autres seraient en posture avec un droit égal d'empêcher celui qui voudrait l'exercer. »[35]

Contre Hobbes, Pufendorf fait porter ici sur la notion de droit un principe de cohérence. On pourrait formuler ce principe de la manière suivante (en tenant x pour un agent quelconque et F pour une action quelconque) : si x a le droit de F, alors il n'est pas que les autres agents que x aient un droit égal d'empêcher x de F. Pufendorf argumente donc de façon quelque peu indirecte :

> « Nous admettons que naturellement le pouvoir de l'homme le porte à user des êtres inanimés et des bêtes. Mais cette faculté, considérée précisément, ne peut pas être appelée « droit » ; tant parce dans celles-ci il n'y a nulle obligation de se prêter à ces usages ; que parce que, en vertu de l'égalité naturelle des hommes entre

[34] « Sic, uti in fabulis, facultatem naturalem habebat equus pascendi in prato, habebat eandem et cervus ; neuter tamen jus habuit, quod illa utriusque facultas alterum non afficeret. Sic homo, quando res sensu destitutas, aut bruta in usum suum adhibet, meram duntaxat facultatem naturalem exercet ; siquidem illa praecise consideretur in ordine ad res, & animantes, quibus utitur, citra respectum ad alios homines. » DJNG, III.v.3, p. 263, l. 23-27. Le cheval et le cerf sont ceux d'Esope, repris par exemple chez Jean de la Fontaine, *Fables* IV.13.

[35] « Sed quae tunc demum in juris proprie dicti naturam evalescit, quando in caeteris hominibus hic effectus moralis producitur, ne alii eum impedire debeant, aut ipso invito ad easdem res usurpandas concurrere. Ineptum quippe est, eam facultatem juris nomine insignire velle, quam exercere volentem alii omnes pari jure impedire queant. » DJNG, III.v.3, p. 263, l. 27-31.

> eux, aucun d'entre eux ne peut à juste titre (*recte*) exclure les autres de l'usage de ces choses, si ce n'est en vertu du consentement exprès ou présumé qu'il aurait acquis en particulier. Là où cela s'est fait, c'est alors seulement que l'on peut dire à juste titre qu'on a un droit sur une chose. »[36]

11. Conclusion

Ma thématique m'a permis de présenter des matériaux propres à alimenter une interrogation de caractère logique au sens large, celle de la coupure logique entre le fait et le droit. Mon appréciation d'ensemble est que Hume, avec le sens aigu de la synthèse et de la focalisation qui est le sien, ramène à un point transversal de logique, ce qui est encore chez Pufendorf une préoccupation relativement diffuse, dans l'horizon d'un système de droit naturel. De cette préoccupation dispersée, j'ai recueilli dans ce qui précède quelques points d'application, avec ce que Pufendorf dit de Grotius et de Hobbes. Dans la mesure où ces points d'application procèdent clairement d'une préoccupation systématique qui a pour effet de découpler le fait et le droit, il conviendrait – sur ce thème de la coupure entre fait et droit – de parler davantage d'une « loi de Pufendorf-Hume » que seulement d'une « loi de Hume ». Bien entendu, si on définit un domaine de faits à partir des *entia moralia*, on retrouvera en revanche un Pufendorf proche de John Searle[37], ce qui n'est pas surprenant.

[36] « Igitur hoc admittimus, naturaliter competere homini facultatem ad usus suos adhibendi res quasvis sensu carentes, ut & bruta. Verum ea facultas, ita praecise considerata, jus proprie vocari nequit, tum quia in istis nulla est obligatio ipsius sese usibus praebendi ; tum quia propter aequalitatem naturalem hominum inter se non potest unus caeteros ab iisdem rebus recte excludere, nisi ex eorum consensu expresso aut praesumto id sibi peculiariter comparaverit. Quod ubi factum est, tunc demum recte jus se ad eam rem habere dicere potest. » DJNG, III.v.3, p. 263, l. 32-37.

[37] Sur ces questions, voir le travail de l'A. mentionné en note 14 ci-dessus, ainsi que l'étude de J.-L. Gardies signalée au même endroit.

50

A challenge for moral rationalism: why is our common sense morality asymmetric?

STÉPHANE CHAUVIER

What is a moral rationalist? A moral rationalist doesn't need to be a moral realist, believing that Good and Evil or Right and Wrong have a *transcendent reality* and could exist even if human agents or at least conscious agents did not exist. But a moral rationalist should at least believe that Good and Evil or Right and Wrong have an *objective reality* or are objects of moral truths of an universal validity. But why is moral rationalism a *ratio*nalism? What is the place or the role of reason in moral rationalism? We could be tempted to say that a moral rationalist believes not only that moral values have an objective reality but also that they are cognitively accessible to human reason and accessible in a non-sophisticated way. For, of course, nothing precludes that an objective reality could remain unknown, being only an object of uncertain opinions. But is the belief in the cognitive accessibility of objective moral values sufficient to characterize moral rationalism? One thing is, for an objective fact, to be cognitively accessible, even easily accessible, accessible to common sense, and another is, for that same objective fact, to be *comprehensible*, to be consonant with the structural laws of reason and deducible from them. But nothing preclude Good and Evil or Right and Wrong to be 1) objective, 2) cognitively accessible *but* 3) irrational or incomprehensible to human reason. Suppose for instance that, somehow, we find out that it is morally good to benefit people from 8 to 12 o'clock, morally good to harm them from 12 to midnight and morally indifferent to benefit or harm them during the night. Perhaps some shrewd philosopher could propose a justification of that complex moral fact. But it is doubtful that such a justification would pass for

a genuine expression of Reason, because rational comprehensibility is consonant with simplicity, evidence, common sensibility and so on. There are therefore three dimensions to moral rationalism: a moral rationalist believes that Good and Evil or Right and Wrong are plainly objective ; he also believes that they are cognitively accessible in a non-sophisticated way, that moral knowledge is not reserved to the happy-few. But he must also believe that Good and Evil or Right and Wrong are fully comprehensible, that moral truths are rational truths or that Good and Evil or Right and Wrong don't appear to human Reason as the result of an objective but irrational *Fiat*.

Bearing in mind the recent and admirable book of Pascal Engel[1], where he pleads for a sort of super-rationalism that one may label "neo-bendatism", I'd like to address to his rationalist zeal what seems to me a challenge or, at least a puzzle for any moral rationalism : why is our common sense morality asymmetric with regard to these two dimensions of altruism that are, to borrow Quine's phrasing[2], "the passive respect for the interests of others" and "the active indulgence of their interests to the detriment of one's own" ? This asymmetry may be summarized in the following manner :

— On the one hand, it is morally worse not to passively observe the interests of others than not to actively support their interests.

— On the other hand, it is morally better to actively support the interests of others than to passively observe their interests.

Or, to put it in a less precise, but more striking way :

— On the one hand, it is morally worse to do harm others than not to benefit them.

— On the other hand, it is morally better to actively benefit others than to restrain from harming them.

The general asymmetry of good and evil could thus be represented in the following manner :

	Good	**Evil**
Maxi	Benefiting	Harming
Mini	Not harming	Not benefiting

[1]. Engel (2012).

[2] . Quine (1987), p. 4. There are few studies devoted to the normative asymmetry between Good and Evil. We could mentioned Hurka (2010), Mayerfeld (1996) and McMamara (1996). The *normative* asymmetry between Good and Evil must be firmly distinguished from the *epistemic* asymmetry, the fact that Evil is much more recognizable or knowable than Good.

This asymmetry is not only present in our common sense morality : it is also validated by the main part of the traditional moral philosophy, with the notable exception of utilitarianism, when it distinguishes duties of right and duties of virtue, perfect duties and imperfect duties or erogatory and supererogatory duties.

How can a moral rationalist reconcile himself with this asymmetry, since, as I will try to show, systems either symmetric or inversely asymmetric are perfectly conceivable or are in no way repugnant to reason ?

1. Illustrations of the asymmetry

I'll start by giving some intuitive illustrations of this double asymmetry:

1. First, subject to qualifications to which I shall come back in a moment, the idea that it is more serious, morally speaking, to harm others than not to benefit them can be widely illustrated.

— Making someone poor seems more serious than not helping him to get out of his poverty.

— Depriving someone of his favourite music seems more serious than not playing him a tune of his favourite music.

— Punishing an innocent seems worse than not forgiving a guilty party.

— Abandoning our children seems worse than not adopting orphan children.

— Giving birth to a child that we know to be genetically condemned to a short life of intolerable suffering seems worse than not giving birth to a child who is devoid of any known genetic problem.

Of course, when we say that it is more serious to do A than to do B, we don't intend to say that it is indifferent to do B. But we indicate that there is a hierarchy in evil, that there is a maxi-evil and a mini-evil.

2. These various asymmetries in the registry of moral blame can also be found in the registry of moral praise.

— If it is more serious to make someone poorer than to not help him to get out of his poverty, it is most laudable, more meritorious to help someone to get out of his poverty than to aggravate his poverty.

— If it is more serious to deprive someone of his favourite music than not playing him a tune of his favourite music, it is more praiseworthy, more meritorious to play someone a tune of his favourite music than not depriving another of his favourite music.

— If it is more serious to punish an innocent than not to forgive a guilty party, it is much more laudable, more meritorious to forgive a guilty person than to not punish an innocent.

— If it is more serious to abandon our children than not to adopt children of others, it is more commendable to adopt orphan children who need us than not to abandon our children.

— If it is more serious to give birth to a child that we know to be genetically sentenced to a life of intolerable suffering than not giving birth to a child that we know to be devoid of any genetic problem, it is or should be more praiseworthy, more meritorious to give birth to a child without serious genetic disease than not to give birth to a child with severe genetic disease.

3. I don't deny the fact that, by presenting these illustrations as obvious, one can imagine concrete situations where these various hierarchies lose their obviousness. But the reason is that the asymmetry of maleficence and beneficence in the registry of praise as of blame can be disturbed by the effect of an extra-moral distinction : the distinction between doing and letting happen or between acting and not intervening.

Suppose a person is in an urgent need of some basic goods. One can either exacerbate her situation by stealing the little that remains to her, or one can not improve her situation, although one could. When we say that it is more serious to aggravate her case than not to improve it, we do not say, as we have already pointed out, that it is indifferent to do nothing for her. However, it seems intuitively less severe to lack compassion than to aggravate shamelessly the situation of an already miserable person. But what disturbs this intuitive moral hierarchy between harming and not benefiting, is that there are situations that could be called dynamic, as opposed to static situations. In the previous example, a person is in need and can remain statically in this state. But the intuitive appreciation changes if one considers a person whose need is gradually increasing up until the point of causing her death. Because of this dynamic situation, the clear duality between aggravating and not ameliorating gives way to a much tighter distinction between accelerating the worsening of the situation of the person and letting her fate become worse. In other words, "not improving" becomes in this case equivalent to "letting things worsen".

Are these dynamic situations a challenge to the general asymmetry between harming and not-benefiting or between non-harming and benefiting ? Not really. They only indicate that, due to the dynamic nature of a situation, we sometimes have to make a semantic decision when we evaluate from a moral point of view the contrast between worsening and not improving. In

the literature devoted to the difference between acting and letting happen, the main cases considered are dynamic situations with fatal issues, as when a baby is drowning in his bath : in such a case, it seems difficult not to equate "not improving" with "letting aggravate" and "letting aggravate" with "aggravating". But if we consider less tragic dynamic situations, it is not clear that the asymmetry between harming and not benefiting cannot find its place. For example, there is a clear difference of moral seriousness, when we set aside any fatal perspective, between misleading someone into lying and not disabusing him or letting him sink into error. If we introduce death into the issue, it will tend to crush the asymmetry on the most serious side. But if the ultimate issue of the dynamic situation is not death but a benign outrage, we will maintain the asymmetry, we will estimate that deceiving others is worse than not having disabused them.

2. Semi-formalization of the asymmetry

Before attempting to interpret this asymmetry, I'll try to characterize it in a semi-formal way.

We can be tempted to compare the asymmetry between harming and non-benefiting with another (alleged) asymmetry, the asymmetry between duties and virtues or between the language of duties and the language of virtues. Thus, if it is virtuous to sacrifice our life for others, it seems difficult to argue that there is a duty to sacrifice oneself, something like a "You must sacrifice your life to save the one of others". In reality, as suggested by this example, the asymmetry between virtue and duties has its source in the existence of supererogatory conducts, *i.e.* conducts that are worthy of moral praise, which are virtuous, but which are not duties or are beyond the duty, beyond the *debitum*. But it does not seem that the moral difference we are interested in here, the difference between not harming and benefiting overlaps with the difference between the erogatory and the supererogatory. A very large number of conducts of active altruism are traditionally regarded as *debita*, and not just as heroic conducts. To take some Kantian examples, benevolence, charity, sympathy are not heroic conducts: they are genuine erogatory conducts. So if we accept that the moral difference between passive and active altruism or between non-maleficence and beneficence falls within the field of *debita*, we have to conclude that the only way to formally express the difference between the duty not to precipitate others into misery and the duty to contribute to alleviating their misery is to introduce a difference between two modalities of

duty or between two degrees of "moral imperativity", if you will.

In standard systems of deontic logic, we only find a single concept and a single operator of duty : Ob (¬p) ⇔ ¬Pe (p) ⇔ For (p). "It is obligatory that non-p" is logically equivalent with "it is not permissible that p" and with "it is forbidden that p". And we understand these equivalences if *p* represents the action of taking from others the little that they have. But such an unique operator of moral obligation does not capture what we have previously expressed in terms of a scale of moral gravity or, conversely, in terms of a scale of moral merit. If the obligation to respect the life of others gives rise to an absolute prohibition of killing them, the obligation to actively advance the interests of others produces only a *conditional or modulated prohibition* of not helping, which can go until a legal and even a moral tolerance of its opposite.

Therefore, it does not seem possible to represent the logic of obligation to help others, at least as it is commonly understood when we set aside the tragic dynamic situations we have evoked, as involving an absolute prohibition of not helping. "You must help others" means here: "You must strive to help others" but not "It is forbidden not to help others". Doubtless it may seem paradoxical, given standard logical habits, to say that the obligation to do something is not equivalent with the prohibition of not doing that thing. But it is also quite surprising that deontic logicians have been inclined to believe that such a logic of obligation could apply to all our effective and not supererogatory moral duties.

Without going into the complexity of the construction of a more adequate formalism, we propose, in order to capture the intuitive difference between the obligation of not-harming and the obligation of benefiting to introduce two concepts and therefore two operators of moral obligation:

— the strict or narrow obligatory, the contrary of which is prohibited.

— the wide or large obligatory, the contrary of which is excusable.

What these two forms of moral obligation have in common is that the obligatory, whether strict or wide, is the non-optional or the non-indifferent. But the difference is that the strict obligatory is also the non-omissible, which the wide obligatory is not. However, and this is where there is a problematic lacuna in the standard systems of deontic logic, when we say that the wide obligatory is different from the non-omissible, we don't mean that the opposite of the wide obligatory is permitted. What we express by the proxy notion of the Excusable designates a medium between the Prohibited and the Permitted, a medium which corresponds to reproach or blame, in contrast with punishment on one side, indifference on the other.

These two forms of obligation which I'll henceforth label the Mandatory Obligation and the Latitudinal Obligation relate mostly, for the first, to the conducts of non-harming and, for the second, to the conducts of beneficence. However, we say "mostly" because there are limit-cases:

— On the one hand, there are forms of beneficence that, as we have seen, fall under the Mandatory Obligation: those that are connected to dynamic situations where a person is in pressing danger and that are discussed in the literature under the name of the problem of the Bad Samaritan.

— On the other side, there are forms of side-effects, that are intentional or internal and not external, but that, because of their lightness and inevitability, fall within the scope of the Latitudinal Obligation and form the thin register of the *sub*erogatory[3]. For instance, pushing people to make space in a crowded subway seems an excusable violence.

3. Four moral systems

If we accept the distinction between the two regimes of obligation and the idea that, essentially, these two regimes cover the distinction of non-maleficence and beneficence, we can define, in a purely formal manner, four possible moral systems, four way to deal morally with the both sides of altruism that are the passive compliance and the active support of the interests of others, the non-harming and the benefiting sides of morality. If we note *Obst* "It is strictly obligatory " and *Obla* "It is latitudinarly obligatory", we can obtain the four following combinations :

1. Our asymmetry or **System of Gracious Goodness**.

Obst (not harming) $\Leftrightarrow$ It is forbidden to harm.

Obla (benefiting) $\Leftrightarrow$ It is excusable not to benefit.

2. Reverse asymmetry or **System of Double Effect**.

Obla (not harming) $\Leftrightarrow$ It is excusable to harm.

Obst (benefiting) $\Leftrightarrow$ It is forbidden not to benefit

3. Imperious symmetry or **System of Severe Goodness**.

Obst (not harming) $\Leftrightarrow$ It is forbidden to harm.

[3] Driver (1992).

Obst (benefiting) $\Leftrightarrow$ It is forbidden not to benefit

4. Latitudinal Symmetry or **System of Great Tolerance**.

Obla (not harming) $\Leftrightarrow$ It is excusable to harm.

Obla (benefiting) $\Leftrightarrow$ It is excusable not to benefit.

From a strictly formal point of view, these four systems have no inconsistencies. The second could perhaps be suspected : How can it be excusable to harm and prohibited not to benefit ? But besides the fact that "excusable" does not mean "permissible", one has just to contemplate the cases where one harms somebody while benefiting another. If benefiting were an imperious duty, the cases of double effect could become more frequent. It is the reason why I call this system the *System of Double Effect*.

Then, if we admit that none of these four systems is repugnant to pure reason, the question arises to why we are morally living in the first. And it seems to us that the explanation is mainly extra-moral.

4. Attempts to justify or to explain the asymmetry.

Justifications ?

First, could we justify the moral asymmetry between passive and active altruism ? The main rationale for such a justification is the one proposed by Kant who distinguishes duties of right and duties of virtue. This justification can be thus summarized : the imperious character of the obligation of passive altruism, by contrast with the latitudinarian character of the obligation of active altruism derives from the fact that the duty of not harming others has its basis in the rights of others of not being harmed, while the duty to actively support the interests of others is not correlated to a right of others to be supported in their interests. The problem with this justification is that it presupposes that rights are knowable independently of duties. But nothing can exclude that the inverse is true, that it is the binding force of moral duties which permits us to identify natural rights. Let us indeed consider the system of the Imperious Symmetry : it prohibits both harming and not benefiting. Why couldn't it be said that, in such a system, each person has the right of not being harmed *and* the equal right of being supported in the advancement of his interests? Of course, if others, but not me, had this right, I should become in some way

the slave of others. But if these rights are reciprocal, what we obtain is only another system of natural rights than the one that is involved in the System of Gracious Goodness, a denser system, but not necessarily an inconsistent system. On the one hand, I could have a full right of ownership to certain things, such that others should not undermine it. But, on the other hand, others may be entitled to demand of me that I use my things for the advancement of their interests. Why should all these rights not be compatible? It is therefore doubtful that the asymmetry could be justified by appealing to a substructure of rights, since nothing demonstrates that rights are *logically* prior to duties.

The same treatment can be reserved to another justification advanced by Kant, in conjunction with the preceding one, namely that the difference between mandatory obligation and latitudinal obligation is related to the difference between immoral acts that may be objects of a public repressive constraint and acts that may not. This rationale raises a problem that is of the same nature as the previous one : what does it mean "to be [the] object of a public repressive constraint"? Is it because the duty not to harm is imperious that people *may* be publicly compelled to respect it or is it because we *can* compel people to not harm others that they have an imperious duty of not harming? In the first case, we have a "may" with a normative meaning. In the second case, we have a "can" with a physical sense. But we can reject the second interpretation : the normative status of an obligation couldn't be founded on the contingent fact that only a certain kind of acts can be objects of a public repressive constraint. The difference between Mandatory Obligation and Latitudinarian Obligation cannot be founded on the contingent fact that some acts are punishable and that others are not or that some acts are more readily punishable than others. It is therefore more likely to say that the existence of a public constraint is itself a consequence of the importance attached to certain duties and therefore of the severity attached to some moral transgressions, rather than conversely. But, if the existence of a public constraint is a consequence of the hierarchy between two kinds of obligation within the system of Gracious Goodness, it follows that the public constraint cannot serve as a justification of that hierarchy. Within another system, the public constraint could be attached to none, to both or to another kind of duties. For instance, in the system of Double Effect, the transgressions of the duties to actively support the interests of others would be sanctioned by a public repressive constraint, but not the transgressions of the duties to passively respect the interests of others.

Therefore neither public constraint nor rights can justify the difference between imperious duties and latitudinal duties, since it is that difference that

permits us to identify the rights and to define the area of the public constraint. The system of rights and the scope of legitimate public constraint would vary according to the moral system that could be adopted.

Are other explanations conceivable ? For instance utilitarian justifications ? Would the System of Gracious Goodness be the one in which the general well-being would be the greatest ? This seems clearly false. From the standpoint of the maximization of the overall utility, the System of Severe Goodness outperforms the System of Gracious Goodness, since the latter differs from the former by the fact that non-beneficence is prohibited in the former and excusable in the latter. And in fact, it is to such a system of Severe Goodness that the utilitarian doctrine leads, by promoting universal beneficence as the basic moral imperative and by treating non-harming as a consequence of the principle of maximization of overall welfare.

So the appeal to the principle of utility is not more relevant than the appeal to natural rights or to public constraint in order to justify the asymmetry. Perhaps some other justifications could be tested, but the most common are ineffective or inconclusive.

Explanations

If there are no moral reasons that can justify the fact that we are living and reasoning within the System of Gracious Goodness rather than within one of the three other systems, could we at least *explain* the fact that we are leaving and reasoning in that system, that we treat asymmetrically passive and active altruism ?

We will mention three explanations, but, in our view, only a combination of the three can be a sufficient explanation.

1. It is worse to suffer than it is good to enjoy.

The first possible explanatory factor is well known : there is a natural or affective asymmetry that, with the concourse of sympathy, may explain the normative asymmetry. As Popper writes in a footnote in chapter 9 of *The Open society* : "there is no symmetry between suffering and happiness or between pain and pleasure[4]". And he infers from that natural fact that, from a moral point of view, it is more urgent, more pressing to alleviate the suffering of others than to increase their happiness.

[4] Popper (1966), p. 284.

How can this natural asymmetry be mobilized to explain the normative asymmetry? At the condition that we admit the validity of the following oblique inferences:

(1) It is worse to suffer than it is good to enjoy.

Therefore :

(2) It is worse to inflict pain than it is good to give joy.

Therefore :

(3) It is worse to inflict pain than not to give joy.

And therefore :

(4) It is better to give joy than not to inflict pain.

Obviously "better" and "worse" mean here “sympathetically more pleasant” and “sympathetically more painful”. A complete explanation of the asymmetry would then be that there is, in morality, an unnoticed influence of sensitivity on the fixation of the meaning of the normative moral concepts of Good and Evil.

2. *The passive respect for others is necessary to society while the active support of their interests is only a social ornament.*

This very different explanation was advanced by Pufendorf :

> “That some things should be due to us perfectly and others imperfectly, the reason amongst those who live in a state of natural liberty is the great diversity of precepts of nature’s laws, of which some conduce to the very being, others only to the well-being of society. And therefore since there is less necessity of performing these latter than the former, reason shows that the former may be required and executed by more severe courses and means, whereas in regard to the latter; it is mere folly to apply a remedy more grievous than the disease.” Pufendorf, *The law of nature and people*, I, vii, 7[5].

One finds an echo of that argument in a passage of Kant’s *Doctrine of Virtue* :

[5]. *Pufendorf (1729), p. 81.*

> "Casuistic question. Would it not be better for the well-being of the world generally if human morality was limited to duties of right, fulfilled with the utmost conscientiousness, and benevolence were considered morally indifferent (adiaphora) ? It is not so easy to see what effect this would have on human happiness. But at least, a great moral adornment, love of man, would be missing from the world. Love of man is, accordingly, required by itself, in order to present the world as a beautiful whole in its full perfection, even if no account is taken of advantages (happiness)." Kant, *Doctrine of Virtue*, §35[6].

The explanation involved in these two passages of Pufendorf and Kant seems to us very strong. It makes the structure of our morality depend on an essential feature of human society: namely that the course of social life renders it largely possible to serve the interests of others by first taking care of our own. The asymmetry may thus be explained by the fact that nature, in part, takes charge of the reciprocal support of interests, while it does nothing to help their passive respect. In other words, according to this view, the urgency of the obligation of non-maleficence would result from the *social necessity* to prevent maleficence artificially, while the latitudinarity of the obligation to actively support the interests of others would be the expression of the fact that, in most cases, voluntary support to the advancement of the interests of others is a simple adornment of social life.

3. Maleficence is very recognizable, whereas beneficence is subject to dispute.

That third explanation may be compared with one of the ways by which Kant explains the difference between duties of right and duties of virtues :

> "Ethical duties are of wide obligation whereas duties of right are of narrow obligation. [...] for if the law can prescribe only the maxim of actions, not actions themselves, this is a sign that it leaves a latitude (latitudo) for free choice in following (complying) with the law, that is, that the law cannot specify precisely in what way one is to act and how much one is to do by the action for an end that is also a duty. But a wide duty is not to be taken as permission to make exceptions to the maxim of actions, but only as permission to limit one maxim of duty by another (e.g. love of one's neighbour in

[6]. *Kant (1991), p. 251.*

> general by love of one's parent). [...] Imperfect duties are, accordingly, only duties of virtue. Fulfilment of them is merit (meritum = +a); but failure to fulfil them is not in itself culpability (demeritum= −a), but rather mere deficiency in moral worth = 0, unless the subject should make it his principle not to comply with such duties" Kant, *Doctrine of virtue,* introduction, VII[7].

Besides the appeal to the rights and to the public constraint that we have already mentioned, Kant also explains the difference between two kinds of duties by distinguishing duties that command certain actions and duties that command maxims of action or that command one to target some ends. Thus we can distinguish the duty to give back to others what we have borrowed from them and the duty to contribute to advancement of the happiness of others. In the first case, the agent must only remember what he has borrowed from others, for instance how much they have lent to him, whereas in the second case, he has to determine a) whose interests he has to actively support b) by doing what and c) by taking into account his others obligations and purposes.

We could summarize this distinction by saying that passive compliance with the interests of others requires less expenditure of epistemic energy than the active support of their interests. The difference between the Imperious and the Latitudinal obligations would derive from a difference of epistemic order.

5. Conclusion

To the question we asked: "Why do we judge that it is worse to harm others than not to benefit them and better to benefit others than not to harm them" we can now propose a synthetic answer by mixing the three preceding explanations : "Because it is more painful to suffer than it is pleasant to enjoy, because it is more expensive, epistemically and pathologically, to benefit others than not to harm them and because also it is socially less necessary to support actively and voluntarily the interests of others, since an invisible hand accomplishes, partly, that task."

If, on the one hand, these three concurrent explanations are plausible and if, on the other hand, the system of Gracious Goodness is only one of the systems of morality that are logically possible, then it follows that our common

[7]. *Kant (1991), p. 194.*

sense morality cannot be seen as a system of rational moral truths, the contrary of which is inconsistent. Alternative moral systems are conceivable and could have been ours if the contingent features that explain our adoption of the system of Graceful Goodness had been different.

The pluralism that we find at a systematic level in Morals makes a pair with the pluralism we also find in Ontology. Systematic pluralism doesn't condemn rationalism, but it introduces a component of decision at the heart of the reason, and a decision that cannot be itself rational or, at least, *a priori*[8]. As a geometry, a moral system is an axiomatic possibility that we have adopted for reasons of empirical convenience, but that cannot be proved to be the only deontic articulation of Good and Evil that is rationally conceivable.

6. References

Driver Julia (1992), "The Suberogatory", *Australasian Journal of Philosophy*, 70, 286-295.

Engel Pascal (2012), *Les lois de l'esprit. Julien Benda ou la raison*, Paris, Éditions d'Ithaque.

Kant Immanuel (1991), *The Metaphysics of Morals*, transl. Mary Gregor, Cambridge, Cambridge University Press.

Hurka Thomas (2010), "Asymmetries in Value", *Noûs*, 44 (2), 199-223.

Mayerfeld Jamie (1996), "The Moral Asymmetry of Happiness and Suffering", *The Southern Journal of Philosophy*, 34 (3), 317-338.

McNamara Paul (1996), "Doing Well Enough. Toward a Logic for Common-Sense", *Studia Logica*, 57, 167-192.

Pufendorf (Samuel), *Of The Law of Nature and Nations*, transl. Basil Kennett, London, 1729. https://openlibrary.org/books/OL6563001M/Ofthelawofnatureandnations.

Popper Karl (1966), *The Open Society and its Ennemies*, vol. 1 "The Spell of Plato", London, Routledge.

Quine W.V.O. (1987), *Quiddities*, Cambridge (Mass.), Harvard University Press.

Von Wright Georg H. (1951) "Deontic Logic", *Mind*, 60 (237), 1-15.

[8] It is the leading idea of Vuillemin (1986).

Vuillemin Jules (1986), *What are Philosophical Systems ?*, Cambridge, Cambridge University Press.

Part Five

The Dispute

51

Tool-Box or Toy-Box? Hard Obscurantism in Economic Modeling

JON ELSTER

1. Introduction

Pascal has written incisively on "soft obscurantism" in the humanities, notably in the wonderful *Instructions aux académiques*, written by him but published under the penname Federico Tagliatesta, with a preface signed by Pascal. When he sent me the book and I read it, I was taken in by the hoax – I believed the story about the young Italian philosopher who died tragically in an accident. More recently, at a conference for which I wrote this paper, he presented a hilarious, erudite and conceptually sharp analysis of the distinction between stupidity and folly. It's good to have him as a partner in the struggle against obscurantism, although "mit der Dummheit kämpfen selbst Götter vergebens".

In the present article I consider the less frequently phenomenon of "hard obscurantism", a species of the genus scholarly obscurantism. In academic debates, a more common term for obscurantism is "bullshit", first identified as an intellectual pathology by Harry Frankfurt (1988); later discussions include Cohen (2002) and Gjelsvik (2006). I shall not enter into the conceptual debate concerning the fine grain of bullshit. The project is perhaps hopeless, since, as any reader of undergraduate essays will know, confusion resists precise capture. One may perhaps, distinguish between *obscure* writers and *obscurantist* writers. The former aim at truth, but do not respect the norms for arriving at

truth, such as focusing on causality, acting as the Devil's Advocate, and generating falsifiable hypotheses. The latter do not aim at truth, and often scorn the very idea that there is such a thing as the truth.

The preceding remarks apply in obvious ways to *soft* obscurantisms. The best way of characterizing this cluster is by *extension*: it includes post-modernism, subaltern theory, queer theory, deconstruction, psychoanalysis, functionalist explanations, multiculturalism, structuralism, and several others (see Elster 2012 for examples). Beyond the weak or strong disregard for truth, these schools or movements have so little in common that a definition by *intension* seems impossible.

So far, bullshittologists have not focused their attention on hard obscurantism. The term "hard" is vague, but points in the direction of quantitative, mathematical, computer-based, or formal analyses. In the broadest sense, hard obscurantism can be found in a number of academic disciplines. I believe that analytical philosophy and linguistics sometimes exhibit pointless or excessive formalism. Much work is model-driven, not world-driven. These disciplines are not, however, my focus here. I shall consider hard obscurantism in the social sciences, notably in economics and in the increasing body of political science that relies on economic models.

Before I proceed, a comment on the *relation* between the two forms of obscurantism may be in order. Historically, the introduction of quantitative or formal modes of analysis may have been a reaction to prevalent forms of soft obscurantism. Analytical philosophy arose to some extent as a reaction to the perceived soft obscurantism of Continental philosophy. Mathematical economic models arose to some extent as a reaction to purely verbal economics, which prevented economists from estimating the *net effect* of the many causal mechanisms at work in the economy. Many of these reactions were salutary. Yet by virtue of psycho-sociological mechanisms that I shall not consider here (but see Elster 2009 a, 2012), the models often took on a life of their own and became increasingly dissociated from their original aims. Once hard obscurantism took off, its existence seemed to justify soft obscurantism. In political science, for instance, the "perestroika" movement that briefly flourished around 2000 was a soft-obscurantist reaction to what was correctly perceived as a pernicious turn of the discipline towards hard obscurantism. Each extreme might seem to justify the other. Scholars who fought a two-front war against both ran the risk of being attacked by both or enlisted as allies by both.

To characterize hard obscurantism in the social sciences, let me once again begin with an extensional definition. The case I consider in this article is, as in the title of a book by Green and Shapiro (1994) that usefully supple-

ments my analysis, *pathologies of rational-choice theory*. I want to emphasize, however, that I believe rational-choice theory in general, and game theory more specifically, have immense *conceptual* value. The simple budget-line-cum-indifference-curve analysis found in any microeconomics textbook created light where previously there was semi-obscurity. The observation that it can be strategically rational to burn one's bridges or one's ships made sense of behavior that had seemed unintelligible or irrational. The distinction among the Prisoner's Dilemma, the Assurance Game (or Stag Hunt), the Battle of the Sexes and the Game of Chicken has illuminated complex structures of social interaction that were previously only dimly understood. We understand today why a bad state may persist indefinitely if it is an *equilibrium*, in which no single agent has an incentive to change behavior unilaterally. In fact, I have suggested elsewhere (Elster 2009 a) that hard obscurantism may itself be a bad equilibrium.

Rational-choice theory can also have great *practical* value, as a routine and indispensable tool for ministries of finance, central banks and similar institutions. Assuming that consumers and producers respond rationally to incentives, institutions can set interest rates or tax rates to achieve socially desirable goals. Typically, though, the theory only allows predictions of *short-term effects of small changes*. Since reforms are usually small, incremental and reversible, this is sufficient for most practical purposes. The following contrast may illustrate the limits of the approach. It is possible, or so I assume, to estimate pretty accurately the effect of a one per cent increase in the price of alcoholic beverages on the legal purchase of such liquids. I do not think, however, anyone can estimate the effect of tripling the price, which might induce smuggling and home brewing to an incalculable extent. For an even more dramatic example, I do not think anyone can estimate the effects of legalizing hard drugs. To do so, one would have to estimate the effect of legalization on preference formation, and not simply the effect on consumption for given preferences of rational individuals. Nobody understands how preferences are formed and transformed.

The U. S. Constitution (Art. I.8) presupposes rational, incentive-responsive agents when it gives Congress the power to "promote the progress of science and useful arts by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries". More recently, Nobel prizes in economics have been awarded for the design of auctions and for matching medical interns to hospitals (among other pairs). The designers of incentives systems should always keep in mind, however, the possibility that the schemes may be *gamed* by rational agents. The perverse effects of us-

ing citation counts to allocate funds to academic institutions and individual scholars offer a well-known illustration. Such examples, which could be multiplied, do not of course amount to an objection to rational-choice theory, only to its practical usefulness.

Rational-choice models are not the only sources of hard obscurantism in the social sciences. I shall briefly and without much nuance mention three other theoretical approaches that can – but do not necessarily – favor hard obscurantism.

Statistical models can illustrate hard obscurantism in economics and political science. In his criticism of the use of such models in the social sciences, David Freedman (2005, 2010) identifies *gross errors* in six articles published *in the leading journals* of these disciplines. It goes without saying that any model or theory can be used in sloppy and irresponsible ways. The use of statistics in the courtroom is a notorious example. *When elite journals publish bad science, however, the situation is serious.* In a famous article on "Statistical models and shoe leather", Freedman (2010, p. 46) comments on regression analysis in the following terms:

> A crude four-point scale may be useful: 1. Regression usually works, although it is (like anything else) imperfect and may sometimes go wrong. 2. Regression sometimes works in the hands of skillful practitioners, but it isn't suitable for routine use. 3. Regression might work, but it hasn't yet. 4. Regression can't work. Textbooks, courtroom testimony, and newspaper interviews seem to put regression into category 1. Category 4 seems too pessimistic. My own view is bracketed by categories 2 and 3, although good examples are quite hard to find.

In my discussion of rational-choice pathologies, I shall follow Freedman's example by criticizing work by prestigious economists and political scientists published in elite journals or by elite publishers. It would be an easy and pointless victory to criticize work that is bad by the standards of the discipline.

Agent-based models (a form of simulation) can also illustrate hard obscurantism in the social sciences. As originally introduced by Schelling (1971) in a study of race-based residential segregation, these models had great appeal, because of their transparency. One could follow, step by step, the mechanism by which individuals who prefer to live in a mixed neighborhood, but wish that a majority belong to their own race, make residential choices that lead

to complete segregation. More recent models are more complex and highly opaque. As they usually have many parameters, which can take on many values, one can fiddle with the settings to get virtually any outcome. In this respect agent-based models resemble the data-mining and curve-fitting of many statistical models. The results rarely have the ex-post obviousness that characterizes good science and that Schelling's model possessed. Agent-based models, from being a tool-box, have largely become a toy-box.

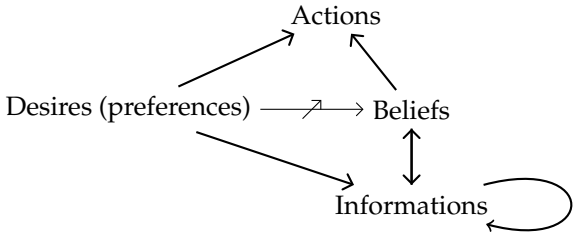
Behavioral economics, too, is running the risk of becoming a toy-box. There is probably not a week without some highly sophisticated experiment demonstrating a new "mechanism" or a new "effect" appearing on the Internet. Many of the findings seem to have little relevance outside the laboratory; at least, many authors do not try to demonstrate relevance "in the wild". When authors cite real-world cases to demonstrate, say, the sunk-cost effect, they rarely take the time to explore (and refute) alternative explanations of the alleged examples. The sunk-cost effect certainly seems to "fit" the Vietnam war or the construction of the Concorde airplane, but so do many other hypotheses. It is ironical that behavioral economics, which was largely inspired by the explanatory failures of rational-choice theory, is running into similar problems.

A rough intensional definition of hard obscurantism is that models and procedures become ends in themselves, dissociated from their explanatory functions. The term "toy-box" that I used to characterize (some instances of) agent-based modeling and behavioral economics can be extended to (some instances of) rational-choice modeling and statistical modeling. Concerning the latter, Ragnar Frisch deplored the tendency of econometrics to become "playometrics". (As he was the founding editor of *Econometrica*, this was a rare instance of *insider criticism* of hard obscurantism.) Concerning the former, the common defense of rationality as an "as-if" assumption also confirms this characterization. In the Summary, I propose a more fine-grained analysis.

I shall now proceed as follows. In Section **2** I set out the basic structure of rational-choice theory, and consider the two ways in which it is liable to fail: by the *indeterminacy of the models* and by the *irrationality of the agents* whose behavior the models try to explain. In Section **3** I exemplify and criticize hard obscurantism by examining the writings of eminent economists and political scientists, including three winners of the Nobel Prize in economics and two winners of the John Bates Clark Medal who have not (yet) received the Nobel prize. Section **4** offers a brief Summary.

2. Rational-choice theory and its failures.

Rational-choice theory is first a normative theory, advising agents about what to do to realize their aims as well as possible, and then an explanatory theory trying to account for the behavior of the agents by assuming that they follow this advice. The idea of realizing your aims as well as possible can be disaggregated into three maximization operations: among the available options, choose the one you believe will best realize your aims; use the information available to you to form the beliefs most likely be true; if necessary collect an optimal amount of new information. Diagrammatically:



A few comments may be useful for clarification. (i) A direct influence of desires on beliefs (wishful thinking) is incompatible with rationality. (ii) Because desires can (rationally) influence the amount of information one should collect, high-stake decisions requiring more information, and because information shapes beliefs, an indirect influence of desires on beliefs is compatible with rationality. (iii) In addition, the optimal amount of information to collect is also shaped by the agent's beliefs about the value of the information and the cost of acquiring it. (iv) The loop reflects the fact that the expected benefits may be modified by what is found in the search itself.

Philosophers tend to use this belief-desire model of rational action. Economists tend to use a discounted-expected-utility model. For my purposes here, they are equivalent. In the economist model, agents attach cardinal utilities to each possible outcome of each possible action, and subjective probabilities to the occurrence of each possible outcome of each possible action. The expected utility of an action is defined as the weighted sum of the utilities of its possible outcomes, with the probabilities serving as weights. If outcomes of an action will occur in the future and the agent has a positive rate of time preference, the expected utility must be discounted to its present value. A rational agent chooses the action that has the greatest discounted-expected utility.

This thumbnail sketch of the theory of rational choice ignores a host of sub-

tleties that are irrelevant for my purposes. In preparation for the next Section I should, however, mention one complication. The optimal "action" predicted by a theory need not be an ordinary physical action from which the agent derives immediate utility, such as eating an apple rather than an orange. Rather, it can be the decision to use a lottery device that will direct her to eat an apple with probability **p** and to eat an orange with probability (1-**p**). The use of such *mixed strategies* can be important in game-theoretic contexts.

Rational-choice theory as thus defined can fail in two ways. First, it can fail to yield a sharp determinate prediction. Second, it may yield a sharp prediction that fails the confrontation with the observed facts. The first problem is one of the *indeterminacy* of the theory, the second that of the *irrationality* of the agents.

Theory indeterminacy can arise for several reasons. I shall distinguish three sources: *an infinite regress* in the determination of the optimal amount of information, *uncertainty* (brute or strategic), and *cognitive limitations*.

In some cases, rational-choice theory cannot offer agents determinate advice about how much information to gather. I suspect but cannot prove that this failure is the rule rather than the exception. Winter (1964, p. 252) observed that the idea of reducing satisficing to a form of maximizing creates an infinite regress, since "the choice of a profit-maximizing information structure itself requires information, and it is not apparent how the aspiring profit maximizer acquires this information or what guarantees that he does not pay an excessive price for it." Along the same lines, Johansen (1977, p. 144) characterized the search process as "like going into a big forest to pick mushrooms. One may explore the possibilities in a certain limited region, but at some point one must stop the explorations and start picking because further explorations as to the possibility of finding more and better mushrooms by walking a little bit further would defeat the purpose of the hike. One must decide to stop the explorations on an intuitive basis, i.e. without actually investigating whether further exploration would have yielded better results". When rational belief formation is indeterminate, one does indeed have to rely on intuition. Even assuming that an observer can *predict* the outcome of intuition as based on heuristics and biases, that prediction will not reflect a normative *prescription* for the agent.

Uncertainty in the technical sense means that agents (i) know all possible actions and (ii) all possible outcomes of each action, but (iii) cannot assign precise numerical probabilities to the outcomes. Rational-choice theory does offer some weak advice about what to in such situations: follow some decision rule based on the worst and the best outcomes of each action (Arrow and Hur-

wicz 1971). The maximin rule is commonly recommended, but "maximax" is equally compatible with the (weak) demands of rationality. For all practical purposes, therefore, the theory does not offer any advice.

Brute uncertainty can arise in "games against nature", as in earthquake predictions or predictions of global warming (Weitzman 2009). It can also arise in complex social systems, such as financial markets. (The latter are also of course subject to strategic uncertainty.) Taleb (2007, p. 198-200) offers a general analysis of the issue:

> This problem [estimating the shape and the parameters of a distribution] has been seemingly dealt away with the use of "off-the-shelf" probability distributions. But distributions are self-referential. Do we have enough data? If the distribution is, say, the traditional Gaussian, then yes, we may be able to say that we have sufficient data – for instance the Gaussian tells itself us how much data we need. But if the distribution is not from such a well-bred family, then we may not have enough data. But how do we know which distribution we have on our hands? Well, from the data itself. [...]
>
> So we can state the problem of self-reference of statistical distributions in the following way. If 1) one needs data to obtain a probability distribution to gauge knowledge about the future behavior of the distribution from its past results, and if, at the same time, 2) one needs a probability distribution to gage data sufficiency and whether or not it is predictive outside its sample, then we are facing a severe regress loop. We do not know what weight to put on additional data.

Strategic uncertainty arises when agents have to form beliefs about one another, including beliefs about beliefs etc. In theory, one can short-circuit the looming infinite regress by the notion of an *equilibrium* set of strategies, defined as strategies hat are optimal against each other. In non-zero-sum games, these often involve mixed strategies that are only *weakly optimal* against each other, in the sense that an agent can do *just as well* by adopting one of the strategies in the mix as a pure strategy. In that case, however, why should other agents believe she is playing her mixed equilibrium strategy? Would it not be rational to assume that she plays it safe and adopts a maximin pure strategy? Strategic uncertainty can also arise when a game has several equilibria in pure strategies and none of them is weakly Pareto-superior to the others.

In that case, even agents who are perfectly informed about the nature of the game and about each other (common knowledge) have no *rational* grounds for tacitly coordinating on one equilibrium rather than on another.

While real and important, the issues of infinite regress and uncertainty are minor compared to a much more elementary, indeed almost trivial question: *How can one impute to real-life agents the capacity to make in real time the calculations that occupy many pages of mathematical appendixes in the leading journals and that can be acquired only through years of professional training?* Actually, the question is even more elementary. The calculations assume that the utilities and subjective probabilities of the agents are well-defined, stable and fine-grained. A large body of psychological literature completely undermines this assumption. Let me just focus on the fineness of grain. To be sure, agents may prefer immediate rewards to delayed rewards, but the assumption of exponential discounting has no psychological reality: it is due only to mathematical convenience. To be sure, they may also have an idea that some outcomes are more likely, perhaps much more likely than others, but the assumption of a continuous probability distribution has no psychological reality. Agents do not have access to such data, and could not draw the proper inferences from them even if they had. *Mental states that do not exist cannot have causal efficacy or enter into explanations.*

Some rational-choice theorists are aware of this problem and try to address it. I shall discuss four possible answers to the italicized question, three of which have actually been proposed.

The first answer – which is mainly hypothetical – is to invoke the precedents of Newton's law of gravitation and of quantum mechanics. Early critics of Newton objected to the law of gravitation that it presupposed the metaphysically absurd notion of action at a distance. Eventually, however, everybody accepted the theory because it worked, with an amazing degree of precision. The even more incomprehensible theory of quantum mechanics, which involves not only action at a distance but objective indeterminacy, is also accepted because its predictions are verified with nine-decimal accuracy. Similarly, in spite of the general objections to rational-choice theory that I have proposed, one might be willing to accept it if its predictions were verified with comparable precision. However, anyone with the slightest acquaintance with economics or political science will dismiss the idea as laughable. Often, scholars are happy if they "get the sign right".

The second and most frequent defense of the explanatory relevance of rational-choice theory would appeal to a causal mechanism capable of *simulating rationality*. Just as economists are fond of arguing that self-interest

can simulate altruism, they often claim that non-intentional mechanisms can simulate intentional optimizing. These mechanisms will generate behavior with utility-maximizing or profit-maximizing consequences even though the agents are incapable of deriving it from an intention to maximize. Generally speaking, there are two mechanisms that might be capable of this feat: reinforcement and selection. The former works by causing given behavioral units to optimize, the latter by eliminating non-optimizing units. As defenders of rational-choice theory rarely if ever appeal to reinforcement, and since the mechanism in any case doesn't simulate optimality very well, I shall ignore it.

Natural selection has of course produced the kind of rough-and-ready and cognitively undemanding rationality that serves us well in everyday life. As an example, consider the Norwegian proverb: "Don't cross the river to the other bank when you go to fetch water". An organism that engaged in such wasteful behavior would quickly be eliminated. There is no reason to believe, however, that natural selection could produce the highly sophisticated strategic behaviors that the models predict. Evolutionary game theory may have some uses, but that of sustaining the models is not one of them.

Models of "economic natural selection" do have some empirical relevance. The work of Nelson and Winter (1984), in particular, shed some qualitative light on economic development. Yet they do not provide the sought-for simulation of rationality, for several reasons. As with agent-based modeling in general, it is often hard to know the extent to which the results are artifacts of the assumptions. Moreover, these results do *not* show optimizing behavior. In a population of firms that evolve by innovation and imitation there is always a substantial proportion of non-optimizing firms. Since firms are adapting to a rapidly changing environment, they are aiming at a moving target. In any case, there is no hope whatsoever that the simulations could mimic the models *all the way down* to the mathematical appendices. Finally, and even more important, bankruptcy-driven or takeover-driven elimination of inefficient agents could never generate optimizing behavior in non-market societies or in non-market sectors in market societies. I conclude that appeal to selection is pure hand-waving.

A third defense is that although cognitively limited agents are liable to make mistakes, these will cancel each other out in the aggregate. If we required each person in a group to carry out calculations of the order of difficulty, say, of multiplying 49 and 73 in at most 30 seconds, we would expect there to be some mistakes, but also that these would be symmetrically distributed around the correct answer. For some purposes, this fact might justify the rationality assumption. (Note, however, that Dalton's famous ox-weight-

guessing experiment used the median rather than the average of individual guesses to arrive at a correct conclusion.) When, however, the answer requires solving differential equations or carrying out other complicated operations, there is no reason to expect answers or guesses to be symmetrically distributed around the correct answer. The burden of proof is on those who claim they will.

Friedman (1953, p. 11-12), finally, offered two seductive analogies to persuade his readers of the reality of maximizing behavior that does not rely on maximizing calculations. First, "leaves [on a tree] are positioned as if each leaf deliberately sought to maximize the amount of sunlight it receives, given the position of its neighbors, as if it knew the physical laws determining the amount of sunlight that would be received in various positions and could move rapidly or instantaneously from any one position to any other desired and unoccupied position". Second, "excellent predictions would be yielded by the hypothesis that the [expert] billiard player made his shots as if he knew the complicated mathematical formulas that would give the optimum directions of travel, could estimate accurately by eye the angles, etc., describing the location of the balls, could make lightning calculations from the formulas, and could then make the balls travel in the direction indicated by the formulas".

While seductive, the analogies are unpersuasive. The leaves simulate maximization because natural selection eliminated trees that didn't. To assume that a similar mechanism exists for economic behavior is to beg the question. Expert billiard players are experts because ten thousand hours of practicing enable them somehow (we don't know how) to make the right shots on an intuitive basis. This is, of course, a tightly constrained situation. To extrapolate the argument to business decisions in a fluid and opaque environment is unwarranted. Nor does the metaphor work for consumer decisions. The only attempt known to me to transform the billiard metaphor into a theory found that "individual learning methods can reliably identify reasonable search rules only if the consumer is able to spend absurdly large amounts of time searching for a good rule" (Allen and Carroll 2001, p. 255). This would be a case of *hyperrationality* (and therefore of irrationality) – searching for the decision that would be optimal if one were to ignore the costs of decision-making itself.

This concludes my discussion of model indeterminacy. I shall spend less space on questions of agent irrationality, which are empirical rather than conceptual. For the sake of brevity, I simply offer a list of irrationality-generating mechanisms (the ones I believe to be the most important are in boldface):

- **loss aversion**
- **hyperbolic discounting**
- **the sunk-cost fallacy** and the planning fallacy (especially deadly in conjunction)
- the tendency of unusual events to trigger stronger emotional reactions
- the cold-hot and hot-cold empathy gaps
- trade-off aversion and ambiguity aversion
- **anchoring** in the elicitation of beliefs and preferences
- **the representativeness and availability heuristics**
- the conjunction and disjunction fallacies
- the certainty effect and the pseudo-certainty effect
- choice bracketing, framing, and mental accounting
- **sensitiveness to changes from a reference point rather than to absolute levels**
- status quo bias and the importance of default options
- meliorizing rather than maximizing
- **motivated reasoning and self-serving biases in judgment**
- flaws of expert judgments and of expert predictions
- self-signaling and magical thinking
- **non-consequentialist and reason-based choice**
- **overconfidence** and the illusion of control
- **spurious pattern-finding**

The list draws heavily on findings from behavioral economics (sources cited in Elster 2009 a). It should be supplemented by the tendency of **emotion** to induce distorted belief formation and low-quality belief formation. These mechanisms all seem to be robust, in the sense of affecting behavior "in the wild" and not only in the artificial context of the laboratory.

Confronted with apparently irrational behavior, economists often try, sometimes successfully, to make sense of them in the rational-choice framework. Some cases of revenge are probably rational reputation-building; some cases of addiction may be rational self-medication; and some suicides may be rational self-euthanasia. At the same time, some revenge-seekers have an urge to act immediately, rather than bide their time to increase the likelihood of success; some addicted smokers persist because of their wishful thinking about the dangers of smoking; and some people commit suicide when they are overwhelmed by shame and the shame causes them to believe, irrationally, that it will endure forever.

3. Hard obscurantism in practice

I shall I present five examples of hard obscurantism in economics and political science:

> Cognitive dissonance theory as used by economists (Akerlof and Dickens, Rabin)
>
> Endogenous patience and altruism (Becker, Mulligan)
>
> Warm-glow theories (Kahneman and Knetsch, Caplan, Andreoni)
>
> Mixed strategies (Dixit and Skeath)
>
> Political transitions (Acemoglu and Robinson)

Cognitive dissonance theory as used by economists

It is not surprising that the theory of cognitive dissonance reduction appeals to economists. It is a *quantitative* phenomenon: the reduction, perhaps even the minimization, of a disutility (dissonance) caused by the coexistence of several mental states or attitudes. The word "cognitive" is somewhat misleading, as the states can be beliefs, desires or, as in the case that originally inspired the theory (Festinger 1957, p. vi), emotions.

George Akerlof and William Dickens (1982, p. 38) argue that workers form motivated beliefs about job safety "according to whether the benefit [of holding the belief] exceed the cost, or vice versa. If the psychological benefit of suppressing one's fear in a particular activity exceeds the cost due to increased chances of accident, the worker will believe the activity to be safe". Similarly, Rabin (1994, p. 178) models "a person's difficulty of maintaining 'false' beliefs with a cost function such that a utility-maximizing person will trade off his preference for feeling good about himself with the cost of maintaining false beliefs". In these two models, costs of holding false beliefs arise from different sources. Akerlof and Dickens refer to the fact that unjustified sanguine beliefs about workplace safety increases the risk of workplace accidents. In Rabin's model, the cost is psychic rather than material, and involves the dissonance that arises when "engaging in immoral activities [that conflict] with our notion of ourselves as moral people" (*ibid*.). He offers the example of wearing furs at the expense of animal suffering.

Akerlof and Dickens (1982, p. 108) explicitly assume that there are no constraints on belief formation: the worker "can believe whatever he chooses irrespective of the information available to him". This flies in the face both of common sense and of the evidence: "people are not at liberty to espouse any attitude they want to: they can do so only within the limits imposed by their prior beliefs" (Kunda 1990, p. 484). The authors recognize that their assumption represents a "polar case", but offer no reasons for thinking that the conclusions generalize to more realistic cases in which belief formation is *constrained* by prior beliefs.

Rabin recognizes that people cannot just adopt *any* belief they would like to have, but he models the obstacles in terms of *costs* rather than constraints. (See Elster 2004 for this distinction.) "I shall assume that the person believes that there is some morally legitimate level of the activity [e.g. wearing furs], Y, such that the person suffers from cognitive dissonance if he chooses level X greater than Y. [...] To capture the difficulty [of believing that Y is high], I let the function C(Y) represent the psychic cost of holding beliefs Y, where C(0) = 0 and C′(y) > for all Y" (Rabin 1994, p. 180-81). In his example, suppose Y=0 is the state in which no animals are killed for the purpose of making furs. For a person wearing furs, the cost of believing that painful killing of animals is morally acceptable is higher than the cost of believing that painless killing of animals is acceptable but that painful killing is not. At the same time, the benefits from holding the former belief are greater than those of holding the latter, since the person who believes that animal suffering is morally acceptable can wear his fur coat with a clear conscience, whereas the person who

can only bring himself to believe that painless killing is acceptable will experience some painful (pun intended) cognitive dissonance. For a given functional form and given parameters of the cost and benefit functions, it might be the case, for instance, that the belief that painless killing is acceptable but that painful killing is not is the one that maximizes the agent's utility.

I shall not dwell on the surreal character of these arguments, but only make some conceptual and empirical objections. In the Akerlof-Dickens argument, the benefits of motivated beliefs come now and the possible costs later. Both enter into the agent's decision to adopt the belief. For that argument to go through, we have to assume that the unconscious is capable of making such intertemporal tradeoffs. There is no evidence that it is. Usually the unconscious is seen as guided by the Pleasure-Principle of seeking immediate satisfaction, whereas the capacity to make intertemporal tradeoffs is characteristic of the conscious mind. It is a conceptual truth that for future benefits or costs to shape present behavior, they must be mentally *made present* (represented) on some mental screen. One can argue, moreover, that consciousness can be *defined* behaviorally by the capacity to represent what is temporally or physically absent. This definition is routinely adopted, for instance, in debates about animal consciousness.

For my purposes here, I need not enter the debate over these complex issues, since I can rest my case on what I believe to be an empirical fact: *there is no evidence for unconscious intertemporal tradeoffs*. The unconscious does not seem to weigh the present benefits of false beliefs against the future costs of holding them. Nor does it seem to weigh the present costs of false beliefs against the future benefits of holding them. If the former trade-off is possible, the latter should be, but there is no evidence for either. Although Winston (1980) made a case for the latter idea, he offered no empirical evidence. Briefly summarized, he argued that if I want to quit using drugs but find that my beliefs about their dangerous effects are insufficiently dissuasive, I should adopt the belief that they are more dangerous than I currently believe they are, since this belief would motivate me to suffer the withdrawal pains. Everything we know about addiction suggests, however, the opposite: addicts persuade themselves that the drug is *less* dangerous than they have reason to think it is.

Both models assume that the agent first identifies the correct belief and then modifies it in a utility-maximizing way. In other words, they assume that the mechanism at work is self-deception rather than wishful thinking. In the latter case, the agent just adopts the belief she would like to be true (assuming no constraining prior beliefs) without confronting it with the evidence that would induce a rational belief. It might indeed be the case that she adopts

by wishful thinking *the very same belief* that she would have formed by considering the evidence, although this could of course only happen by accident. In self-deception, however, it could *never* happen, since it is the very discrepancy between the rational belief and the belief that the agent would like to be true that causes her to adopt the latter. Neither Akerlof and Dickens nor Rabin pay any attention to the huge psychological and philosophical literature on self-deception, part of which denies the very existence of the phenomenon (see the symposium on Mele 1997).

Without stating or it defending it, the authors seem to adopt a model that has been discarded by psychologists and philosophers alike. Their arguments make sense only on the assumption of an *homunculus* - the unconscious as a small inner person capable of behaving like a strategic conscious agent.

Endogenous patience and altruism

In a rational-choice model, preferences are usually seen as *given*, and certainly not as *chosen*. I have just argued that an alcoholic cannot *choose to believe* that drinking is more dangerous than it actually is; nor can she *choose to dislike* alcohol by a mere act of the conscious or unconscious motivational machinery. She can, of course, use indirect strategies, such as taking a drug (Disulfiram) that will make her sick if she drinks, or announcing publicly - to make her incur social disapproval in the case of backsliding - that she has quit drinking.

Such strategies make no sense, however, for ordinary consumption goods. The quip, "I'm glad I don't like spinach, because then I would eat it and I hate the stuff" is laughable precisely because there is no reason not to eat spinach if you like it, and therefore no reason to desire not to desire it. If tomorrow I learn that spinach is strongly cancer-inducing, I shall simply cease to have a first-order desire for spinach, but *not* as the result of a second-order desire not to have the first-order desire. Alcohol, nicotine and other drugs are different. The desire not to desire consuming drugs may be causally inefficacious if it is swamped by cravings or withdrawal symptoms.

I shall call preferences for spinach, alcohol and other consumption goods, in the broadest sense of the term, *material preferences*, and oppose them to *formal preferences*. The latter include *risk attitudes* (risk-preference or risk-aversion) and *impatience* (a preference for earlier reward over later reward). I also include *altruism* and *selfishness* in this category. We may now ask the following question: can second-order preferences over first-order formal preferences be causally efficacious? If I am a risk-lover but desire to be risk-averse, can that desire change my first-order preference? If I have a short time horizon but

desire to be able to defer gratification, will the desire help me to resist temptations? If I am selfish (altruistic) but desire to be altruistic (selfish), will the desire have causal efficacy? A positive answer to these questions would justify the idea of endogenous *and rational* preference formation.

It is probably true that our lives as whole would go better if we were risk-neutral - adopting an attitude of "You win some, you lose some". This fact has inspired the idea of endogenizing risk-attitudes (Palacios-Huerta and Santos 2004). Here, I focus on efforts to endogenize patience (Becker and Mulligan 1997) and, more briefly, on efforts to endogenize altruism (Mulligan 1997). Before I engage with their arguments, I shall explain why the answer is indeed positive in the special case of hyperbolic discounting.

Rational-choice models usually assume that agents discount future rewards exponentially, meaning that there is a constant period-by-period discount rate. For instance, if the agent is indifferent between one unit of utility in the next period and 0.9 unit today, then he is different between one unit of utility in the period after next and 0.81 unit today. This assumption ensures *time-consistency*: if an agent at time t faces the choice between getting a small reward at time t+n and a larger reward at time t+n+m and prefers (at time t) getting the larger delayed reward, he will still prefer the larger reward at time t+n. By contrast, if the agent discounts the future hyperbolically, he will - to simplify - be indifferent between one unit of utility t periods into the future and 1/(1+t) units today. This agent is not time-consistent: well ahead of time he may prefer the larger delayed reward over the earlier smaller reward and then suffer a preference reversal when the time of delivery of the early reward approaches. Mulligan (1996) shows that this behavior is irrational according to a standard money-pump criterion: a person with hyperbolic time discounting could be made to ruin himself by a sequence of stepwise "improvements". This argument does not, of course, invalidate the assumption of hyperbolic time discounting, which is strongly supported by the empirical evidence (Fredericks, Loewenstein and O'Donoghue 2004). Moreover, the money-pump argument presupposes, implausibly, the existence of a "money-pumper" who has full information about the shape of the agent's discounting function.

An agent who discounts the future hyperbolically *and knows it*, can have an incentive to change his discounting function if there is a technology for doing so – a discounting pill, or perhaps psychotherapy (see Elster 2007, p. 211 note 16 for a numerical example). In a calm and reflective moment he wants, let us assume, to choose the larger delayed reward in a certain category of choices, but knows that in each choice he will succumb to tempta-

tion and choose the earlier, smaller reward unless he manages to change his discounting function. This idea is, to be sure, a mere conceptual possibility, which arises only within the artificial framework of economic theory. In reality, self-control problems stem from other sources. I discuss the idea only as a background for my claim, discussed below, that if discounting is exponential a deliberately induced change of preferences is not even a conceptual possibility.

Becker and Mulligan (1997) claim that people can and do choose their rate of (exponential) time discounting in order to improve their lifetime utility. Their basic assumption is that "people have the option to put forth effort to increase their appreciation of the future" and that "[m]ore resources spent on imagination increase the propinquity of future pleasures and therefore their [present] value" (p. 734). For instance, a "person may spend additional time with his aging parents *in order to* appreciate the need for providing for his own old age" (p. 735; my italics). Similarly, because "schooling can communicate images of the situations and difficulties of adult life [...], educated people should be more productive at reducing the remoteness of future pleasures" (p. 735-36). In fact, this *effect* of schooling may also provide the *motivation* to seek higher education: "more patience may be the reason why some people choose to continue their schooling" (p. 751).

In what seems like an independent reinvention of an argument offered by Tocqueville (2004, p. 615), Becker and Mulligan also claim, if I understand them correctly, that if individuals invest in information about the afterlife there might be a *spillover effect* to life on earth: "To the extent that future-oriented capital [due to investment in imagining life in heaven] is 'general' – it facilitates the imagination of events at a variety of distances into the future – a higher utility after death [sic] will even encourage consumption growth *before* death" (p. 741). They add that this effect obtains only for those who believe they will go to heaven, which seems inconsistent with the earlier assertion that their "afterlife [will be] affected by what they do while alive" (p. 740). Surely, in equilibrium the optimal investment in imagination and the optimal amounts of good deeds should be determined simultaneously. Their discussion of this (non-trivial) issue is so brief, however, that it is hard to tell exactly what the argument is.

Once again, I shall not comment on the surreal aspects of some of these assertions, but offer two general arguments against the theory. The first is a purely conceptual objection, whereas the second criticizes the as-if character of their argument. In developing the first objection, I rely on Skog (1997, 2001). I should notify the reader that in the late 1990s I debated these issues

orally and by e-mail with several economists, notably Gary Becker and Peter Diamond. While I did not succeed in persuading them, the reciprocal was also true. (My best answer to their objections is in Elster 2007, pp. 209-11.) My failure to be persuaded may well have been due to my lack of technical competence. The much-regretted Ole-Jørgen Skog did possess that competence, but so far no one of the Becker-Mulligan persuasion has tried to rebut his arguments.

The following argument (Skog 2001, p.211) offers what seems to me to be a knock-down argument against the Becker-Mulligan theory:

> For instance, consider a person with exponential discounting, valuing tomorrow's rewards at 40 per cent of their instantaneous value. He would always prefer one chocolate bar at T = $t + s$ to two chocolate bars at T = $t + s + 1$, whatever the delay s. Suppose that he was offered a pill that would increase his discount factor to 60 per cent. This obviously would induce him to wait for the two bars. But why should the impatient self want to do that? For him one bar with a small delay is better than two bars with a bigger delay. In this example, the myopic actor has no real *motive* for reducing his discount rate (increasing his discount factor). According to his utility function, one chocolate now is the best option.

In my own discussion of the same problem, I suppose that "scientists came up with a discounting pill, which would increase the weight of future rewards in present decisions. If I take the pill, my life will go better. My parents will be happy I took the pill. In retrospect, I will be grateful that I did. But if I have a choice to take the pill or not, I will refuse if I am rational. Any behavior that the pill would induce is already within my reach. I could stop smoking, start exercising or start saving right now, but I don't. Since I do not want to do it, I would not want to take a pill that made me do it" (Elster 2007, p. 210). The investment in more vivid impressions of aging or of the afterlife seems to me exactly analogous to the discounting pill.

Even supposing that Skog and I are mistaken on this point, there is a more elementary objection to the Becker-Mulligan account, viz. their neglect of the distinction between intentions and consequences. It may be true that people who choose higher education learn to value the future more highly, and that as a result their life goes better. These two causal claims provide, however, no evidence that they *intentionally choose* higher education in order to learn to value the future more highly. Becker and Mulligan are simply telling a just-so story.

To use a phrase that can be applied to many other cases, they are engaged in *rational-choice functionalism*. Most functionalist explanations apply to collective behavior. To cite a case that may seem extreme but is actually quite representative, it has been alleged that feuding and vendettas can be explained by their effect of keeping population size at a sustainable level (Boehm 1984). Becker and Mulligan (and many others) apply similar arguments to individual behavior. The fallacy can also be stated as that of *neglecting non-explanatory benefits*.

Of course, the objective consequences of behavior are easier to identify than subjective motivations. That fact does not, however, justify an exclusive focus on the former, any more than the proverbial drunk who had lost his key was justified in looking under the lamppost simply because the light was better there. Explanations of behavior *must* appeal to antecedent mental states. The latter *cannot* be imputed to the agent solely on the basis of the consequences of the behavior. In a Chicago-style reply (Friedman 1953), Becker and Mulligan might counter that they are only testing an implication of their theory, and that the realism of the motivational assumptions is irrelevant. I agree, however, with Gibbard and Varian (1978, p. 671) when they say that "On [our] reading of Friedman, when a model is applied to a situation, all that is hypothesized is that the conclusions of the applied model are close enough to the truth for the purpose at hand. According to us, something further is hypothesized: that the conclusions are sufficiently close to the truth *because* the assumptions are sufficiently close to the truth". In the case at hand, the assumptions, such as visiting aging parents *in order to* form a more vivid impression of what aging is like, seem very far-fetched.

Finally, we may extend the analysis to the idea of "endogenous altruism" proposed in Mulligan (1997). This kind of extension from an intrapersonal case to the corresponding interpersonal case – from "future selves" to "other selves" – is often tempting and sometimes useful. In the present case, the extension fails for the same reasons that explain failure in the intrapersonal case.

Mulligan (1997, p. 73) argues that "parental actions affect their [sic] willingness to sacrifice their own consumption for consumption by their child. Parents are aware of the effect of those actions on their 'preferences' and take those effect into account when determining what actions to take". More specifically, a "parent's concern for a child's consumption is assumed to depend on the quantity of resources – mainly time and effort – directed to the accumulation of concern. [...] People may naturally tend to be selfish, but parents *may also spend time and effort in self-reflection to overcome such a natural bias*"

(p. 77).

Given what I said about endogenizing time preference, my response to the statement I have italicized is obvious: *why* would selfish parents want to become less selfish? If they want to, aren't they already unselfish? Hence I agree with the following common-sense objection in a review of Mulligan's book:

> If parents in some sense want to behave more altruistically than they would with zero expenditures on child-oriented resources, why do they not simply shift more resources to their children, perhaps by investing more in their human capital? If they start facing a marginal tradeoff of 1.50 for their own versus their children's consumption but want to be more altruistic, why don't they simply shift consumption from themselves to their children to change this tradeoff rather than divert resources away from both generations' consumption by using them for child-oriented resources? (Behrman 1998, p. 1508)

In addition, again analogously to the time preference case, I would ask: *where is the evidence* that parents intentionally behave in this way? Mulligan's response (p. 123) can only be characterized as lame: "The effect of child-oriented resources is mechanical and well understood by parents in my model, but a precise understanding by parents is not necessary for parents to willingly purchase child-oriented resources and for my results to obtain. Advertising is an example where the effect of something on preferences is modeled as well-known, but those models clearly provide insights into the 'real world' where advertising has effects that are not always predictable."

At the very least, I agree with Bowles (1998, p. 80) when he writes that "We know that [in preference formation] intentional motivations are sometimes involved; one learns to appreciate classical music because one notices that aficionados appear to enjoy it [. . .]. But instrumental motivations may be of limited importance compared to other influences", such as conformism or, for that matter, anti-conformism. I believe the case should be put more strongly. For instrumental motivations to matter, they must be empirically demonstrable on a scale large enough to have socially important consequences. The occasional anecdote about visiting aging parents or wishing to learn to appreciate classical music is too much like the made-up examples that have brought parts of analytical philosophy into discredit.

Warm-glow theories

Rational-choice models often assume that agents are *selfish*, in the sense of being motivated by material, usually monetary gains. Many critics and some defenders of rational-choice theory mistakenly assume that rationality implies egoistic selfishness. Although there is no logical connection, there may be a sociological one. In the words of Frank (1988, p. 21), "[t]he flint-eyed researcher fears no greater humiliation than to have called some action altruistic, only to have a more sophisticated colleague later demonstrate that it was self-serving".

Some actions, however, are recalcitrant to such demonstrations. When people vote in national elections with secret ballot, do volunteer work to preserve the environment or give money to Oxfam, material self-interest will rarely if ever be their motivation. Faced with this challenge, the flint-eyed economist can appeal to *egocentricity* rather than to material egoism (Elster 2009 b). Egocentric motivations include egoistic ones, but also the desire for *approval by an audience*. Vanity causes us to seek the approval of an external audience; amour-propre to seek the approval of an internal audience. According to Kant (1996, p. 61-62), it can in fact "never be inferred with certainty that no covert impulse of self-love, under the pretense of [duty], was not actually the real determining cause of the will".

Kant did not make a positive claim about the motivation for any specific actions. Some flint-eyed economists do, when they explain good-doing by the "warm-glow" from doing good – the applause of the internal audience. I shall examine three arguments along these lines, two of them briefly and the third at greater length.

Voting in national elections with secret ballot poses two puzzles. First, why would a selfish and rational individual bother to vote at all? Second, how can we explain the well-supported empirical fact that voters by and large vote "sociotropically", to promote the public interest rather than their own? Caplan (2007, p. 151) addresses the second issue, and argue that even when people vote for the public good rather than their private interest, they do so to "enhance their self-image". Because a single vote has essentially zero effect on the outcome, affluent voters can afford to *buy altruism* at low or no cost, by voting for redistributive policies that would harm them if they were to be implemented. Although Caplan does not address the first puzzle, I conjecture he would offer the same explanation.

Kahneman and Knetsch (1992) address the question of people's willingness to pay for public goods. In the first part of an experimental study, they asked subjects how much they would pay for environmental services. Three

groups of subjects responded to questions formulated at different levels of inclusiveness, ranging from "improving environmental services" to "improve the availability of equipment and trained personnel for rescue operations". As the last improvement is a small subset of the first, one might expect subjects to state willingness to pay much larger amounts for the more inclusive measures. Contrary to expectations, the amounts were roughly similar. To explain this finding, they carried out a second experiment to test the hypothesis that "the moral satisfaction associated with an inclusive cause extends with little loss to any significant subset of that cause" (p. 64). The hypothesis was confirmed, suggesting that subjects were more concerned with feeling good about themselves than with doing good.

Andreoni (1990, 2006) argues that philanthropy, too, must be explained by the desire of donors to obtain a warm glow rather than by their desire help recipients. He assumes that donations must be in a Nash equilibrium, in which each citizen donates an amount that is optimal given the donations of everybody else. Since the welfare of the recipients is a public good for the donors, however, they have a collective action problem. In the words of Andreoni (1990, p. 465), one can deduce from the assumption of rational equilibrium behavior that "in large economies virtually no one gives to the public good, hence making the Red Cross, the Salvation Army and American Public Broadcasting logical impossibilities". However, if donations are motivated by the fact that they are good for the donor rather than for the recipients, they produce a private good rather than a public good. Since the donor internalizes all the benefits from his gift, there is no collective-action problem.

A succinct statement of the warm-glow effect in the context of a public good experiment, in which each member of a group had the opportunity to benefit other members, is that "the *act* of contributing, independent of how much it increases group payoffs, increase a subject's utility by a fixed amount" (Palfrey and Prisbrey 1997, p. 830). Psychologically, this seems implausible. Since the warm glow is supposed to come from doing good, it is presumably enhanced by doing more good rather than less. A donation that's *known to be pointless* cannot produce a warm glow any more than a *costless donation* can. The greater the benefit to others and the greater the cost to oneself, the warmer the glow. In practice, and perhaps in principle, it would be hard to distinguish between the enhanced welfare of others as the *altruistic goal* of the donor and its role as a *condition for achieving his egocentric goal*. In the regression equation of Palfrey and Prisbrey (1997), which uses as variables the egoistic cost to the agent, his egocentric benefits and his altruistic benefits, the last is found to be "not significantly different from zero". My reaction to their claim is one

of frank disbelief, not because I can identify any error in their analysis but because this is not how human beings are and behave. Even if they ultimately care only about the warm-glow, they cannot get it unless they believe that they do good for others, and *the more good they believe they do the warmer is the glow*.

As further evidence of the lack of intellectual sophistication by economists in this field, one may cite the following argument that Andreoni (2006, p. 1220 n. 16) makes for his theory: "the fact that people do get a joy from giving is such a natural observation as to be nearly beyond question". Yes, but as the neuroeconomists know (de Quervain et al. 2004), getting pleasure from doing X and going X for the sake of the pleasure are two entirely different things.

As the point is fundamental, it may be worthwhile mentioning that it has a venerably ancestry. Thus in *On the Happy Life* (9.1-2), Seneca affirms that

> In the first place, even though virtue is sure to bestow pleasure, it is not for this reason that virtue is sought; for it is not this, but something more than this that she bestows, nor does she labor for this, but her labor, while directed toward something else, achieves this also. As in a ploughed field, which has been broken up for corn, some flowers will spring up here and there, yet it was not for these poor little plants, although they may please the eye, that so much toil was expended - the sower had a different purpose, these were superadded - just so pleasure is neither the cause nor the reward of virtue, but its by-product, and we do not accept virtue because she delights us, but if we accept her, she also delights us.

I am unable to propose a positive theory of why people donate to good causes. They probably have all sorts of reasons, in proportions and combinations that may defy precise analysis. The operation of social norms – giving because one is observed by others - is relatively uncontroversial. That of "quasi-moral norms" (Elster 2007, Ch. 4) – giving because one observes others giving - also seems plausible, especially perhaps in the wake of natural disasters. I would not exclude irrational phenomena such as magical thinking, but evidence might be hard to find. Targeted altruism ("adopt a child") is observed in many contexts. Donations to the Salvation Army or to the Red Cross may be expressions of untargeted altruism, because many *donors do not go through the Nash equilibrium reasoning*. In fact, the assumption that they do is highly implausible, since they would need information about the donation behavior of others that most people almost certainly do not have. Warm-glow motivations probably also have a role to play, although for the reasons indicated above they may be hard to separate from altruism.

The assumption that agents are *rational* and *self-interested* is both parsimonious and capable of yielding sharp predictions. Yet when the predictions of the model fail, something has to give. Generally speaking, the natural reaction to an explanatory failure is to try to explain the observed facts by departing as little as possible from the original model. In the case that concerns me here, the smallest deviation from *rational egoism* might seem to be that of *rational egocentricity*. In this model, the utility function of the typical agent would include both her own material benefit and the degree to which she can think of herself as a moral agent. As in other cases, there would be a trade-off between these aims: she might be willing to sacrifice some material welfare to get the warm glow from the enhanced self-image.

Warm-glow theorists of philanthropy and similar non-selfish actions (voting, preserving the environment, etc.) do not seem to realize that this small adjustment to the model, *substituting egocentricity for egoism*, requires another and more radical one: *substituting irrationality for rationality*. Specifically, agents have to deceive themselves about their own motives. This substitution is required by what I take to be a conceptual truth: one cannot derive a warm glow from an action unless the agent believes that the action was performed at least in part to benefit others. *An egocentric agent who performs for the inner audience has to believe that she is altruistic.* An agent who performed "good actions" only for the *conscious end* of enhancing her self-image could not achieve that aim, any more than one can enhance one's self-image by paying another person to praise oneself.

Thus a common denominator of the arguments offered by Caplan, Kahneman and Knetsch, and Andreoni is the assumption of self-deceiving agents. I have no empirical objection to the assumption: there is no doubt that we often deceive ourselves about our motives. Envy may be transmuted into righteous indignation or guilt into anger ("who has offended cannot forgive"). These cases differ, though, from warm-glow cases. Transmutation is a form of dissonance-reduction, which is triggered by the urge to escape from a negatively valued emotion. The warm-glow effect is triggered by the search for a positively valued emotion: there is no dissonance to be reduced. It is not clear, however, how much one has to do to produce the warm glow. Surely, the act of giving a penny to the panhandler in the street isn't enough. Similarly, I find it hard to believe that the *costless* response to a question about how much one is willing to pay for the provision of a public good could create a worm glow. The assumption that people act to create moral satisfaction may predict what we observe, but the assumption must also have independent credibility.

Mixed strategies

Sometimes, individuals or groups make decisions by using a randomizing device, when it is impossible or prohibitively expensive to gather enough information to make a determinate rational choice. We may flip a coin, for instance, to determine who is going to clean up after dinner. To select the foreperson of a jury, one can put the names of all jurors in a hat and select one at random. Ordinary decisions rarely if ever use devices that deviate from equiprobability, meaning that each option or candidate has an equal chance of being chosen.

In other cases, an agent may randomize, possibly deviating from equiprobability, to prevent others from anticipating or guessing her decision. In poker, rational-choice theory may for instance tell the agent to bluff one third of the times she's dealt a Queen. In military affairs, a rational military commander may use a randomizing device to select strategies for fighter pilots in dog fights (Luce and Raiffa 1957, p. 76). These are zero-sum games. For the reasons noted above, there is no similarly plausible rationale for randomizing in non-zero sum games (see also *ibid*. p. 93).

Yet even in non-zero-sum cases, sometimes agents act *as if* they randomize. Consider the case of Kitty Genovese. In the standard presentation, this young woman was stabbed to death in 1964 in the presence of 38 neighbors who heard her cries, none of whom called the police. Although this account is probably apocryphal (see Manning, Levine and Collins 2007), it has been claimed that there is also experimental evidence for bystander passivity (Darley and Latane 1968). Some experiments also suggest that the more bystanders there are, the less is the chance that *any* of them will intervene. There is no safety in numbers. I need not consider the empirical veracity of these claims, since my only purpose here is to consider the explanation of bystander passivity offered in a highly regarded textbook on game theory (Dixit and Skeath 2004, p. 416).

Citing the Kitty Genovese case, these authors argue that one may try to justify the idea of mixed strategies by appealing to a causal mechanism: "[M]ixed strategies are quite appealing in this context. The people are isolated, and each is trying to guess what others will do. Each is thinking, Perhaps I should call the police ... but maybe someone else does ... but what if they don't? Each breaks off this process at some point and does the last thing that he thought of in this chain, but we have no good way of predicting what that last thing is. A mixed strategy carries the flavor of this idea of a chain of guesswork being broken by a random point." So far, so good. The authors then go on, however, to commit a simple quantifier fallacy: from the correct premise that for every

person there is a probability **p** that he will not act, they reach the false conclusion that there is a **p** such that each person will abstain from acting with probability **p**. Moreover – a second unjustified step – they assume that when all abstain from acting with probability **p**, their choices form an equilibrium.

In this example, the probabilities completely lack microfoundations. The authors give no reason why all subjects should come up with the particular probability that has the property of generating an equilibrium. *A blind causal process cannot mimic a conscious strategic choice.* When Harsanyi (1977, p. 114) offered a similar argument, he limited his claim to equiprobabilistic mixed strategies, "generated [. . .] by what amounts to an unconscious chance mechanism inside [the player's] central nervous system". Dixit and Skeath (2004, p. 417 note), by contrast, are willing to entertain the idea that when the benefits and costs from calling the police are 10 and 8 respectively, the equilibrium probability that a given person in a group of 100 will *not* call is 0.998. In their model, "increasing [the size of the group] from 2 to infinity leads to an increase in the probability that not even one person helps from 0.64 to 0.8". Yet although the model has the seemingly attractive feature of predicting the counterintuitive fact that there is no safety in numbers, the argument is undermined by the absurdity of the assumptions. Their stylized explanation is nothing more than a "just-so" story.

Political transitions

In a number of acclaimed publications, Daron Acemoglu and James Robinson have tried to provide rational-choice foundations for the transition to democracy. I shall focus on one of their articles (Acemoglu and Robinson 2001), published in the flagship journal of the American Economic Association. Since I have not read all their other writings on the subject, it is possible that they have already addressed and countered the objections I shall make. Even if this were so, the relevant fact for my purposes is that *an obscurantist article was published in the most prestigious journal of the profession*, not that two individual scholars made mistakes that they may or may not have corrected later.

As a preliminary comment, let me say that I find the article a monumental expression of explanatory hubris. The idea that one could explain transitions to democracy in West European and Latin American countries by a game-theoretic model involving only two actors, "the rich" and "the poor", is too far-fetched to be taken seriously – except for the present polemical purposes.

I shall not try to address all the issues discussed in the article, but only comment on the basic conceptual framework and its empirical support or lack thereof. As noted, they reduce the question of class struggle to the conflict between rich and poor, thus neglecting, for instance, possible conflicts of interest between poor peasants and poor urban workers. The former have an interest in high prices on food products, the latter an interest in low prices, a phenomenon that mirrors the conflict between landowners and industrialist capitalists in the 19^{th} century England. I shall not pursue this issue further, but take the two-class model for given.

Acemoglu and Robinson assume that all agents have identical preferences, differing only in their capital endowments. All poor agents are assumed to have the same endowments, as do all the rich. Having already swallowed the two-class assumption, I shall swallow these simplifications as well. I am not equally willing, however, to accept the assumption that agents discount the future exponentially. Although mathematically convenient and seemingly justified by the axiom that agents are rational, the assumption has "little empirical support" (Frederick, Loewenstein and O'Donoghue 2004, p. 210). To adopt it without trying to justify it or defend it against criticism, which was surely well-known to the authors, is a cavalier procedure.

Compared to other issues, the assumption of exponential discounting is nevertheless a minor problem. A more troublesome issue is the idea that in any given period aggregative productivity A can be modeled by assuming that A is either high with probability (1-s) or low with probability **s**. I shall ignore the starkly dichotomous and starkly implausible character of the assumption and focus instead on its interpretation. When Acemoglu and Robinson assert (p. 940) that that the level of income is "stochastic", they are presumably using the term in the dictionary sense of a process involving the operation of chance, such as the onset of cancer. Although scientists may today be able to quantify the probability that a given person will develop a given kind of cancer in a given period, the person herself may not – and a hundred years ago certainly could not - have any idea about the magnitude of the risk. By contrast, Acemoglu and Robinson (p. 944, Equation 4) impute knowledge of the value of **s** *to the agents*, in order to calculate the "discounted expected net present value [...] of a poor agent after the revolution but before the state A is revealed". They would presumably defend this imputation by some version of the theory of rational expectations. Whatever the defense they might offer, the imputation is indefensible. The idea that, say, the rural poor in France in 1789 or the urban poor in 1848 attached a sharp probability to *aggregate* productivity being high or low is a piece of science-fiction.

For the game-theoretic model of revolution to get off the ground, each class - the rich and the poor – must be viewed as a *unitary actor*. Acemoglu and Robinson are of course aware of the problem of free riding in revolutionary situations, but claim that "[b]ecause a revolution generates private benefits for a poor agent, there is no collective action problem" (p. 941). In a footnote they add that

> Although a revolution that changes the political system might seem to have public-good like features, the existing empirical literature substantiates the assumption that revolutionary leaders concentrate on providing private goods to potential revolutionaries (see Gordon Tullock 1971). There could also be a coordination problem in which all poor agents expect others not to take part in a revolution, so do not take part themselves. However, since taking part in a revolution imposes no additional costs irrespective of whether it succeeds or not, it is a weakly dominant strategy.

The reference to Tullock (1971) is strange. *Tullock does not offer or cite any empirical evidence* concerning actual revolutions. In fact, he does not cite a single empirical study of any revolution. Tullock merely asserts that his "*impression* is that [revolutionaries] generally expect to have a good position in the new state which is to be established by the revolution. Further, my *impression* is that the leaders of revolutions continuously encourage their followers in such views" (p. 98; my italics). To cite this armchair speculation as a decisive piece of "empirical literature" is to offer very weak support, in fact, no support at all. Instead, Acemoglu and Robinson should have cited primary empirical sources. There is a vast literature on the motivations of revolutionary leaders and followers, some of it supporting the idea of private benefits and some of it, perhaps the greater part, contradicting it. The French peasants who triggered the abolition of feudalism on August 4 1789 by burning the castles of nobles in the second half of July were not motivated by the desire to achieve leading positions in a post-revolutionary society, nor were the East Germans who mobilized in the streets of Leipzig in October 1989. My aim, however, is not to offer a counter-generalization to the reckless generalization of Acemoglu and Robinson, only to point out its recklessness.

The final sentence in the quoted passage is so opaque that I shall not comment on it, except to remark that it *seems* to ignore the vast costs and risks often incurred by revolutionaries. Instead I shall conclude by a brief comment on the formal analysis of *counterrevolution* proposed by Acemoglu and Robinson.

As they state, their "formalization implies that, as with a revolution, there is no free-rider problem with a coup" (p. 942). Again, the empirical evidence is relegated to a footnote: "This seems plausible. For example, in Venezuela in 1948, Guatemala in 1954, and Chile in 1973, landowners were rewarded for supporting the coup by having the land returned to them". They provide no references to support this claim. I know little about Latin America, but a brief literature search on Chile suggests that for this country at least their claim may be incorrect.

The claim seems to presuppose a selective political motivation for land restitution: landowners who supported the coup, but only those who supported it, got their land back. It also presupposes that in lending their support to the coup, landowners anticipated and were motivated by these selective rewards. As I observed in the discussion of Becker and Mulligan, ex post consequences cannot substitute for ex ante intentions. Since I do not know of any evidence about intentions, however, I can only refer to consequences. The extensive discussion of land restitution after 1973 in Bellisario (2007) does not cite any selective political criteria. By contrast, the "individual allotment of parcels to a portion of the asentados [beneficiaries] of Agrarian Reform" did discriminate against those "who had participated in land takeovers and in other political 'crimes' before the military coup" (Silva 1991, pp. 22, 26). In other words, supporters of the coup were not selectively rewarded, but some beneficiaries of the Allende regime were selectively punished. Needless to say, I cannot vouch for the accuracy of these analyses, which are the fruit of half an hour's search on the Internet. I cite them only to indicate the *kind* of fine-grained analysis that would be needed to make good on the claim.

4. Summary

Although I believe that the cases I have selected for analysis are somewhat representative of mainstream economic theorizing, I cannot make strong claims about *how* typical they are. What I can assert with great confidence is that the authors I have singled out are far from marginal, and in fact are at the core of the profession. Their numerous awards testify to this fact.

These writings have in common a somewhat uncanny combination of *mathematical sophistication* on the one hand and *conceptual naiveté* and *empirical sloppiness* on the other. The mathematics, which could have been a tool, is little more than toy. The steam engine was invented by Hero of Alexandria in the first century A. D., but he considered it mainly as a toy, not as a tool that could

be put to productive use. He did apparently use it, though, for opening temple doors, so his engine wasn't completely idling. Hard obscurantist models, too, may have some value as tools, but mostly they are toys.

I have pointed to the following objectionable practices:

1. Citing empirical evidence in a cavalier way, in the form of anecdotes, "impressions", and unsubstantiated historical claims (Becker and Mulligan, Acemoglu and Robinson).

2. Adopting huge simplifications that make the empirical relevance of the results essentially nil (Acemoglu and Robinson).

3. Assuming that the probabilities in a stochastic process are known to the agents (Acemoglu and Robinson) or even in some sense optimal (Dixit and Skeath).

4. Assuming that intentions can be inferred from outcomes (Becker and Mulligan, Kahneman and Knetsch).

5. Assuming that the unconscious has the capacity to carry out intertemporal tradeoffs (Akerlof and Dickens).

6. Assuming that agents can choose optimal beliefs on the basis of the consequences of having them rather than on the basis of the evidence supporting them (Rabin, Akerlof and Dickens).

7. Assuming that agents can choose optimal preferences (Becker, Mulligan).

8. Ignoring well-established facts such as hyperbolic discounting or limited cognitive capacities (Acemoglu and Robinson).

9. Assuming self-deceiving agents (Rabin, Andreoni, Caplan, Kahneman and Knetsch), without engaging in the literature on this controversial subject.

10. Assuming that agents can enhance their self-image by taking trivial, even costless altruistic actions (Andreoni, Caplan, Kahneman and Knetsch).

11. Adhering to the instrumental Chicago-style philosophy of explanation, which emphasizes as-if rationality and denies that the realism of assumptions is a relevant issue.

These features do not, of course, amount to sufficient and/or necessary conditions. If I were to single out one cluster of issues that seem to be the most important, I would mention (1), (2), (4) and (11). Whereas the other issues are problem-specific, these four questions seem to be more recurrent. Many of the other issues have a common root, which is the neglect of elementary conceptual analysis. The mixed-strategy case is perhaps the best example. Also, the obsession with optimization operates across the board.

5. References

Acemoglu, D. and Robinson, J. (2001), "A theory of political transitions", *American Economic Review* 91, 938-63.

Akerlof, G. and Dickens, W. (1982), "The economic consequences of cognitive dissonance", *American Economic Review* 72, 307-19.

Allen, T. and Carroll, C. (2001), "Individual learning about consumption", *Macroeconomic Dynamics* 5, 255-71.

Andreoni, J. (1990), "Impure altruism and donations to public goods: A theory of warm-glow giving", *Economic Journal* 100, 464-477.

Andreoni, J. (2006), "Philanthropy", in S.-C. Kolm and J. M. Ythier (eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity*, vol. II, Amsterdam: North-Holland, pp. 1202-69.

Arrow, K. and Hurwicz, L. (1971), "An optimality criterion for decision-making under uncertainty", in C. F. Carter and J. L. Ford (eds.), *Uncertainty and Expectation in Economics*, Clifton NJ: Kelley, pp. 1-11.

Becker, G. and Mulligan, C. (1997), "The endogenous determination of time preference", *Quarterly Journal of Economics* 112, 729-58.

Behrman, J. (1998), Review of Mulligan (1997), *Journal of Economic Literature* 36, 1508-9.

Bellisario, A. (2007), "The Chilean agrarian transformation: Part 2", *Journal of Agrarian Change* 7, 145-82.

Bowles, S. (1998), "Endogenous preferences", *Journal of Economic Literature* 36, 75-111.

Caplan, B. (2007), *The Myth of the Rational Voter*, Princeton University Press.

Cohen, G. A. (2002) "Deeper into bullshit," in S Ross and L Overton (eds.) *Contours of Agency. Essays on Themes from Harry Frankfurt*, Cambridge, MA: MIT Press, pp. 321-39.

Darley, B. and Latane, J. (1968), "Group inhibition of bystander intervention in emergencies", *Journal of Personality and Social Psychology* 10, 215-21.

Dixit, A. and Skeath, S. (2004), *Games of Strategy,* New York: Norton.

Elster, J. (2004), "Costs and constraints in the economy of the mind", in I. Brocas and J. Carillo (eds.), *The Psychology of Economic Decisions*, vol. 2, Oxford University Press, pp. 3-14.

Elster, J. (2007), *Explaining Social Behavior*, Cambridge University Press.

Elster, J. (2009 a), "Excessive ambitions", *Capitalism and Society* 4, 1-30.

Elster, J. (2009 b), *Le désintéressement,* Paris: Seuil.

Elster, J. (2012), "Hard and soft obscurantism in the humanities and social sciences", *Diogenes* 58, 159-70.

Festinger, L. (1957), *A Theory of Cognitive Dissonance*, Stanford University Press.

Frank, R. (1988), *Passions within Reason*, New York: Norton.

Frankfurt, H. (1988), "On bullshit", in H. Frankfurt (ed.), *The Importance of What we Care About,* Cambridge University Press, pp. 117-33.

Fredericks, S., Loewenstein, G. and O'Donoghue, T. (2004), "Time discounting and time preference", in C. Camerer, G. Loewenstein and M. Rabin (eds.), *Advances in Behavioral Economics*, New York: Russell Sage, pp.162-222.

Freedman, D. (2005), *Statistical Models*, Cambridge University Press

Freedman, D. (2010), *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, Cambridge University Press

Friedman, M. (1953), *Essays in Positive Economics*, University of Chicago Press.

Gibbard, A. and Varian, H. (1978), "Economic models", *Journal of Philosophy* 25, 664-77.

Gjelsvik, O. (2006), "Bullshit illuminated", in J. Elster et al. (eds.), *Understanding Choice, Explaining Behaviour*, Oslo Academic Press, pp. 101-11.

Green, D. and Shapiro, I, (1994), *Pathologies of Rational Choice Theory*, Cambridge University Press

Harsanyi, J. (1977), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press.

Johansen, L. (1977), *Lectures on Macroeconomic Planning*, Amsterdam: North-Holland.

Kahneman, D. and Knetsch, J. (1992), "Valuing public goods: The purchase of moral satisfaction", *Journal of Environmental Economics and Management* 22, 57-70.

Kant, I. (1996), *Groundwork of the Metaphysics of Morals*, in Immanuel Kant, *Practical Philosophy*, Cambridge University Press, pp. 37-108.

Kunda, Z. (1990), "The case for motivated reasoning", *Psychological Bulletin* 108, 480-98.

Luce, R. and Raiffa, H. (1957), *Games and Decisions*, New York: Wiley.

Manning, R., Levine, M. and Collins, A. (2007), "The Kitty Genovese murder and the social psychology of helping", *American Psychologist* 62, 555-62.

Mele, A. (1997), "Real self-deception", *Behavioral and Brain Sciences* 20, 91–136.

Mulligan, C. (1996), « A logical economist's argument against hyperbolic discounting », at http://home.uchicago.edu/cbm4/hyplogic.pdf.

Mulligan, C. (1997), *Parental Priorities and Economic Inequality*, University of Chicago Press.

Nelson, R. and Winter, S. (1984), *An Evolutionary Theory of Economic Change*, Cambridge, MA: Harvard University Press.

Palacios-Huerta, I. and Santos. T. (2004), "A theory of markets, institutions, and endogenous preferences", *Journal of Public Economics* 88, 601-27.

Palfrey, T. and Prisbrey, T. (1997), "Anomalous behavior in public goods experiments: How much and why?", *American Economic Review* 87, 829-46.

Quervain, J. F. de et al. (2004), "The neural basis of altruistic punishment", *Science* 305, 1254-58.

Rabin, M. (1994), "Cognitive dissonance and social change", Journal of *Economic Behavior and Organization* 23, 177-94.

Schelling, T. (1971), "Dynamic models of segregation", *Journal of Mathematical Sociology* 1, 143-86.

Silva, P. (1991), "The military regime and restructuring of land tenure", *Latin American Perspectives* 18, 15-32.

Skog, O. (1997), "The strength of weak will", *Rationality and Society* 9, 245-71.

Skog, O. (2001), "Theorizing about patience formation", *Economics and Philosophy* 17, 207-19.

Taleb, N. (2007), "Black swans and the domain of statistics", *The American Statistician* 61, 198-200.

Tocqueville, A. de (2004), *Democracy in America*, New York: Library of America.

Tullock, G. (1971), "The paradox of revolution", *Public Choice* 11, 89-99.

Weitzman, M. (2009), "On modeling and interpreting the economics of catastrophic climate change", *Review of Economics and Statistics* 91, 1-19.

Winston, G. (1980), "Addiction and backsliding", *Journal of Economic Behavior and Organization* 1, 295-324.

Winter, S. (1964), "Economic 'natural selection' and the theory of the firm", *Yale Economic Papers* 4, 225-72.

52

Philosophy in a Dark Time: Martin Heidegger and the Third Reich

TIMOTHY O'HAGAN

Like Oscar Wilde I can resist everything except temptation. So when I received Anne Meylan's tempting invitation to contribute to this *Festschrift* for Pascal Engel I accepted without hesitation, before I had time to think whether I had anything for the occasion. Finally I suggested to Anne the text of a public lecture which I delivered in 2008 and which I had shown to Pascal, who responded to it with his customary enthusiasm and barrage of papers of his own on similar topics. But when I re-read it, I realized that it had been written for the general public rather than the professional philosophers who would be likely to read this collection of essays. So what was I to do with it? I've decided to present it in two parts. In Part One I reproduce the original lecture, unchanged except for a few minor corrections. In Part Two I engage with a tiny fraction of the vast secondary literature which has built up over the years and which shows no sign of abating.

1. Part One: The 2008 Lecture

Curtain-Raiser

Let us start with two dates, 1927 and 1933. In 1927 Adolf Hitler's *Mein Kampf* (volume II) was published. So too was Martin Heidegger's magnum opus *Being and Time*. In 1933 two appointments were made: Hitler as Chancellor of the German Reich and Heidegger as Rector of Freiburg University. In 1927 it was a case of sheer coincidence; in 1933 the two events were closely linked.

Heidegger's life up to 1933

First, a brief sketch of Martin Heidegger's life up to 1933. He was born in 1889 in Messkirch, a small provincial town in the Baden. His father was a cooper and church sacristan in this reactionary, deeply Catholic backwater. Throughout his life Heidegger sustained a self-image of the provincial outsider, even peasant, within the cultural élite of Germany.[1] Heidegger went as a scholarship boy to high schools in Konstanz and then Freiburg. This was followed by studies at Freiburg University, from 1909-11 as a student for the Catholic priesthood in the theology faculty. But he lost his faith and soon turned against the Catholic Church. He pursued his studies in philosophy and science, gaining his doctorate in 1913 and his *Habilitation* in 1916.

In the Great War Heidegger served in the German army as a postal censor and as a weather forecaster. He married the Protestant Elfride Petri in 1917. He taught at Marburg University from 1926-8, when he was appointed to the Chair of Philosophy at Freiburg University, as successor to the preeminent phenomenologist Edmund Husserl, who had held the post since 1916.)

Following Hitler's seizure of power, Heidegger was elected Rector of Freiburg University on 21 April 1933 after the previous Rector, medical Professor and Social Democrat von Möllendorf, was removed by the Baden Ministry of Education.

The work which ensured Heidegger's appointment to the Freiburg Chair was *Being and Time*. I'll now highlight some specific themes of that book which will emerge, brutally transformed, in the notorious Rectoral address of 1933.

Being and Time: Context and reception

The appearance of Martin Heidegger's *Sein und Zeit* (*Being and Time*) in 1927 would rock the philosophical world to its foundations. It announced the end of centuries of speculation concerning every philosophical conundrum, from the mind-body problem to the nature of language and truth, from skepticism in epistemology to nihilism in morality. Heidegger's strategy was not so much to *solve* these problems as to *destroy* them. He did this by calling for a radical

[1]. We learn from Farias that Messkirch was a site of great tension between Catholics and Old Catholics: the former were conservative and authoritarian, while the latter were more progressive and sympathetic to liberalism and enlightenment; they also tended to be relatively rich and priviliged. Thus Heidegger, poor and Catholic, had his future ideological enemies formed for him from the cradle. A classic case of Nietzschean *ressentiment*, one might say.

shift of focus which would allow us to look with fresh eyes on fundamental questions about our own being, questions which had been neglected since the presocratic Greek thinkers, and had subsequently been buried by generations of metaphysicians. In *Being and Time* Heidegger coined a bizarre philosophical jargon, partly idiomatic, partly scholastic.

Yet somehow the powerful critical thrust of Heidegger's message got through, first to his own pupils and then farther afield, so that even as commonsensical a philosopher as Gilbert Ryle recognized the importance of *Being and Time* in his review in *Mind*, although he expressed forebodings about the eventual outcome of Heidegger's thought. As we shall see in a moment, Ryle's anxiety was well-founded.

Being and Time: Phenomenology meets existentialism

Being and Time carries the Dedication "To Edmund Husserl, in friendship and admiration" and at least in the first half it is true to the phenomenological programme of patiently describing, ever more deeply, the world we encounter, until, by stripping away all received assumptions, we reach its fundamental nature. In his early work Husserl had bracketed off the traditional idea of a substantial self as bearer of consciousness, leaving only consciousness itself and its stream of intentional objects. He was soon to retreat from that ultra-radical position and reinstate the subject of consciousness.

Heidegger rejected Husserl's revisionist move, holding that any philosophy which starts with a self existing in isolation from its world is doomed to end in solipsism. In Heidegger's new version of phenomenology, Dasein (human existence) and its world are given as a whole. The phenomenological task now becomes that of describing the essential character of that world as a whole, the world as it discloses itself to us, infused with meaning, irreducibly temporal. It is accessible to us primarily as creatures with projects and concerns, only derivatively as theorists and scientists.

Even truth becomes a process of disclosure. Thus in searching for truth our goal is not an increasingly accurate correspondence between an explanatory model and a set of phenomena. Instead we are seeking an authentic way of being, such that the world discloses itself to us as it truly is.

As the focus shifts to authentic being in Division II of the book, phenomenology begins to fuse with existentialism. The influence of Kierkegaard is felt and the tone becomes more declamatory. Forgetful of being, we lose ourselves in "idle talk", in the fog of cliché and catch-phrase which conceals from us who and what we are; most importantly it makes us forget our own mor-

tality. To live an authentic life one must achieve an authentic "being-towards-death".

Authentic Dasein now comes to occupy centre stage. According to Heidegger, it is only from a vantage point of authenticity, in particular an authentic "being-towards-death" that one can finally emerge from the fruitless search to validate knowledge, truth and values by reference to standards that might somehow transcend time. For Heidegger, this daunting vision that there is nothing *beyond* Dasein's essentially temporal being, far from inducing nihilistic despair, releases our capacity for making resolute choices; choices which are always to be made within a *historical*, horizon. Thus we *choose* our identities, *choose* how to interpret our place in the world, but that choice is made along with others (*Mitsein*) by choosers who are always aware of the historical significance of their choice.

Authenticity grounds *resoluteness,* which "as *authentic-Being-one's-self* does not detach Dasein from its world, nor does it detach it so that it becomes a free-floating "I". And how should it, when resoluteness as authentic disclosedness is authentically nothing else than Being-in-the-world?" Whew.

But worse is to come:

> "... if fateful Dasein, as Being-in-the world exists essentially as Being-with-others, its historizing is a co-historizing and is determinative for it as *destiny*. This is how we designate the historizing of the community, of a people (*Volk*). Destiny (*Geschick*) is not something that puts itself together out of individual fates (*Schicksal*) ... Our fates have already been guided in advance in our Being with one another in the same world and in our resoluteness for definite possibilities. Only in communicating and in struggling does the power of destiny become free ..."

From the innocent idea that we find our identities in being together with others, and that this process unfolds in time, we have moved to a conception of a people (a *Volk*) which at once invents and discovers its destiny, as it invents and discovers its history. A people forges itself by "handing down" traditions in "repetition" from generation to generation. This allows Dasein, now identified with the people, to go back into its history and "choose its hero".

It is not hard to see how this vocabulary of people (*Volk*) and hero could all too easily become part of an unphilosophical jargon of political ideology. And that was exactly what happened. Was Heidegger responsible for, even complicit in this hijacking of his work?

The events of 1933[2]

To seek an answer, we must fast-forward six years to the Rectoral address of 1933. Once more some dates:

> 30 Jan 1933: Hitler appointed Chancellor of Germany.
>
> 21 April 1933: Heidegger appointed Rector of Freiburg University.
>
> 1 May 1933: Heidegger joined the NS Party.
>
> 27 May 1933: Heidegger delivered his Inaugural Lecture as Rector.
>
> 23 April 1934: Heidegger resigned from the Rectorate.

The *Rektoratsrede*

With those dates in mind, we now glance at the contents of the Rectoral Address (*Rektoratsrede*). It is entitled "The self-assertion (or Self-affirmation – *Selbstauffassung*) of the German university". In it Heidegger sought to combine a defence of the autonomy ("self-governance") of the university with its precise opposite, the demand that the teachers and students of the university should now be part of the "following" (*Gefolgerschaft*). There is no vulgar, explicit reference to the brutal reality of the Führer. In his place we encounter an impersonal "mission": "the leaders (of the university) are themselves led - led by that unyielding spiritual mission that forces the fate of the German people to bear the stamp of its history".

With the "return" to primordial "science", we are bidden to distance ourselves from subsequent ideologies, both "Christian-theological" and "mathematical-technological". True science "is not a cultural good, but all that binds the individual to people and state" (473). *Dasein*, the key term of *Being and Time* is here used repeatedly to denote the being of the individual as organic part of a people (*Volk*).

In *Being and Time* Heidegger had invoked the "call of conscience". There we were called to be self-aware, to shake off the cosy comforts of *Alltäglich*keit (everyday life). Now the call is issued impersonally to science, which must submit to the command to

> "become the fundamental happening of our spiritual being as part of a people ... The concept of the freedom of the German student

[2]For the chronology see Victor Farias, (English trans. p.84) and Hugo Ott (English trans. p.136).

> is now brought back to its truth. Henceforth the bond and service of the German student will unfold from this truth."

The "bond", according to Heidegger, is three-fold, requiring "labour- service" (*Arbeitsdienst*), "military service" (*Wehrdienst*) and "knowledge-service" (*Wissensdienst*). Students will be required to perform all three services, but the third is their privileged mission. And that mission is particularly infused with ideological content.

The *Rektoraatsrede* is marked, even crippled, by a tension between the phenomenological image of Dasein as fundamentally self-questioning, and a political rhetoric which puts a brutal end to that questioning. It does that by laying down limits to any further questions concerning the nature of knowledge or science. These limits are revealed to Dasein as it comes fully to be identified with its historical (German) destiny. Three themes dominate the *Rektoratsrede* and mark it off from *Being and Time*: (1) criticism of specialization in the sciences, particularly the natural sciences, leading to a domination of intellectual life by what Heidegger will soon identify as *technology*; (2) proclamation of Spirit, with all its religious overtones, but stripped of religious content, to take the place of the dethroned vulgar sciences, and the identification of Spirit with German Spirit; (3) call for the university to be transformed by fusing students and teachers into a body trained for ideological struggle.

Heidegger's activity during the second world war

After his resignation from the Rectorate, Heidegger remained a powerful, maverick figure in Nazi intellectual circles. His ambition was to transform university education in Germany, to mould it into an ideological force infused by the German Spirit. He attempted to set up a new *Dozentenakademie* to train the new generation of university teachers. When that failed, he participated in party initiatives to revolutionize legal training and the teaching of political science. The aim was to bring these disciplines into line with the *Führerprinzip*. On each occasion Heidegger was sidelined, not because he lacked enthusiasm for the NS cause, but because his views were found to be too extreme, even anarchic by the hard-liners who were now dominant within the National Socialist movement.

Heidegger remained a Nazi party member throughout the war, but he retreated from direct involvement with politics. As Professor at Freiburg University he continued to teach and write until 1944, when he was called up, first to work on fortifications on the Rhine, then for service with the Volkssturm:

the oldest member of faculty to be enlisted, a clear sign, according to Heidegger, of the hostility he had incurred from NS officials.

After the war: Heidegger interrogated

After the allied victory in 1945 the French occupied Baden and, as a NS fellow traveller, Heidegger was brought before a denazification committee. The final verdict of the Committee, after lengthy deliberations and disagreements, was that Heidegger should be allowed to continue as a paid, emeritus Professor, but banned from teaching at the University. That ban was lifted in 1951.

Heidegger's initial line of defence was that he was a lofty intellectual who had acted naïvely, but with the best of motives in accepting the Rectorate, which he did to avoid a worse outcome, the appointment of a party nominee. It was: "to stem the coming development by means of the constructive powers which were still viable". And, as soon as he realized that this would be impossible, he resigned from office and took no further part in politics. While the first of these claims was true, the second was not: Heidegger continued to promote his own version of National Socialism, with little success, for many years to come. As historians continue to unravel the narrative, "... we know now that Heidegger intentionally misrepresented the facts."

One thing is certain: Heidegger held that there was an "inner truth and strength" in the National Socialist movement, namely a vision "some day (to) bring about a gathering of what is German unto the historical essence of the west ..." That heady dream was betrayed, according to Heidegger, by party hacks who had sold out to "technology".

Heidegger's view of what went wrong with National Socialism: the triumph of technology

The theme of technology became increasingly important in Heidegger's later work. In the *Spiegel* interview (1966), Heidegger condemned technology because it "tears men from the earth and uproots them". The term "technology" embodied all that Heidegger found wrong with the modern world, dominated by instrumental rationality and forgetful of Being. But technology also plays a sinister role in Heidegger's own forgetfulness of the brutal reality of the Holocaust: "Agriculture is now a motorized food-industry - in essence the same as the manufacture of corpses in gas chambers and extermination camps ..." (Lecture 1949). Heidegger was pressed repeatedly to accept that as a party member with a high public profile, however distant he may have been

from the centres of power, he shared responsibility for Nazi atrocities. He rejected all such demands. The mistake had been made by others, who had put technology in command. The dominance of technology would lead to many undesirable results, including battery chicken farms and also to the extermination camps. And these two examples are, from Heidegger's perspective, "essentially the same". The response is breathtaking. At the grotesque level of generality adopted by Heidegger, all morally relevant distinctions between the two cases evaporate, and with them all questions of moral responsibility.

Heidegger and anti-semitism

So was Heidegger an anti-semite? That apparently simple question requires a careful answer.[3]

Heidegger expressed hostility to Jewish influence as early as 1929 (the year he began dabbling in right-wing politics), when he proclaimed in a letter to an official in the Ministry of Education: "We now face a real choice whether we should again provide for our German spiritual life (unserem deutschen *Geistesleben*) talents and educators rooted in our soil, or whether we should surrender it once and for all to an ever-growing "jewing" (*Verjudung*) in both a broad and narrow sense". Heidegger had already used the term *Verjudung* as early as 1916 in a letter to his future wife Elfriede: "The *Verjudung* of our culture and the universities is really frightening and I think the German race (*die deutsche Rasse*) should find enough inner force to reach the summit." Heidegger later disagreed with party spokesmen about their different interpretations of the term, but like them he continued to deplore the *Verjudung* of German culture.[4] Historians have tracked the recurring anti-semitic elements in Heidegger's wartime lectures.[5]

Heidegger disagreed with the Nazi race "theorists" on the question of "biologism". They claimed that Jews were tainted in virtue of inherited physi-

[3]. Safranski notes that, in his judgment on Heidegger's possible appointment to the Chair in Philosophy at Berlin in 1933, the party hack Jaensch criticized him for being "talmudic, rabbinic and Jewish in spirit". One can only assume that Jaensch recognized the hermeneutic turn taken in *Being and Time*. But since the heritage of biblical exegesis had been carried over into Christianity (and carried on by Koranic scholars in Islam), it is absurd to restrict it to only one of the three "religions of the book" (Safranski, p.268).

[4]. This previously unpublished letter was discovered by Ulrich Sieg and published in *Die Zeit* Feuilleton 52 (29 December 1989). It has since been widely quoted. I found it first in Safranski, ch.14. My thanks to Richard Maguire for getting hold of the German text for me.

[5]. Faye has assembled the most detailed and systematic inventory of Heidegger's anti-semitism that I have encountered. He quotes the recently published 1916 letter on p.10.

ological characteristics. Determinism of that kind was never part of Heidegger's philosophy. In *Being and Time* to subscribe to such a doctrine would be a mark of inauthenticity. Yet even in that text authentic Dasein, in resolving to identify itself with its people, has only one choice. Those and only those who have a shared history can make that choice.

There is still nothing in this story that would exclude Jews from identifying themselves with the *Volk*. Heidegger's biographer Safranski reports that in his lectures given in the 1930s he explicitly denied that there existed a "Jewish spirit" in philosophy.[6] In theory at least, German Jews, no less than German "aryans" could have identified themselves with the Volk; provided they had absorbed its historical destiny.

Yet Heidegger showed himself indifferent to the fate of his Jewish colleagues and students, most of whom would soon flee Germany. Husserl remained in increasing isolation in Freiburg until his death in 1938. His widow managed to find refuge in Belgium until the end of the war.

Heidegger remained silent about the Nazis' persecution of the Jews throughout the time of their rise to power until the end of the war. He retained that attitude of indifference after the war, as is clear from his judgment that the extermination camps were "essentially the same" as industrial farming. Now there is nothing in *Being and Time* that *entails* Heidegger's attitude to the Jews. Rather there are yawning gaps. There is no space for the discursive realms of moral responsibility and interpersonal relations. So when he was interrogated about these matters after the war, Heidegger gave the impression of someone who did not understand the language in which the questions were posed.

If that gives an explanation, though not an excuse, for the behaviour of Heidegger the philosopher, what of Heidegger the man? [7] In the inner circle of Heidegger's star students in the 1920s many were Jews, or at least would be classified as Jews by the Nazis, though few had previously thought of themselves as such. They included Karl Löwith, Herbert Marcuse and, for two tempestuous years preceding the publication of *Being and Time*, Hannah Arendt. The story of the passionate clandestine affair between the 36 year old professor nearing the zenith of his acclaim and his beautiful Jewish student, seventeen years his junior, is only now coming to light.[8] Hannah Arendt en-

[6]. Safranski, p.256.

[7]. In making that distinction I echo Herbert Marcuse, who wrote to Heidegger in 1947: " I - and many others - have learnt an immense amount from you as a philosopher, but we cannot separate Heidegger the philoso-pher and Heidegger the man. For to do that would be to contradict your own philosophy ..." (Reprinted in Martin, p.156)

[8]See Lilla's brilliant narrative of the relationship between Arendt, Heidegger and Jaspers. For

tered the relationship an impressionable, insecure teenager, and emerged from it a woman of great moral and intellectual power. After the nazi takeover Heidegger did no more for his former lover than he did for any of his other Jewish students. She married, escaped to Paris and thence to New York, where she became one of the leading figures in post-war social thought. She published an angry denunciation of Heidegger in 1946, but was later reconciled with him and publicly defended his philosophical legacy. The story of a doomed love affair that turned into a lifetime friendship lies for the most part outside the domain of philosophy. But it constitutes a fitting inconclusive conclusion to this evening's lecture!

2. Part Two: Afterthoughts, 2013

In 2008 I presented the opposition between the two poles of interpretation as I have just done, and then proceeded swiftly and baldly to conclude that the second "continuist" pole was correct, in other words that there was a connection between *Being and Time* and Heidegger's political stance in the 1930s. Five years later I still think that judgment was right, but I have used the intervening time to return to a number of the more important commentators and to recast my conclusion in the light of their work.

Those denying continuity have rightly pointed out that *Being and Time* is a powerful extended reflexion on the human condition, a radical attempt to go beyond dilemmas of epistemology and philosophy of mind. *Being and Time*, on this reading, is no more concerned with politics than Wittgenstein's *Philosophical Investigations* or Ryle's *Concept of Mind* are. Marcuse expressed this response most clearly in his 1947 letter to Heidegger, in which he distinguished "Heidegger the philosopher" from "Heidegger the man". From this perspective Heidegger's political activity in 1933 would be a brief personal aberration, wholly detached from his philosophy.

Those asserting continuity included figures as diverse as Kolnai, Löwith, Lukacs and Adorno, all of whom found direct links between the philosopher of *Being and Time* and the National Socialist Rector of 1933. From this short list I shall say no more about Lukacs's *The Destruction of Reason*, in which Heidegger makes a brief appearance along with Jaspers, only to be summarily dismissed. For Aurel Kolnai, Jewish convert to Catholicism, writing in 1935, it was its exclusion of personal relations, along with privacy and autonomous morality. Karl Löwith sought refuge from the Nazis in Rome. He describes his

the full story see Ettinger, Grunenberg and Young-Bruehl.

neeting there with Heidegger in 1936, during the dourse of which he, Löwith, told his former teacher that in his opinion "his commitment in favour of National Socialism was in the essence of his philosophy. Heidegger unreservedly approved of my judgment and added that his notion of "historicity" was the basis of his political *commitment* ... . He also left me in no doubt about his faith in Hitler ..." (p.77). After the war Theodor Adorno argued in *The Jargon of Authenticity* (1964) that there was a direct link between the idea of authentic Being, involving "rootedness" and a sense of belonging, elaborated in *Being and Time*, was taken up in the vulgar rhetoric of the *Rektoratsrede*. Underlying the jargon, accorfing to Adorno, is a profound error, namely the doctrine of "reflected unreflectedness" The latter is a philosophical thesis asserting that unmediated Dasein has ontological primacy, yet that thesis, like any other piece of philosophizing, is itself an act of reflection. Much of the rest of Adorno's book consists of more polemical swipes, more or less *ad hominem*. But there too, in exposing the hypocrisy of the intellectual supposedly most at ease in the company of simple peasants.

But since 2008 I have begun to think that the more interesting commentators cannot easily be assigned to one or other of the two camps I distinguished. Most of them, in other words, have found certain elements in *Being and Time* which find echoes in the *Rektoratsrede*, even though the latter, along with other texts from that period betray a radical change of direction.

An indispensable work here is Ernst Tugendhat's article "Heidegger's idea of truth" (1969), not least because in it the author addresses a purely philosophical problem, leaving his readers to draw their own conclusions about its relevance to Heidegger's politics. Tugendhat's starting point is*Being and Time*, 44.221: "... Dasein discloses itself to itself in and as its ownmost potentiality for Being. This *authentic* disclosedness shows the phenomenon of the most primordial truth in the mode of authenticity.". Following Tugendhat we find that in this section Heidegger [a] started by applying the notion of truth as disclosedness to the truth of assertions and then ;b] extended it to all that can be uncovered, that is to all disclosure of "the world". Everything follows from [a], from the notion that the truth of an assertion lies in its disclosedness. This in turn is Heidegger's version of Husserl's theory of truth in *Logical Investigations*, in which truth was the correspondence between [i] the state of affairs as it is intended in signifying givenness and [ii] that same state of affairs as it is in itself. A relation of identity. Heidegger transformed Husserl's theory by removing "in itself". So now "The assertion is understood as its disclosedness. The truth of an assertion now consists simply in the pointing out, uncovering, disclosing of Being, with no reference to "as it is in itself".

Disclosure Is now understood as an *occurrence*. In Husserl the act of assertion is understood *statically*. In Heidegger it is understood *dynamically*. It is actively relayed to its opposite – "closedness", so that "we lift it out of concealment!., Thus Heidegger abandons "as it is in itself" because uncovering, disclosing "must be true if it really is an uncovering". But in normal usage uncovering is not equivalent to truth, since one may uncover the false. On this point Heidegger prevaricates: in the false assertion "the false is in a sense already uncovered and still not represented". In short "if one limits oneself to the two concepts concealment and unconcealment, there remains absolutely no possibility of determining the specific sense of falsehood, and therefore also of truth". "Because the truth of an assertion does not lie in the way that it is uncovered but only in the fact that it is uncovered, [Heidegger] is able to carry truth over to all truth in general". There is no way of distinguishing between what is true and what is false once you accept that all disclosure is true. Disclosure of Dasein is itself "the most primordial truth". In other words "self-manifestation is itself truth". So "there is no place for critical consciousness to assess truth claims" (238). Against Heidegger Tugendhat defends "the regulative idea of certainty and the postulate of a critical foundation". In Tugendhat's final judgment, "Heidegger does not just set aside the notion of truth. He holds on to it and deforms it".

Tugendhat's reading of this crucial section of *Being and Time* is so valuable because, instead of dismissing the so-called "apophantic" account of truth out of hand, he takes it seriously and shows how Heidegger produced his own "deformed" theory from the sober phenomenological approach of Husserl. Although Tugendhat restricted himself in this paper to a purely philosophical question.

For Jürgen Habermas Tugendhat's account of the apophantic theory of truth formed a key part of his own picture of the vulnerability of Heidegger's philosophy to ideological subversion, a process in which philosophy gave way to *Weltanschauung*. Equally important, as Habermas saw it, were elements deep within the ontology of *Being and Time* which effectively precluded the serious study of intersubjectivity and society and of real historical processes. The process of subversion, argued Habermas, gathered momentum from 1929 onwards, coming ever closer to a diagnosis of the disorders of our time, rather than serious philosophical reflection.

It is at this point that I screw my courage to the sticking point and dare to mention the name of Derrida in a collection of essays dedicated to Pascal. I do it only because, if we cut through the fancy verbiage (or, as Pascal would have it, the bullshit), we find at the heart of Derrida's *De l'esprit* an important insight

into Heidegger's Nazism. Derrida spotted that the term *Geist* (spirit), which Heidegger vowed to avoid in *Being and Time* (1927), and there used only rarely and always in quotation marks, plays a prominent role in the *Rektoratsrede* (1933) and in the *Introduction to Metaphysics* (1935), texts in which Heidegger espoused National Socialism, while distancing himself from "vulgar Nazis" who had failed to grasp the spiritual dimension of their movement. In Derrida's words, *Geist* "is regularly inscribed in contexts that are highly charged politically". In such contexts spirit is essentially German and it summons German academics, teachers and students, to identify themselves with the historical (spiritual) destiny of the German people. Insofar as Derrida draws our attention to the sinister tone of Heidegger's invocation of *Geist* as bearer of a spiritualized *Führerprinzip*, his analysis rings true. But his diagnosis of Heidegger's error will convince only true believers in the Derridean gospel. His argument, as far as I can understand it, goes like this. Heidegger marks off his version of National Socialism from biologism and racism by identifying it with spirit. But that strategy "risks" turning spirit into a *subject*, something which should have been definitively replaced by *Dasein*, the central element of *Being and Time*. To Derrida it was self-evident that if you refer to a subject in a philosophical context, you are doomed to return to a metaphysical doctrine of the self as substance. It was so self-evident to Derrida that it barely merited an argument. According to Derrida, if the substantial subject embodied in *Geist* is free, as Heidegger said it was in the *Rektoratsrede*, then that freedom "always runs the risk of [turning into what Hegel called] a merely formal liberty of an abstract universality". This comment occurs in a footnote to a page in which Derrida mentions "those who state their opposition to racism, totalitarianism, Nazism, fascism etc in the name of ... the freedom of (the) spirit, in the name of an axiomatic — for example that of democracy or 'human rights' (Derrida's quotation marks) which ... comes back to this metaphysics of *subjectité*" (Derrida's coinage). This passage suggests that Derrida agrees with Heidegger in rejecting any version of ethics involving "rights talk", based on the idea of a morally autonomous agent. Derrida seems to accept, with Heidegger, that anyone who might think of doing serious moral philosophy will give up as soon as they have read Nietzsche's *Genealogy of Morals* or *Beyond Good and Evil*. As a result, Derrida, like any rational person, assumed that "racism, totalitarianism, Nazism, fascism etc." are deplorable, while rejecting any attempt to demonstrate their unacceptability by means of a systematic moral theory. Monique Canto-Sperber has provided a clear account of the eclipse of moral philosophy in post-war France, a process in which Derrida played a leading part.

Up to this point my second thoughts had modified my first thoughts, without wholly shaking them. But after reading Hans Sluga's *Heidegger's Crisis* I began to wonder if I had entirely misjudged the Heidegger case from start to finish. In his book Sluga argues that it is pointless for historians to pass moral judgments on events in the past and, worse, that it is likely to distort their understanding of those events. To hisnrather sweeping dismissal of all moral judgments about the past, it could be objected that the historian he is certainly entitled to report the moral judgments of Heidegger's contemporaries. And anyway Sluga's dictum is open to challenge. His own teacher Michael Dummett evidently did not feel bound by it when he wrote in the Prefaxe to *Frege's Philosophy of Language*: "There is some irony for me in the fact that the man about whose philosophical views I have devoted ... a great deal of time to thinking, was, at least at the end of his life, a virulent racist, specifically an anti-semite." Dummett reports on the effect of reading the previously unpublished section of Frege's diary, which "shows Frege to have been a man of extreme right-wing political opinions, bitterly opposed to the parliamentary system, democrats, Catholics, the French and, above all, Jews, who he thought ought to be deprived of political rights and preferably, expelled from Germany. I was deeply shocked, because I had revered Frege as an absolutely rational man, if, perhaps, a not very likeable one ... From it I learned something about human beings which I should be sorry not to know; perhaps something about Europe, also." And the distinction between "the man" and "the philosopher" is very much easier to draw in Frege's case than in Heidegger's.

But Sluga's more important point about Heidegger's critics is that they have focused on him in isolation from more general trends of the time. So in entitling his book *Heidegger's Crisis* he drew attention to the fact that there was general agreement among German intellectuals between the wars they were living through a time of crisis and that the situation called for a drastic solution. When political parties and their leaders offered such solutions, they found a ready audience, particularly in intellectual circles/ Sluga concentrates on Germany, but what he says applies equally to most European countries. Sluga shows that numerous other German philosophers, of various schools of thought, had been critical of the parliamentary democracy of the Weimar Republic since its inception, and ended up by subscribing to National Socialism. The DPG (*Deutsche philosophische Gesellschaft*) representing conservative philosophers, gave its allegiance to the party at its meeting in Magdeberg in 1933. By 1938 roughly half of German philosophy professors were party members.

Against that background Sluga argues that there is no particular link between Heidegger's philosophy and National Socialism since numerous other philosophers committed themselves to the cause while holding philosophical views contrary to Heidegger's. Critics of Heidegger point to three theses on his philosophy which make it susceptible to political subversion: [a] its rejection of transcendental norms and values; [b] its irrationalism; [c] its decisionism. But other supporters of the Nazis based their support on reason and universal values and, on the other hand some of its opponents subscribed to one or more of those theses. The list would include positivists, existentialists and many others.

Sluga devotes much of his book to a detailed investigation of these pro-Nazi philosophers. Most of them, with a few exceptions like Nicolai Hartmann, were previously unknown outside Germany. A particularly influential figure was Max Wundt, who was dedicated to German idealism in philosophy. Once he had espoused National Socialism he fabricated an ideology of the German (idealist) spirit which needed to be cleared of all contamination, especially Jewish, but also Catholic. In such company Heidegger appears relatively innocent, untainted by either pseudo-scientific idealist accounts of German racial superiority. In explaining the widespread cultural conservatism of the tme, Sluga diagnoses a particularly severe case of nationalism present in Germany, partly due to the late emergence of the unitary state, exacerbated by defeat in the Great War and the economic collapse. So in the *Rektoratsrede* Heidegger "was not initiating a new kind of discourse but merely inserting himself into one that already had a long history … None of [its] ideas was original and he made little of them in his philosophical thinking."

Sluga makes a compelling case for understanding Heidegger's engagement with National Socialism in its historical context. He has demonstrated that the myth of Heidegger's unique contribution to the Nazi cause is just that – a myth. But one is left wondering whether Sluga has thereby rendered that contribution *banal* (to echo Hannah Arendt's judgment that Eichmann embodied *The Banality of Evil*). But the historical record tells us that Heidegger, unlike Eichmann, was not a banal figure. The publication of *Being and Time* in 1927 had established his reputation throughout Germany and abroad. By the time of the Davos encounter with Cassirer in 1929 he had already become a celebrity, his fame having passed beyond the confines of academic philosophy. After the *Rektoratsrede* and his public espousal of National Socialism, despite his unorthodox and controversial version of its creed, he still received invitations from Berlin and Munich to take the Chairs in Philosophy at their respective universities. In short, while accepting Sluga's outstanding contri-

bution to establishing the cultural milieu in which Heidegger found himself, we can remain unapologetic in putting him in the foreground of the picture, both because his philosophy was and still is so interesting, and also because he played such a prominent role in the cultural politics of his time.

3. References

This is a list of some of the works I consulted when writing the lecture, along with one or two others I have encountered since then. It represents a microscopic fraction of the immense literature devoted to the topic whose scale can be appreciated by a glance at the bibliography, running to over 30 pages, in Faye's book. Moreover it makes no pretence at scholarship. I have simply used the versions of the texts which I had to hand, a motly mixture of items, some in their original languages, others in translation.]Heidegger, Martin, Being and Time, trans. Macquarrie and Robinson, Oxford: Blackwell, 1962 (Sein und Zeit, first publishes 1927).

Heidegger, Martin, "The Self-assertion of the German university" (*Rektoratsrede*), delivered 1933)

Heidegger, Martin, *Introduction to Meta[hysics*, trans. Manheim, New Haven: Yale University Press, 1959 (*Einführung in die Metaphysik*, lectures delivered in 1935)

Heidegger, Martin, "Only a god can save us" (Interview with Der Spiegel, conducted ?, published 1976, trans, Richardson I *Heidegger, the Man and the Thinker*, 1981111)

Adorno, Theodor, *The Jargon of Authenticity*, trans. Tarnowski and Will, London: Routledge and Kegan Paul, 1973 (*Jargon der Eigentlichkeit: zur deutschen Ideologie*, Frankfurt: Suhrkamp, 1964)

Derrida, Jacques, De l'esprit: Heidegger et la question, Paris: Galilée, 1987. Partial trans. Bennington and Bowlby, in *Critical Inquiry*, 1989

Dummett, Michael, *Frege: Philosophy of Language*, London: Duckworth, 1981

Farias, Victor, *Heidegger et le nazisme*, trans. Benaroch and Grasset, Paris: Verdier, `1987

Faye, Emmanuel, *Heidegger, l'introduction du nazisme dans la philosophie*, Paris: Albin Michel, 2005

Gordon, Peter E., *Continental Divide: Heidegger, Cassirer, Davos*, Cambridge, MA: Harvard University Press, 2010

Habermas, Jürgen, *The Philosophical Discourse of Modernity*, trans. Lawrence, Cambridge: Polity Press, 1987 (*Der philosophische Diskurs der Moderne*, 1985)

Habermas, Jürgen, "Work and Weltanschauung: the Heidegger from a German perspective", trans. McCumber, in *Critical Inquiry*, 1989

Löwith, Karl, *Ma vie en Allemagne avant et après 1933*, trans. Lebedel, Paris: Hachette, 1988

Lukacs, Georg, *The Destruction of Reason*, trans. P. Palmer, London: Merlin Press, 1980 (*Die Zerstörung der Vernunft*, 1962)

Martin, Bernd (ed), Martin Heidegger und das "Dritte Reich": ein Kompendium, Darmstadt: Wissenschaftliche Buchgesellachaft, 1989

Mulhall, Stephen, *Heidegger and "Being and Time"*, London: Routledge, 1996

Ott, Hugo, *Martin Heidegger: Eléments pour une biographie*, trans.Beloeil, Paris: Payot, 1990 (*Martin Heidegger: unterwegs zu seiner Biographie*, 1988)

Safranski, Rüdiger, *Martin Heidegger: between Good and Evil*, trans. Ewald Osers, Cambridge, MA: Harvard University Press, 1998 (*Ein Meister aus Deutschland: Heidegger und seine Zeit*, 1994)

Sluga, Hans, Heidegger's Crisis: Philosophy and Politics in Nazi Germany, Cambridge, MA: Harvard University Press, 1993

Tugendhat, Ernst, "Heidegger's idea of truth" in R. Wolin (ed), *The Heidegger Controversy*, Cambridge, MA: the MIT Press, date ? ("Heideggers Ideee von Wahrheit", 1969)

Young, Julian, *Heidegger, Philosophy, Nazism*, Cambridge: Cambridge University Press, 1997

On Hannah Arendt

Ettinger, Elzbieta, *Hannah Arendt, Martin Heidegger*, New Haven: Yale University Press, 1995

Grunenberg, Antonia, *Hannah Arendt und Martin Heidegger: Geschichte einer Liebe*, Munich: Piper, 2006

Lilla, Mark, , "Ménage à trois" and "The perils of friendship" in *New York Review of Books*, 18 November, 2 December 1999

Young–Bruehl, Elisabeth, *Hannah Arendt: for Love of the World*, New Haven: Yale University Press, 1982

53

Philosophy as Literature: The non-argumentative tradition in continental philosophy

NENAD MIŠČEVIĆ

Being's poem, just begun, is man.
Martin Heidegger

Abstract Pursuing a line from Pascal Engel's remarkable dialogue "*La Dispute*", the paper discusses the non-argumentative tradition within contemporary philosophy. The tradition encompasses some very successful and famous 20^{th} century philosophers, like Heidegger and Derrida (and theoreticians, like Jacques Lacan in his later phase), who systematically avoid any sort of explicit argumentation in their work. Philosophizing without argument here means doing philosophy without any visible argumentation-like steps. How did the non-argumentative writing gain its place in twentieth-century philosophy? The paper proposes a philosophical account, resting on the assumption that the authors in question are following an intellectual strategy. Assuming that a-rational aspects of human existence (desire, passion, and the like) are of central interest they accept an implicit methodological principle: the cognitive style, the language, style and the method of studying an a-rational domain D should follow the language, style and the manner of D itself. In particular, for such a-rational domains, the cognitive style and the linguistic expression should minimize the use of (or perhaps completely eschew) traditional rationalist methods of enquiry and presentation.

1. Introduction

Pascal Engel (I shall call him in the sequel just "Pascal", as I did for decades) has been struggling for analytic philosophy, its importance and its status, in the middle of an atmosphere that has been all but friendly to it. His effort, quite successful to my knowledge, needs to be praised, and this is what I intend to do, by dedicating this paper to him.[1] Almost two decades ago, Pascal has produced a fine philosophical dialogue *La Dispute* (1997); two characters, Analyphron and Philoconte discuss analytic and continental philosophy, the former defending the first, and the later the second; Mésothète, a third character, tries to mediate . In the later work (Rorty and Engel, 2007), the imaginary dialogue is replaced with the actual polemic between Pascal and Rorty, to whom we shall refer often in the sequel. But let me start with *La Dispute*. Early in the book Analyphron diagnoses an important contrast between the two schools:

> Là où le brio du continental se manifeste dans l'écriture littéraire, les· jeux de mots et les formules, les vastes synthèses,

[1] I feel honored by the invitation to contribute, so I thank the organizers for their invitation, and in particular Anne Valérie Meylan Massin for her support and patience.

> le brio de l'analytique se manifeste dans la manipulation des langages logiques, mathématiques, des concepts scientifiques. (1997 :23)
>
> Whereas the brilliance of a continental philosopher manifests itself in the literary style of writing, the play with words and formulations, joined to enormous works of synthesis, the brilliance of the analytic philosopher manifests itself in the manipulation of logical and mathematical languages and scientific concepts. (my translation)

I would like to follow Pascal's interest in this stylistic analytic-continental contrast and propose an *homage* to him, focusing on the first part of Analyphron's diagnosis and raising the question: how did continental philosophy become so prone to "the literary style of writing, the play with words and formulations"? After all, the tradition did not start in such a style, witness Hegel and post-Hegelians; and even at the time when, due to Kierkegaard and Nietzsche, a literary sub-variety of continental thought has been born, the central continental figures, like the members of the Brentatno school, Husserl, Dilthey and most of their immediate disciples, did write in an argumentative style, with keen interest in logic (in the wide sense, relevant here), and the desire to follow the model of science to a significant extent, rather than taking poetry as their paradigm and indulging themselves in "the play with words and formulations". I shall be talking about authors like Kierkegaard, Nietzsche, Heidegger, Derrida, or Žižek, i.e., merely about one line in continental philosophy, albeit a quite central one. (So, I leave authors like Habermas and Apel, or even Ricoeur aside for this occasion).[2]

Does philosophy centrally involve arguments? Many of us would like to think so, but there is a strong tradition that favors less argumentative, and often non-argumentative style. The paradigms of this tradition are the quasi-literary, ironic, stylistically reach works of Kierkegaard, Nietzsche's "*Thus Spoke Zarathustra*", Heidegger's late poetic-sounding works, like "*The Experience of thinking*", Lacan's highly complex, often playful and often very opaque "*Ecrits*", Derrida's experiments with language, and these paradigms are being imitated and varied by a long row of followers and pupils, whose work characterizes the post-modernist and/or deconstructionist scene . (Of course, this

[2] For a very different approach to the issue see also Samuel Wheeler's paper "Philosophy as Art", on his web-site.

is only one tradition within continental philosophy, not all of it; we shall return to this in a moment). Just as a reminder and an illustration, let me quote a distinguished follower of Derrida, J. L. Nancy, talking about alterity.

> The alterity of the other is its being-origin. Conversely, the originarity of the origin is its being-other, but it is a being-other than every being for and in crossing through [à travers] all being. Thus, the originarity of the origin is not a property that would distinguish a being from all others, because this being would then have to be something other than itself in order to have its origin in its own turn. (2000: 11)

Who and what counts as the other? What is exactly the alterity of the other? And why would the alterity be connected, let alone be identical to "being-origin"? We might try to guess. Maybe "the other" of the eurocentric culture are us (me and my co-nationals), Slaves, or Muslims and so on. (Nancy was extremely engaged in helping us, former Yugoslav intellectuals in the difficult time of the war; I have fond memories of talking to him about our plight). But we are certainly not "origin". So, other must be something else. The important point is that no explanation is offered. Deep, or at least deep-sounding thesis of the first sentence quoted is left without any discursive support. So, back to the question of who is the other. Maybe it is God. The text point in this direction:

> This is the most classic of God's aporias, and the proof of his nonexistence. In fact, this is the most immediate importance of Kant's destruction of the ontological argument, which can be deciphered in a quasi-literal manner ; the necessity of existence is given right at the existing of all existences [l'exister de tout l'existant], in its very diversity and contingency. In no way does this constitute a supplementary Being. The world has no supplement. It is supplemented in itself and, as such, is indefinitely supplemented by the origin. This follows as an essential consequence (Ibid.)

So this is how a respectable later-day continental philosopher talks about proofs and consequences. And Nancy is a serious academic, a rather strict university professor, not a poet nor a public figure seducing a wide cultured audience. The tradition we just briefly introduced is our object of study in this paper. The paper discusses a non-argumentative tradition within contemporary philosophy. Philosophizing without argument, here means doing

philosophy without any visible argumentation-like steps. Of course, some examples can be reconstructed, in fact re-interpreted in terms of argument, but the argument form is strictly avoided. Late Heidegger, and Lacan systematically avoid any sort of explicit argumentation in their work, and Derrida in some works comes close to the ideal. On the other hand, philosophy cannot do completely without argument; so when these philosopher have one, they hide it into a more poetic text.

Before moving on I want to stress that this is just one current within continental philosophy, not the whole of it. Husserl, Max Scheler and Gadamer are subtly argumentative, Foucault is passionate about historical evidence, and its role in making well-argued points about the unrecognized dark history of the last two centuries (see an interesting discussion in Smokrović, (2013)) Althusser sees philosophy as close to science, and writes in a clear argumentative manner. The mainstream Frankfurt school production has been quite argumentative, and Habermas straddles the continental-analytic divide. So, there is an argumentative, even highly and subtly argumentative tradition within the continental philosophical culture. I will be talking about the other one. (I hope that Nancy's general statement is not right about this other, non-argumentative tradition, and that "its alterity" is not its "being-origin", that its non-argumentative character and radical difference with the arguers is not the original sin of continental philosophy.) Of course, the tradition is very important, very widely read and taught, and worth studying by anyone interested in philosophical issues of argumentation.

The literary character of the tradition has been remarked by authors very sympathetic to it. Richard Rorty goes as far as classifying it as non-philosophical, which is a compliment in his jargon: philosophy itself is moving "from a philosophical to a literary culture" since the time after the death of Kant. And his diagnosis is a bit dramatic, although he is very optimistic about it:

> In the literary culture which has been emerging during the last two hundred years, the question "Is it true?" has yielded to the question "What's new?" (2007:91-2).

And here are the consequences for redefining philosophy, this time in terms of " philosophy as a kind of writing" as the title of one essay in (2007) suggests:

> All that "philosophy" as a name for a sector of culture means is "talk about Plato, Augustine, Descartes, Kant, Hegel, Frege, Russell . . . and that lot." Philosophy is best seen as a kind of writing.

> It is delimited, as is any literary genre, not by form or matter, but by tradition (...), (2007:143).

Note a subtle ambiguity. On the one hand, almost any intellectual activity involves writing, and even mathematics can be described as manipulation of a certain kind of written symbols, as formalist have been eager to do. In this sense, philosophy is unproblematically a kind of writing, in this very wide and non-dramatic sense. On the other hand, Rorty probably means much more; namely that philosophy is close to literary writing, and that this is central to it. Let me mention another author, Michael Weston. In his book on *Kierkegaard and modern continental phil*osophy (1994) notes the following:

> Post-metaphysical thought in Nietzsche, Heidegger and Derrida shows certain central characteristics which have their parallels in Kierkegaard: a 'style' of writing at variance with that of the metaphysical tradition which has its rationale in the 'situatedness' of the thought whose intention is, not the representation of 'the truth', but an 'intervention' into that situation. (1994:136).

His examples are vey well chosen:"Nietzsche's use of aphorisms, stories, poems, the fictional character of Zarathustra, Heidegger's 'etymologies' and 'poetic' thinking, Derrida's 'double-reading' (Ibid.). He notes that all this continued and strengthened today in some of the mainstream continental work, in cultural studies, continental feminist philosophy. Why are these non-argumentative moves important for the thinkers mentioned? In his judgment these " are strategies of writing demanded by the essentially 'situated' character of their thought. "(136). I don't' see why one cannot be essentially situated and still arguing, but I leave it at that. But mere "situatedness" explains little; why would one use etymologies merely because one is situated? If the answer is that the use of etymologies is dictated by our situatedness in time, then why not use General theory of relativity, given our situatedness in space-time?

So, there is a strong non-argumentative tradition in continental philosophy, and it is worth being analyzed. In this paper I want to address three questions: how is the non-argumentative discourse typically structured? I shall do it very, very briefly in the next section. Next, where did it all come from in the nineteenth and how it developed in the twentieth century? This will take most of the space, and still will be done quite sketchily, given the huge material available, in section III. Finally, in the conclusion I summarize the main findings, and briefly address the question of what one should do assuming that one is into argumentative style.

2. Depicting the non-argumentative tradition: the allusive philosophising

Let us start with a passage from the central continental thinker of the 20th century, Martin Heidegger:

> But what is it that touches us directly out of the widest orbit? What is it that remains blocked off, withdrawn from us by ourselves in our ordinary willing to objectify the world? It is the other draft: Death. Death is what touches mortals in their nature, and so sets them on their way to the other side of life, and so into the whole of the pure draft. Death thus gathers into the whole of what is already posited, into the positum of the whole draft. As this gathering of positing, death is the laying-down, the Law, just as the mountain chain is the gathering of the mountains into the whole of its cabin. (1971:123).

If you were a discourse analyst and were given the quotation as homework what would you first notice? First, pronounced literary form, and none or very few indicators of any kind of arguing ("so", "therefore" and the like). Second, the texts is seriously polysemous (without indications about decoding). You might miss a central point if you don't look at the German original: the word "draft" stand for German "Entwurf", and the etymology of *Entwurf* has to do with "werfen", to throw; so the innocently looking "draft", is in fact a way in which ones existence is "thrown" into the world and history. So, the original gives you "Entwurf" which is both simple "draft" and "the thrown"; the translator has opted for one, and lost the other. Thirdly, we have central use of poetic figures, the use of "Entwurf" pointing to a philosophically wide-reaching metaphor. Again, the reader is not told how to interpret the metaphor, so that even the translator, at the end of the day, chose not even to suggest it to the reader; the translation "draft" makes life easy for the reader, but misses the main point of the author.

What about the pragmatics of the paragraph? Well, an important, if not the most important, goal seems to be suggesting and evoking. Mentioning death is by itself significant, but death is being characterized in a deeply suggestive and passionate way: "It is the other draft: Death. Death is what touches mortals in their nature, and so sets them on their way to the other side of life, and so into the whole of the pure draft." "Taking a way on the other side of life" is not a usual matter; how many of us think that there is a way "on the other side

of life"? We are being invited to imagine a journey; personally I was reminded of a beautiful journey of the soul of the hero in the Russian movie "Cuckoo" ((Kukushka) by Rogozhkin); his soul takes "a way on the other side of life", but is called back by the women in love in an immensely poetic sequence. But what about the philosophy in the passage? There seems not much left of any argumentative point. It is rather an invitation to thinking following the poetic figures.

In fact, the text is typical. Very often the following features will be easily spotted. First, as to form, there is no explicit argument-form; and often one finds pronounced literary form. Second, as to semantics, on encounters a seriously multiply ambiguous text without clear indications how to disambiguate it. Given a long tradition of the search for definitions in philosophy, from Socrates and Aristotle, thought Leibniz and Kant to Frege and the analytical philosophers (or at least search of either necessary of sufficient conditions for something to fall under the given concept) the contrast is quite dramatic. What or who is exactly "the other"? Maybe the philosopher has five meanings in mind, maybe only three. But he does not tell us; at best we might get a discreet indication. Derrida is explicit about polysemy: first, any text is polysemous, second, polysemy is indefinite, not to be captured by making distinctions, third, this is a very positive state of affairs, repressed by the logocentric metaphysical tradition, and fourth, philosopher-writer should multiply meanings way beyond necessity. [3] Even more importantly, we encounter massive and central use of poetic figures without indication about decoding.

In fact, we should distinguish between weakly and strongly non-argumentative style. The strongly non-argumentative style eschews any argument form, proliferates meanings, sometimes very vague and allusive ones, and straddles into poetry. It is a deeply allusive philosophy. The weakly non-argumentative style hides the arguments it uses. Typically, in the Heideggerian tradition, the philosopher would appeal to the authority of some great predecessor, e.g. a Pre-Socratic. But, the appeal would not be done in the form of explicit *argumentum ad verecundiam*. The pre-Socratic would be quoted, with a suggestion

[3]Here is how Derrida expresses his view that a non-figurative treatment of metaphor is impossible:

> I am obliged to speak of [metaphor] more metaphorico, to it in its own manner. I cannot treat it (entraiter) without dealing with it (sans traiter avec elle) ... I do not succeed in producing a treatise (une traite) on metaphor which is not treated with (traite avec) metaphor which suddenly appears intractable (intraitable). (1998102–3.

that his quote is extremely important, and carries a deep message. Then, some erudition and some poetic temper would be brought to the deciphering of the quote, resulting in a meaning quite surprising to the novice. The strong suggestion is that the meaning is deep, and true in a deep way.

Thirdly, on the side of pragmatics, in the strongly non-argumentative line the main goal is suggesting, often by non-rational, evocative means. The text is often just invitation to thinking following the poetic figures. In the weakly non-argumentative variant, suggestion and evocation is a goal, not always the main one, and the reader is given a bit more clear indication in which direction to go on thinking.

While we are at the pragmatic, it is worth while mentioning an important additional strategy for subverting the argumentative, namely the judicious use of pseudonyms. You read Kierkegaard on Abraham, the *Fear and Trembling*, and you recognize a pleading in favor of Abraham and his forming the intention of killing his son. The pleading is not merely emotional; it contains interesting arguments, for instance from the transcendence of God. Naively, you start agreeing with Kierkegaard, like many of my students routinely did. But a sophisticated interpreter, like Stephen Mulhall and Geoffrey A. Hale will immediately tell you that it is not at all clear that this is what Kierkegaard meant, in contrast to what Johannes de Sylentio, the pseudonymous author meant (I witnessed such a discussion between Professor Mulhall and a young interpreter of Kierkegaard Bojan Blagojević in a Budapest conference).

Let me conclude very quickly with another example of allusive philosophizing or theorizing, this time from a text that is not poetic, and that attempts some kind of arguing. It is the famous "The Instance of the Letter in the Unconscious", by Lacan:

> Is the place that I occupy as subject of the signifier concentric or eccentric in relation to the place I occupy as subject of the signified? That is the question.
>
> The point is not to know whether I speak of myself in a way that conforms to what I am, but rather to know whether, when I speak of myself, I am the same as the self of whom I speak. *(*2006*:* 430)

In other places Lacan even apologizes for being allusive « je m'excuse d'être aussi allusive » (1973:21); I will argue that allusivenes is essential for the whole tradition. Let us return to his question. It is indeed reasonable enough. Lacan will be proposing a negative answer; no, when I speak of myself, I am not the same as the self of whom I speak. A naïve reader would be probably

shocked; taken in a literary way, the answer suggests that I can never refer to myself. Heraclitus and Buddhism come to one's mind. So, how does Lacan refer to himself? How does he refer to his patients when he builds a theory about them? One would expect these kinds of concern, in the passage introducing his answer. Instead of which, one is offered the following:

> And there is no reason not to bring in the term "thought" here. For Freud uses the term to designate the elements at stake in the unconscious, that is, in the signifying mechanisms I just pointed to there. It is nonetheless true that the philosophical cogito is at the center of the mirage that renders modern man so sure of being himself in his uncertainties about himself, and even in the distrust he has long since learned to exercise regarding the pitfalls of pride. (2006: 430).

Notice that the primarily theoretical question about referring to oneself is placed into a much more emotional content: the self-certainty of the "modern man", hunted by his "uncertainties about himself ", but sure of being himself. Which is "a mirage"; we are not told why. The simple way out is of course to say that I know who I am, but my uncertainties concern my plans, wishes, abilities, and so on. I am not sure whether I really want to criticize continental philosophy, really want to jog in the cold winter day, and the like. This is compatible with, and even requiring that I know who I am in the minimal sense needed for the first-person reference (and the problem does not have much to do with specifically "modern" man, heaving bothered ancient skeptics, as well as Hindu and Buddhist thinkers). Lacan does not address these simple worries and simple proposals. He continues thus:

> Now if, turning the weapon of metonymy against the nostalgia that it serves I stop myself from seeking any meaning beyond tautology, and if, in the name of "war is war" and "a penny's a penny," I resolve to be only what I am, how can I escape here from the obvious fact that I am in this very act? *Ibid.*

And a few lines later, Lacan changes the topic. No serious question of identity has been raised, even less, answered. Instead, we hear that metonymy is a weapon that serves nostalgia. Why? How? Note that this is one of the founding texts of the Lacanian doctrine, not an essayistic sketch. So, this is an example of what I mean by "weakly non-argumentative" line: mixing the poetic, emotional and historical, all in three lines, without explanation, but

with some reasonable sounding questions and attempts to offer a suggestive semi-answers to them.

So much for general characterization, which I share with Pascal. I am aware that it is too brief, and that many readers will find the conclusions overhasty, and the examples too few and/or not enough compelling. I hope some of these flaws can be remedied in the sequel, "with the positum of the whole draft" (as Heidegger would no doubt put it), with some new examples, which, I hope, conform to our brief and all too sketchy portrait of the tradition.

3. Where did it all come from?

Manifesting the a-rational: Kierkegaard, Nietzsche and the Exemplification constraint

A tradition is a practice extended in history; so, one is curious about its origin and forces that have kept it alive and going. How did it all happen is a central question, and I want to address its philosophical aspects, leaving aside social history and similar, in themselves highly interesting concerns. It did happen "shortly after Kant" as Rorty put it. He sees it, I think rightly, as a reaction to Hegel (in the passage from which we have already quoted the last sentence):

> The transition from a philosophical to a literary culture began shortly after Kant, about the time that Hegel warned us that philosophy paints its gray on gray only when a form of life has grown old. That remark helped the generation of Kierkegaard and Marx realize that philosophy was never going to fill the redemptive role that Hegel himself had claimed for it. Hegel's supremely ambitious claims for philosophy were counter-productive. His System was no sooner published than it began to be read as a *reductio ad absurdum* of a certain form of intellectual life. Since Hegel's time, the intellectuals have been losing faith in philosophy. This amounts to losing faith in the idea that redemption can come in the form of true beliefs. In the literary culture which has been emerging during the last two hundred years, the question "Is it true?" has yielded to the question "What's new?" (2007:91-2)

But even if we accept the "What's new?" turn, it is unclear why it would be inimical to argument. His surmise that it is the giving up on truth sounds better, but Heidegger is a prime counterexample; Heidegger wants a deeper

truth, not untruth, or indifference to truth. So, we need much more detail. First, if there is a reaction to Hegel, and indirectly to Kant, what aspects of the huge philosophical projects of the two are its target? If we agree that it is Kierkegaard and Nietzsche who are the purest examples of the tradition we are reconstructing, we shall also notice that a central target of their reaction is the domination of the Reason, and the rational in general. Rorty, coming from a pragmatist tradition, ignores it.

However, will, desire and affect, with specifications like will to power, sexuality, and the like, play a central role in the whole continental tradition. So, I would propose that the first component in the change that lead to the birth of its non-argumentative wing is the (re-)discovery of the a-rational, or even irrational (as contrary to rational) as a central topic for philosophy. (I am using more neutral "a-rationalist" for views that just set aside the rationality, "irrationalist " for explicit enemies of it). The two did play a role before, but in a more tame fashion. Humean desire is a relatively homely matter, and the human passions in Pascal, La Rochefoucauld and other French Enlightment authors lack a cosmic dimension, which they receive only within the post-Kantian tradition. How does this happen? Let us state the central a-rationalist thesis about the forces at work in human mind:

> (**A-RAT-mind**) The central element of human mind is a-rational, it is either will, desire or affect.

This a-rationalizing might take several forms. Typically it involves setting aside pure cognitive (epistemic) rationality. Often one ends up by replacing it with practical one, for instance in some Marxist, Pragmatist (Rorty) and neo-Heideggerian authors (like Dreyfuss). Hume and Rousseau would have subscribed to (**A-RAT-mind**) as would later Schopenhauer and Maine de Biran.

Let me just mention the transformations of the A-RAT in the French and French-inspired philosophy in the 20^{th} century. Let me mention its three main avatars. The first is the appeal to emotions interpreted as modes of existence; it is probably inspired by Heideggers's very strong claim that all interpretation and understanding is founded in and guided by "mood" and "attunement" (Stimmung and Gestimmtheit, in Being and time, §31-32, for a fine discussion see Hatzimoysis 2009). Sartre stresses the role of emotion in apprehending the world; his "*Nausea*" vividly illustrates how the affective state discloses to us (through his character, Roquentin) the deep meaning of the very being-in-itself. The second avatar, also to be found rather early, in Sartre, is the "body"

as the seat of affection and desire. Husserlian phenomenology of the "Leib", the experienced body, with early Merleau-Ponty on the French side, has been stressing the bodily activity and its cognitive role; the more a-rationalistic approach is to stress the bodily aspects of affective states, the force of hunger and sexual desire, and has made the appeal to "corps" practically synonymous with appeal to affect and drive. The third avatar comes with the deployment of psychoanalysis: desire, modeled on sexual desire, becomes the crucial human trait, responsible for understanding of the whole of human thinking and acting. In Lacan it is "jouissance" (enjoyment, with connotations of sexual enjoyment and orgasm), in Deleuze it is "the desiring body" that become fundamental for the whole of what we would call anthropology and metaphysics. Žižek and others (including Deleuze and Guattari) have transferred this model to politics; leftist emancipator politics is defined in terms of desire and "jouissance".

Let us return two centuries back. In the wake of German idealism, the a-rationalist thesis is combined with general anti-realism. Human mind creates or co-creates reality, and the geography of the human mind s at the same time the cosmography of the whole of being. If not the human mind, then an absolute, mind-like entity, Geist, or Absolute. But, if mind creates reality, and the mind is a-rational, then a-rational forces create reality. If the human and historical are directly ontological, then the fierce passions ruling our heart and our political conflicts govern, or co-govern the very Being itself, or are just identical to it. The world is the will, as Schopenhauer proclaimed, it is an artifact of the will-to-power, as Nietzsche claimed. Let me encapsulate the idea and give it a name:

> (**A-RAT-world)** The basic reality of the world is akin to the a-rational element of human mind.

After Schopenhauer, with the late Schelling (A-RAT-mind) and (A-RAT-world) enter the scene of the late German Idealism, in the three initial decades of the nineteenth century. (German historicans of philosophy and culture have dug out interesting connections with the peak of German romanticism, but we cannot enter the topic here). According to the new creed, the central element of human mind is a drive; more importantly, a basic element of reality (including, in the first place God) is a-rational.

Be it as it may, the important ontological turn did not affect the style. Schelling's style is close to Hegel's, as Schoppenhauer's is to Kant's. They argue for the primacy of the will, in the rational framework of their targets,

Hegel and Kant. The third interesting author, their less known French counterpart Maine de Biran argues against French naturalistic philosophers-scientists (*les Ideologues*), with their own rational, argumentative and even naturalistic weapons. The style of our a-rationalists fits the rational style of their opponents; no change is introduced.

One can understand the emergence of the non-argumentative tradition if one compares these early a-rationalists with later-day ones like Kierkegaard and Nietzsche. Indeed, things have drastically changed in the middle of the nineteenth century. With Kierkegaard the affect enters the scene of post-Hegelian thinking (maybe anticipated a few decades earlier by German romanticists, Schlegel brothers and their circle). Kierkegaard has indeed been taken as the thinker of the passion, as opposed to reason, and has influenced the later development precisely in this direction.

Some authors argue it is not the final contrast in Kierkegaard, e.g. Norman Lillegard (2002:251-273): "The passion of his Knight of faith transcend rational understanding „(...) I can understand the tragic hero but cannot understand Abraham, though in a certain crazy sense I admire him more than all other men."Both Vilhelm in *Either-Or* and Abraham in *Fear and Trembling* challenge Kantian and Hegelian moral rationality. Vilhelm by insisting of a kind of absolute choice of oneself, Abraham by his action that is to be condemned within a normal rational framework.[4] But there is more. The crucial point is the publication of his *Either-or* in 1843. There, the passionate is at least *prima facie* contrasted with the rational, but this is no surprise; the true revolution happens with the style. The writer John Updike notices the analogy with the fiction writer:

> Soren Kierkegaard's method, dictated by his volatile and provocative temperament resembles that of a fiction writer: he engages in multiple impersonations, assuming various poses and voices with an impartial vivacity (1987: vii).

Kierkegaard's a-rationalism brings with itself a revolutionary change of style. The first book manifesting it is his "*Either-or*", published in 1843. Famously, three viewpoints are presented there, none of them too rationalistic (although the second one can be related to Kant. These are the hedonistic, moral and

[4] My colleague Majda Trobok asked at this juncture: is it consistent rationally to explain A's action and to claim that it is to be condemned within rational framework, and see oneself as going against the rational? Well, Kierkegaard does not himself see his own account of Abraham's decisions a belonging to rational explanation.

the religious one presented by a seducer, a moralist and a preacher character respectively. But the additional and sensational news is the style of thinking and of writing. The hedonistic viewpoint is presented through the diary of the seducer, the moral one through advice of the elder moralist, Vilhelm, writing very much like Seneca: it is the sincerity of the writer that counts as much as the cogency of the standpoint itself. The moralist speaks in a tone of advisory tracts, not in the cold abstract style of Kant. The final redemption brought by the religious viewpoint is presented through a sermon of a pastor from Jutland. "Either-Or" is the grand monument of domain-adapted style of thinking and writing, as MacIntyre has pointed out in Chapter Four of his *AfterVirtue* . There are no philosophical comments from external, neutral standpoint: the editor character, Victor Eremita, limits himself to factual, archivist information. The characters write in the manner inspired by the domain and topic: the aesthetic attitude is embodied in the seducer's diary, rather than being coldly dissected. Much more importantly, the two more "serious" standpoints are not presented in an argumentative manner at all. The brilliant stylistic exercise anticipates a fundamental turn. The idea is, in the form of a slogan: If you write about passion, write passionately.[5]

Nietzsche contributes to the trend by switching to literary style: aphorism, play with words, etc.. act against traditional (early modern) argumentative style. With "Thus Spoke Zarathustra", the idea becomes an implicit norm for the author. The norm is interesting.[6]

[5]My colleague Nenad Smokrovic objected that Kierkegaard writes in a non-philosophical style. But this is precisely the point, since he is a philosopher and is regarded as one.

Either-Or stands at the beginning of a revolution in philosophical style that has profoundly marked continental philosophy and is responsible for its present profile. It suggests that if you write about a Don Juan, you should do a diary of seduction, if you write about morals you should be moralizing, and if about religion, then preaching. We have already quoted Michael Weston in his book on *Kierkegaard and modern continental philosophy* (1994) who notes the following:

> Post-metaphysical thought in Nietzsche, Heidegger and Derrida shows certain central characteristics which have their parallels in Kierkegaard: a 'style' of writing at variance with that of the metaphysical tradition which has its rationale in the 'situatedness' of the thought whose intention is, not the representation of 'the truth', but an 'intervention' into that situation. (1994: 136).

[6]A recent work on Nietzsche by Rogério Miranda de Almeida carries the consequences to the extreme. Nietzsche should be read as a paradoxical writer, says the Preface:

> Our proposal here is, rather, to focus on paradox, or the paradoxes that Nietzsche expresses through his writing, and thus through the great diversity of perspectives and rereadings operative in the domains of art, science, religion, morality, philosophy, and culture in general.(2006: ix)

But how does one discuss Nietzsche once it is agreed that the meanings are subject to "constant play" of renewals and reevaluations? Every proposal can be turned into its contrary by the "constant play", so that the danger lurks that Nietzsche turns out as saying nothing by saying too much. This might be the price of wanting to write a-rationally about the non-rational.[7] If you write about poetry, write poetically, if you care for the future of the mankind, write as prophets did. If you care about the a-rational, banish rationality from your style. (Political activism also helps: if you write about politics, write manifestoes!). Both the writer and the prospective reader are passionate beings, since all humans are; and the passionate style plays at the deepest cords of their hearts. But, and this is philosophically central, the deep cords of the heart are in unison with the deepest chords of reality: *the passionate, aphoristic, literary style is at the same time deeply philosophical, since it manifests the deepest reality of the world*.

Of course, the turn to non-argumentative style (or at least to the style less argumentative than the style of Descartes, Locke, Leibniz and Hume) has been prepared by predecessors. By his enormous authority Kant made the idea that philosophy may and should be very difficult to read and understand compelling to the academic audience of the next generation; it is the depth that

---

It is wrong, de Almeida claims, to try to clean Nietzsche's text from contradictions:

> To be sure, the traditional commentators on Nietzsche are unanimous in admitting that his oeuvre contains "contradictions" and ambiguities. But these contradictions invoke, as often as not, "apparent contradictions" in the sense that they would be—unknown to Nietzsche himself—a logical thread carrying these texts to a coherent and continuous whole.(Ibid.)

Being contradictory and literally paradoxical is the main virtue of Nietzsche, and it is linked to his understanding of poetry and fiction:

> As a matter of fact, the principal themes of the Nietzschean oeuvre that we develop—that is, the will to power, the relation of forces, nihilism, and the eternal return—are extremely problematic and subject to diverse interpretations. And this is the case because Nietzsche himself continually reiterates, rereads, and creates new perspectives on the art of poetry, fiction, invention, interpretation, and construction. But the art of construction presupposes the force of destruction and imposes a new meaning. This is why a thought that moves in and from one relation of forces, and that is itself force, can only be expressed through the writing of paradox, that is, through the constant play of inclusions, exclusions, ruptures, renewals, and reevaluations. (2006:x)

[7] For a contrasting analytic reading of Nietzsche see Leiter's enjoyable paper (Leiter,2004) on how to recover Marx, Nietzsche, and Freud for analytic philosophy: present them as would-be naturalists, looking for explanation rather than for a "deconstruction" or "subversion" in a post-modernist vein.

counts and not the shallow formal logic (Kant's difficult style is probably the result of a historical and biographical accident, on the one hand, Kant's creativity that brought him new ideas as he wrote, on the other, his need to force the rich flow of ideas into a complicated and rigid patterns of classification, and perhaps even his fears, having to do with religiously provocative and politically challenging ideas; notice how the politically innocent necessary illusions of pure reason are just called mistakes, though in a Latinate terminology, whereas the chapter about the illusion about the provability of God's existence bears the charming title of "Ideal of pure reason"). German idealism continues the line: for it, commonsense is irrelevant (Hegel) and formal logic is alienated and plainly wrong, so, traditional logical tools (from definition to nicely sequenced arguments, with premises and conclusions detailed in full) is out of question.[8] In German idealism, especially in the work of Hegel, holism adds to it: one understands and evaluates parts only by somehow grasping the whole. In Hegel's aftermath, such holism combined with anti-commonsense and anti-scientific attitude, favoring depth over understanding, and religious and poetic influence, the grasping of the whole becomes less and less transparent; this projects on the parts as well. But now, if commonsense is irrelevant, where do you start? Natural science is seen as alienated, so scientific style is not welcome. One alternatives is provided by links to religion and mysticism (German romantics, Schelling), another by poetry. (Holism here becomes less relevant).

It seems that the basic line is that the style should follow the domain. Since Kierkegaard, as we noted, the a-rationalist program becomes methodologically demanding: philosophy should *manifest* the will, desire, the unconscious, i.e. the irrational, and not only think and talk about it. Let me put it in a formula. In particular, since a-rational domains are philosophically central, the style of philosophy should come closer to the reality of the a-rational. Here is the idea generalized and put in a nutshell. Let me call it Exemplification Constraint, EC for short:

> **(EC)** The cognitive style, the language-style and the method of studying a domain D should exemplify and manifest the nature of D itself, by following the language-style and the manner of D in its spontaneous manifestation. In particular, for a-rational domains, the cognitive style and the linguistic expression should minimize the use of (or perhaps completely eschew) traditional rationalist methods of enquiry and presentation.

[8] Thanks go to Urška Mavrić for this point.

How about rational domains, I was asked by my colleague Kati Farkas. In the more radical branches of continental thought, they are disposed with in the following way: the rational is in fact seemingly rational. Logic is just expression of the will to power. Formal logic is part of the alienated, technological world, more recently of the male dominated world: logo-centrism goes with phallocentrism. Of course, not all contemporary continentals follow this lead. But many, and the most vociferous ones do.

The main consequence of EC is that if D is non-cognitive, a-rational, irrational (e.g. the unconscious, will-for-power, desire, poetic language...), then the discourse about D inherits its characteristics, at least as much as it is possible within a professional philosophical discourse.

Consider now how EC interacts with the two principles of a-rationalism (A-RAT-mind) and (A-RAT-world) and AHO. Let me put in a series of three steps.

First, by the a-rationalist assumption (A-RAT-mind) the a-rational or irrational domains– the unconscious, will-for-power, desire– are anthropologically central. Logocentrism is bad, it is the treason of the deepest human reality.

Second, the a-rational is also ontologically central, and we get (A-RAT-world). So, the unconscious, will-for-power and desire should play a central role within ontology as well.

Third, since they are non-rational, they demands non-rational presentation, by EC. Therefore, a central ontological domain has to be presented in a non-rational, non-argumentative way. The whole of philosophical discourse—and most importantly, ontology and epistemology–itself should be passionate, poetic, aphoristic, in short, non-argumentative, at least to some extent.

This moves into the very heart of philosophy since non rational domains are central since Nietzsche, Kierkegaard, Marx and Freud. Marxism has been since its beginnings oscillating between its Hegelian origin and the idea of scientific understanding of social reality and the "scientific socialism" as the alterntive to mere utopia: early Marx vs. *Das Kapital*, Korsch and Bloch vs. dialectical-cum-historical materialism, Heideggerian Marxism vs. Althusser. And the style has been following the characterization of the domain: objective historical development vs. suffering in alienation and appeal to the forces of revolutionary subjectivity and authenticity. On the more popular side, feminism has contributed to a political denigration of the rational as phallocentric and patriarchal; not all feminists claim this, but those that claim have attracted most attention. (Again, I apologize for brevity and generalizations, but I need to paint a big picture on a small canvas. In the next section I mention ???

The crucial role of EC lies in explaining the non-argumentative, poetic and sometimes logic-unfriendly style of a lot of mainstream continental writing. The style is not just the style of writing, it is a matter of the way of thinking. Analytic colleagues get nervous about it, and the malicious among them see it a symptom of craziness. In contrast, EC presents it as a principled choice, far from craziness. Since poetry and literature in general has been traditionally the medium of passion and affectivity, EC will naturally favor a turn to literally culture away from the scientific one. Of course, once the EC has become a norm, it will tend to recruit authors with literary talent, and the circle (virtuous or vicious, depending on the taste) will form itself. Of course, EC is not always followed *a la lettre* but its pressure often results in a discourse that is geared at least in part to exemplifying the passionate, non-rational This is the heritage of the nineteenth century great a-rationalists. The next act happens in the twentieth century, beginning sometime in the late twenties, early thirties, in the troubled, disoriented Germany, poised for a dangerous adventure, that will lead it into a catastrophe.

The thinker as poet: from phenomenology to Heidegger and to the post-heideggerian scene

> PHILOCONTE: Qui irait croire la déclaration ridicule de Carnap quand il dit que le métaphysicien est un artiste raté? Heidegger au contraire nous a montré qu'entre le penseur et le poète il y a des liens si profonds que l'on ne peut plus penser, comme Platon, en termes d'un partage entre ceux qui cherchent la vérité et les producteurs d'apparence.
> Engel (1997 :16)
>
> Who would believe in the ridiculous statement of Carnap that the metaphysician is nothing but an unsuccessful artist? Quite the opposite: Heidegger has shown us that there are deep ties between the thinker and the poet, such that one cannot, like Plato, think in terms of the division between those who search for truth and producers of appearance. (my translation)

We now think of Kierkeggard and Nietzsche as extremely significant authors, but one should bear in mind that they were marginal on the academic scene of their countries. The first never made an academic career, the second started it and abandoned it. Their work was influential, but the academic life was moving in the more boring tracks of neo-Kantianism, until phenomenology

was born; but the phenomenology itself was highly abstract and theoretical, initially geared to answering the same questions that neo-Kantians were addressing, and produced in a very dry, non-emotional, academic Germanic style. It is only with Heidegger that situation changes. How it happened is a matter for historians, but his institutional academic philosophical success was certainly to a large extent due to his erudite investigations into history of philosophy, especially ancient Greek and modern German, that preserved for him the aura of traditional university professor of philosophy, in contrast to outsiders like Kierkegaard and Nietzsche. On the other hand, his daring and original non-traditional ideas have procured to him a talented and responsive audience thirsting for good philosophy in dark times of Nazi Germany; his own survival in the circumstances being, infamously, due to less than impressive political moves of his.

It was Heidegger who turned phenomenological investigation into an analysis of the existential relation between Dasein and Sein, and then into a poetic-hermeneutic investigation into human destiny. He started in *Being and Time* with the idea of human involvement with the world, as an antidote to skepticism. (One route from there is the pragmatist one, taken by many of his American interpreters). In later works the involvement is characterized as "living poetically"(dichterisch). Our motto, "Being's poem, just begun, is man.", taken from his *Aus der Erfahrung des Denkens* combines all the elements we were talking about. First, the idea that human being belongs to the very ground of being, that it is ontologically most intimately connected to it. Second, that the relationship between the two is primarily poetic, as opposed to say, epistemic, or logic. Man is the poem, *Gedicht* of Sein. Which reminds us of the idea that "poetically dwells the mean on the Earth", taken from Hölderlin, and philosophically developed by our philosopher. And of course, the philosopher is expressing this in a poetic way, not in cold theory, nor in a sequence of arguments. Just in case one might think it is an isolated fragment, let me give its context:

> When the early morning light quietly grows above the mountains. . . . /
>
> The world's darkening never reaches to the light of Being./
>
> We are too late for the gods and too early for Being. Being's poem, just begun, is man. /
>
> To head toward a star—this only. /To think is to confine yourself to a single thought that one day stands still like a star in the world's

sky. (1971:4.)

Albert Hofstadter, the translator of *Aus der Erfahrung des Denkens* has entitled it "The Thinker As Poet", because, in his opinion, here "the thinker does what a poet does—dichtet.(Ibid. xi)." Ironically, given his change of the title, he then continues:

> "Heidegger's original title for this piece was "*Aus der Erfahrung des Denkens*—"From the Experience of Thinking"—and one should read it as such, as the uttering of realizations that have come out of a long life of discovery of a way of thinking that belongs to life in its fullness as genuinely human.(Ibid., xii).

So, how did this development from Husserlian phenomenology to the poetic style of the late Heidegger take place. Let us start from phenomenology. Note that the *phenomenological description* was meant as a report on the given, not as any kind of non-argumentative procedure. In the more careful use, it provides evidence for further philosophizing and arguing. But on the more risky side, it offers opportunity for smuggling substantial philosophical views into "pure" describing (analogous to "theory-laden" perception in the debates of philosophy of science). Phenomenology has been promoting a "neutral" description of our experience. However, in Husserl and then in Sartre, the presumed descriptions are very much colored by philosophical theory. Unfortunately, since they are presented as descriptions, this presentation apparently frees the philosopher from the obligation to argue; he is just "presenting evidence" in the form of presumably neutral description. This dogmatism of presumed description is strengthened with increasingly difficult style, acceptable (and perhaps even demanded) in an academic climate formed by Kantian tradition of heavy, convoluted style. The convoluted style of "Being and nothingness" nicely illustrates the danger: a clear line between describing on the one hand and argumentative theorizing is never drawn.

In Heidegger the dogmatism of presumed description encounters EC and the gap widens. In Sartre, it appears in *L'Etre et le neant,* and then meshes with his litteral project. Existentialism continues with linking philosophical writing with (very successful) literature; as Roberto Bernasconi nicely pointed out in a talk, most people have got their first impression of existentialism from *Nausea*, and the novel played the crucial role in its history.

However, before sliding into poetry in late Heidegger, the style went through a very important phase: *non-argumentative hermeneutical reconstruction* of classical sources. It is usually characterized by two features: First, what is re-

constructed are not particular arguments of the classics; in the best case it is a general orientation of arguing, but even this is mostly left implicit. Second, the reconstruction is full of highly suggestive, never explicitly argumentative, and often clearly non-argumentative moves The reconstructed items are in the good case meaning of their main theses, in somewhat less good case, simply meanings of crucial terms, but the reasons for accepting (or rejecting) a view are in the rule not made explicit. The appeals to authority of the great philosophers or thinker in general of the past (ranging from Presocratics, through Plato, to Kant or Hegel) are rarely presented as such, but are masked as invocations of great truths with almost mystical appeal, with no rational explanation of why we should trust, say, Heraclitus rather that Chrysipus, or Plato rather than Aristotle. In all this development, the a-rational is firmly affirmed:

> Thinking begins only when we have come to know that Reason, glorified for centuries, is the most stubborn adversary of thought. (2002:199).

Finally, we get the poetic glaze. Poetry joins philosophy, as illustrated by our motto. Indeed, for Heidegger the traditional forms of rationality are all on the side of the fallen humanity: classical logic, scientific thinking, technological intelligence and rational planning. In contrast, the authentic forms of *Dasein* are famously given in the early work through existential, emotionally colored attitudes, above all the attitude of care. In later work a crucial role will be played by art, and in particular poetry, and the language of philosophy will tend to imitate the poetic language. Here is, for instance, how Heidegger formulates his suggestion about the end of philosophy:

> The old meaning of the word "end" means the same as place:"from one end to the other" means from one place to the other. *The end of philosophy is the place*, that place in which the whole of philosophy's history is gathered in *its most extreme possibility.* End as completion means this gathering (1978:375).

As we would expect, it is seriously multiply ambiguous text; the main term, "end" can mean – finish, goal, place, and the suggestion comes as a surprise: the end of philosophy is the place. Next, we have massive use of poetic figures, with the use of evocative appeals to thing like "*extreme possibility*".[9]

[9]Here is the wider context of the claim:

If we try to reconstruct the deeply hidden argument, we obtain the following:

> 1. The old meaning of the word "end" means the same as place
>
> therefore,
>
> 2. *The end of philosophy is the place,*
>
> (3. *Place is the place of gathering*).
>
> therefore
>
> 4. The end of philosophy is that place in which the whole of philosophy's history is gathered in *its most extreme possibility.* (End as completion means this gathering.)

But how does 2. follow from 1. ? Only because "the old meaning of the word "end" means the same as place"; but the old meaning of the word "silly" is blessed, and nobody would accept this as final evidence that it is a fine thing to be silly. It seems that there is no point in reconstructing Heidegger's thinking in such an argumentative way. Either the reader gets the poetic suggestion, or the labor is lost. In short, what started in 1843 as an experiment in style, has ended in the early 20th century as a transformation of central philosophical disciplines.

Let me briefly further illustrate the working of the same thought through the issue of conceptualizing, conceptual understanding and theory-building

---

> Throughout the whole history of philosophy, Plato's thinking remains decisive in changing forms. Metaphysics is Platonism. Nietzsche characterizes his philosophy as reversed Platonism. With the reversal of metaphysics which was already accomplished by Karl Marx, the most extreme possibility of philosophy is attained. It has entered its final stage. To the extent that philosophical thinking is still attempted, it manages only to attain an epigonal renaissance and variations of that renaissance. Is not then the end of philosophy after all a cessation of its way of thinking? To conclude this would be premature.
>
> As a completion, an end is the gathering into the most extreme possibilities. We think in too limited a fashion as long as we expect only a development of recent philosophies of the previous style. We forget that already in the age of Greek philosophy a decisive characteristic of philosophy appears: the development of sciences within the field which philosophy opened up. The development of the sciences is at the same time their separation from philosophy and the establishment of their independence. This process belongs to the completion of philosophy. Its development is in full swing today in all regions of beings. This development looks like themere dissolution of philosophy, and in truth is precisely its completion. (1978: 375)

in matters of art. Our source is Gadamer, who is the most pro-argumentative of all Heideggerians.[10] In a recent paper the Canadian philosopher Jean Grondin interpreting Gadamer claims "that it is not possible to grasp conceptually the play of art. What we can do is to play along, to participate and to take part in the play" (web:27). Let me call it Impossibility thesis. " When we hear music, we instinctively start singing and dancing." (Ibid.), continues Grondin. If the thesis were taken seriously, as it merits to be taken, it would entail that there is no way to write about the play of art in a distanced, non-playful and non-artistic way, in the manner that is usual for the analytic approaches to art. If one writes about the play of art one should write playfully and artistically, one should "participate and (...) take part in the play". The Impossibility thesis, very much in line with EC, fits nicely with Gadamer's fundamental thesis, according to which is it the play itself, in this case the play of the work of art, that guides our involvement, rather than our subjectivity playing the leading role. If we extend this fundamental thesis to the meta-level of theorizing about art, we get the view that it is the playful nature of the work itself that should guide the way of theorizing about it (although Gadamer himself does not write in playful fashion, and is very much in love with arguments). Moreover, if successful, the work of art changes us, and the change must re-appear in the manner in which we think of it; the manner must bear a stamp of the experienced work itself. And this change is then normally thought of pervading our understanding and our manner of thinking. This is not how many serious philosophers of art, from Kant to Levinson, have proceeded. They have sought precisely conceptual understanding, and their writing is not playful at all. On the other hand, the Impossibility thesis seems to capture nicely a lot of practice in contemporary continental philosophical writing about art and literature, and also shows its bite in the non-philosophical theoretical writings (literary theory, art theory), in which theoretician's often, write in a literary fashion, re-enacting, so to speak, the works of art they are talking about. It fits Derrida's idea of philosophy as *écriture*; what has started in his early work on Husserl, as an examination of the semiotics of the voice as opposed to the letter or writing (the literary sense of "*écriture*"; Derrida would love the pun), has become an invitation to philosophers to pass to *écriture* in the sense of fiction, to become "*é*crivains"; and the followers, have of course, obliged.

Let me conclude this brief sampling by noting the radical variant of the A-RAT and exemplification, to be found in Lacan who, as already mentioned,

[10] thanks go to Darjana Nastić whose thesis introduced me to this debate.

combines the play of words derived from the Freudian tradition of the study of slips of tongue with poetic variations on it, unexpectedly enriched by mathematical looking formulae and diagrams, which, however, in their interpretation offer a wide space to freedom, multiple ambiguity and other typical poetical virtues. What is the link with EC? First, I find Lacan's famous dictum: There is no metalanguage! to be a fine variant on EC. If there is no metalanguage, there is no neutral, rationally controlled, dispassionate point of view from which we can think, speak and write about the non-rational domains (it is not the only reading of the dictum, but it is hopefully a plausible one).

For Lacan's favorite area, the unconscious, the morals is clear: write in the style of the discourse on and around psychoanalyst's sofa, use play of words, form of words inspired by free associations, slip of tongue and similar sources, rather than in the dry, quasi-scientific original Freudian style. Shoshana Felman I think rightly speaks of Lacan's "poetic" rejection of concept(s) and knowledge (where in her writing "poetic" implies "inclusion of *madness into the very style of writing*".(2003, passim).[11]

The crucial role of EC lies in explaining the non-argumentative, poetic and sometimes logic. Thanks to EC, continental philosophy has been vastly more successful in catering to the immediate and pressing concerns of arts and humanities than its analytic rival. Its readiness to tolerate, if not to encourage essayistic style, in particular a mixture of literary and philosophical manner of writing, its constant reference to matters cultural and artistic, its willingness to give up the truth-directedness, the goal of clarity and elimination of ambiguity in the interest of other goals (artistic finesse, political militancy or provocation and the like) has made it much more acceptable to the departments of English, cultural studies or film theory.

Finally an illustration from Derrida. In his *Given Time: I. Counterfeit Money*, he sets himself to investigate the paradoxes of exchange, gift and giving. A

[11]The usual reading of the dictum (from the *Seminar* of November 1966, and repeated for instance, in the preface to the pocket edition of *Ecrits*, and often in *Autres Ecrits*, Seuil, 2001, e.g. at p. 18) stresses that there is not Archimedean point outside of a given discourse, from which one could talk about that discourse. This reading does suggest what we call EC below: if you want to talk about some discourse D (of passion, of politics, of religious exaltation), your own talk will not be "outside" D, less metaphorically, will have characteristics of D. The *Compendium of Lacanian term* (2001: 202) appeals to the following alleged comment that Lacan gives himself:

> 'Any statement of authority has no other guarantee than its very enunciation, and it is pointless for it to seek it in another signifier, which could not appear outside this locus [of the signifier] in any way' (2006: 310).

I was not able to locate the reference, neither in the French original nor in Fink's translation.

In *Autres Ecrits* Lacan comments the slogan with "there is no Other of Other" (325).

real present, a "true" gift should be accompanied by no expectation of return, and accepted with no checking and doubt. But gifts are at the same time caught in expectations of reciprocity, so the true gift is paradoxical and impossible. So, there is an element of madness in giving and reciprocating, the "madness of economic reason" as Derrida characterizes it in the title of the chapter. (The chapter is on Marcel Mauss and his classical book on the gift.) Immediately, EC shows its teeth: Theory, i.e. the distanced, non-mad reflection about gift is powerless (1994:30), in this "sleepwalk at the limit of the impossible". So, thinking about the gift means entering the "destructive circle" of the transcendental illusion. (1994:35). It involves giving "gages", not just tokens of faith, but guarantees, acts of taking "personal risks", and this intellectual "sleepwalk" will reflect on and in the style of writing: „the discourse on madness appears to go mad in its turn, alogos and atopos". (1994:35).

In a way this is the farthest point that a serious non-argumentative strategy could reach apparently following the lead of EC: if you write about madness write (at least a bit) madly. More than that would destroy any seriousness. So much about our main hypothesis, that the (A-Rat) and EC offers a good reconstruction and partial explanation of the birth and success of non-argumentative tradition in philosophy in the last two centuries Further explanations should be historical and sociological, telling us about the external circumstances that made it so successful. Let me just add that no simple-minded explanation in terms of political affiliation is going to work. Some authors (e.g. Emmanuel Faye, 2005) have been offering explanations pointing to Heidegger's extreme right wing sympathies and engagements, others (e.g. D. Eribon, 1992) have mused about the sociology of French a-rationalist scene pointing to the involvements with communism; if we put them together, we see the common mistake of connecting a-rationalism with a particular political agenda. Obviously, the a-rationalist tradition is not politically tied to any particular segment of the extremely wide political spectrum, ranging in its political choices all the way from Hitler through religious center-right and atheist center-left to Lenin, Mao and Gandhi.

4. Conclusion

Honoring Pascal's work on the continental-analytic contrast, this paper discusses the non-argumentative tradition in continental philosophy; it is one

of its central traditions, but not the only one. From Brentano and Husserl to Habermas there have been other lines of thought, bristling with argumentation, but they are not the topic of the paper. Let me first summarize our proposal for understanding the non-argumentative tradition in continental philosophy. The more extreme works in this tradition are sometimes criticized by more argumentatively-minded philosophers as non-philosophy, fiction, or simply as nonsense. In contrast, we have tried to show here that the story is more complex, and have tried to find principled explanation of why good philosophers would turn to a way of writing that is consciously using procedures typical of literary and poetic style, involving, and even praising multiple ambiguity (without indications about disambiguating), massive and central use of poetic figures (again without clear advice about decoding them), blocking reconstruction in argumentative style, and, when using arguments, as all philosophers at the end of the day have to do, hiding it deeply in the poetic text. (Again, I am not claiming that most of continental philosophy just became literature, this would be a caricature.)

The proposal of explanation has three steps. First, it reverts to the importance of the a-rational element in the tradition, say, desire, will to power or drive, which is highly valued and taken to be central for human psychological life. Second, it points to the elements of anti-realism or at least flirtation with it in the main authors: they tend to transfer the diagnosis about the importance of the a-rational element from the mind to the world. Finally, it seems natural that at least some philosophers who made the first two steps, would also have reservations about rational, explicitly argumentative methods of investigating and presenting the central elements and structure of the mind and world as they see it. It the world (or at least our world) is constituted by drive and will to power, if our mind is not only lead by them, but constituted by them, wouldn't a philosopher betray his or her insight by presenting all this in a cold, rational manner? Rather, the style should follow the domain of investigation, the style of philosophy should come closer to the deep reality of the a-rational by exemplifying and manifesting it. If man is the "poem of Being", then the essence of both of them should be expressed poetically. The cognitive style and the linguistic expression should minimize the use of (or perhaps completely eschew) traditional rationalist methods of enquiry and presentation. This is valid for the central authors, whose short quotes we used as our examples (too few, unfortunately, but the space is limited). This way of doing philosophy can lead to caricature, and we have avoided the worse exemplars, often imitators of more serious philosophers. But it can also be used in a more moderate fashion, like for instance, in Adorno and Foucault, where

stylistic brilliance did not destroy the argumentative scaffolding. Most continental philosophers still argue with the reader and with their predecessors and opponents, but arguments tend to be less explicit, and are often being immersed in the medium of non-argumentative style, ranging from poetic flights to political invective. Let me reiterate my main hypothesis: the combination of the preference for the a-rational, and the idea that the style of thinking and presentation should mimic the a-rational domain and exemplify and manifest its characteristics offers a good reconstruction and partial explanation of the genesis and success of the non-argumentative tradition in continental philosophy.

Finally, assuming that we, contributors to the volume, prefer the argumentative style, what can we learn from the non-argumentative tradition and its success? Well, that if you want to persuade a wider audience, it is sometimes best to hide the rigor of one's philosophical argument and add some literary flavor. On the other hand, if there is to be a successful dialogue between analytic and continental philosophy, it is more likely to happen between the analytic philosophers and the more argumentative among their continental colleagues. The dialogue might look as an optimistic continuation of Engel's *La Dispute*; but this is for the moment just a hope.

5. References

Derrida, J. (1994), Given Time: I. Counterfeit Money (Vol 1, University Of Chicago Press.

Derrida (1998)'The Retrait of Metaphor', trans. F. Gasdner, in J .Wolfreys (ed.), The Derrida Reader: Writing Performances, Edinburgh University Press.

Engel, P. (1997), La Dispute, Une introduction, à la philosophie analytique, Les Éditions de Minuit.

Eribon, D. (1992), *Faut-il bruler Dumezil?: Mythologie, science et politique*, Flammarion.

Faye,E. (2005), Heidegger, l'introduction du nazisme dans la philosophie : autour des séminaires inédits de 1933–1935, Albin Michel.

Felman, S. (2003), Writing and Madness, Stanford University Press.

Grondin, J. (web), "Play, Festival and Ritual in Gadamer: on the Theme of the Immemorial in his Later Works", `http://mapagweb.umontreal.ca/grondinj/` pdf/play_festival_ritual_ gadam.pdf

atzimoysis, A. (2009) "Emotions in Heidegger and Sartre", in Goldie, P. (ed.), *The Oxford Handbook of Philosophy of Emotion*, Oxford University Press.

Heidegger, M. (1971), Poetry, language, thought / Martin Heidegger; translated and introduction by Albert Hofstadter. New York : Harper & Row,.

Heidegger, M., (1978), "The End of Philosophy and the Task of Thinking", in Basic Writings Routledge.

Heidegger, M,. (2002). "The Word of Nietzsche: 'God is Dead"' in his *Off the beaten track*, Cambridge University Press.

Heidegger, M. (1971), Poetry, language, thought, Harper & Row.

Kierkegaard, S, (2008), Fear and Trembling, Wilder Publishing.

Lacan, J., (1973), *Seminaires, Livre XI Les quatre concepts fondamentaux de la psychanalyse 1964*, Éditions du Seuil.

Lacan, J. (2001), *Autres Ecrits,* Éditions Du Seuil.

Lacan, J., (2006), *Ecrits*, W W Norton & Company, Inc.

Leiter, B. (2004), "The Hermeneutics of Suspicion: Recovering Marx, Nietzsche, and Freud" in Brian Leiter (Ed.) The Future for Philosophy, Clarendon Press, Oxford,

Lillegard, N. (2002), "Passion and Reason: Aristotelian Strategies in Kierkegaard's Ethics", The Journal of Religious Ethics, Vol. 30, No. 2 (Summer, 2002), pp. 251-273).

MacIntyre , A, (1984), AfterVirtue, University of Notre Dame Press.

Miranda de Almeida,R., (2006), Nietzsche And Paradox, State University of New York Press.

Murphy, S,; Glowinski, H; Marks, Z, M. (2001), *A Compendium of Lacanian Term*, Free Association Books.

Nancy, J-L. (2000), *Being Singular Plural*, Stanford University Press

Rorty R. (1980), "Philosophy as a Kind of Writing: an Essay on Derrida in Consequences of Pragmatism: Essays, 1972-1980, 90-109. Minneapolis: U of Minnesota P, 1982.

Rorty, R (2007), *Philosophy as Cultural Politics*, Cambridge University Press.

Richard Rorty & Pascal Engel (2007), *What's the Use of Truth?*, Columbia University Press.

Smokrović, A., (2013), "Chomsky and Foucault on human nature – a perspective for reconciliation", *Balkan Journal of Philosophy*, v. 5/ 2, 175-181

Updike, J. (1987), *"Foreword" to Kierkegaard, S, The seducer's diary*, Princeton University Press.

Weston, M. (1994), *Kierkegaard and Modern Continental Philosophy~ An Introduction*, Routledge.

Wheeler, S. C.III, (web), Philosophy as Fine Art (draft).

PART SIX

Further papers

54

L'argument solipsiste et sa postérité anachronique

JEAN-MAURICE MONNOYER

La *vérité du solipsisme* a gardé une sorte de prestige négatif, bien qu'elle se trouve entre le scepticisme et le dogmatisme dans une position inconfortable, pour ne pas dire incongrue. — Mais de quoi est-elle la vérité ? Il y a plusieurs manières de l'envisager, comme lorsque l'on se demande "ce que le solipsisme veut dire" : une question que Jacques Bouveresse a examinée en France, de façon presque définitive, depuis *Le Mythe de l'intériorité* (1976). Il ne s'agit pas là d'une question qu'on devrait limiter à la seule question de l'impossibilité du langage privé, ni non plus qu'il faudrait réduire à celle d'un jeu philosophique paradoxal suscité par telle ou telle formulation de Wittgenstein que j'examine ci-dessous pour les tester l'une après l'autre et dans leur consistance respective (qu'on me pardonne par avance la lourdeur de cet examen). Par-delà l'idéologie du langage ordinaire qui a préparé celle du monde ordinaire, c'est bien la métaphysique de la connaissance qui se trouve brutalement mise en balance avec la recevabilité d'un argument de ce genre, si bizarre que beaucoup le considèrent comme assez tordu. Aujourd'hui cependant, avec le renouveau des études sur le premier Wittgenstein, une forme d' indulgence a paru à son égard qui semble conforter la thèse d'une réelle postérité de l'argument solipsiste[1]. Je me propose ci-dessous de le discuter

[1] : Je pense au long article de James LEVINE, "Logic and solipsism", in *Wittgenstein's Tractatus, History and Interpretation*, P. SULLIVAN & M. POTTER eds, Oxfor UP, 2013, pp. 170-238, discutant lui-même Michael KREMER, "To what extent is Solipsism a Truth ?", in B. STOCKER, *Post-Analytic Tractatus* , Aldenshot, Ashgate, 2004, pp. 59-84, et Peter SULLIVAN, "The "Truth" in Solipsism and Wittenstein's rejection of the apriori", *The European Journal of Philosophy* ,4, 1996, pp. 195-219.

en partie et dans la mesure où cet exposé peut au moins présenter le débat sous-jacent.

Que peuvent au juste *délimiter* "les limites de mon monde" ? La phrase allemande est provocante *: die Grenzen der Sprache (der Sprache die allein Ich verstehe) die Grenzen* meiner *Welt bedeuden* (5.62). Ce génitif a quelque chose de furieusement affirmatif, comme chez Lichtenberg, parce que la parenthèse qui le précède est déconcertante : *le seul langage que je comprenne* n'est pas le langage que seul je comprends, mais le langage tout court, celui du béotien comme celui du philosophe. Il est clair, le monde wittgensteinien est spécialement idiosyncrasique, bien qu'il se soutienne aussi d'une contingence déclarée des faits du monde extérieur, lequel n'a rien à voir (semble-t-il) avec cette proclamation singulière. Placée dans un aphorisme mi kantien, mi diplomatique, elle pourrait pour un lecteur superficiel se confondre avec une thèse soutenant qu'il est impossible pour une chose d'exister, par exemple une chose vue, indépendamment du fait qu'une personne ne la pense ou ne la nomme. Car si le solipsisme est vrai, que peut-on identifier comme une connaissance ? La seule connaissance qui en découle serait qu'il y a motif à parler d'un sujet métaphysique *cognitivement identifiable*. Sans quoi, *Das Ich der Solipsimus schrumpft* (...) : "le Je du solipsisme se dissiperait (...)", nous dit le texte, comme si nous courrions un risque quelconque. Wittgenstein s'amuse dans cette allusion onomatopéique et à peine cryptique à réduire le "Je" à *un point sans extension*. Mais c'est le troisième chapitre de la *Theory of Knowledge* que Russell avait rédigé en mai 1913, qu'il vise directement : *the experiencing subject* n'est pas vraiment celui qui réchapperait d'une conception idéaliste, avait écrit Russell, c'est au contraire celui qui ferait une expérience directe à l'instar de celle de Meinong dans sa théorie des objets d'ordre supérieur (p. 43) [2]. Dans ce cas, et si tous les *faits cognitifs* écrit Russell sont ceux qui impliquent une relation d'acquaintance, il n'y a plus de sens à séparer ceux qui m'invitent à une expérience directe au présent, et ceux dans lesquels un acte de pensée, lui-même bien réel, nous dispense d'asserter l'existence d'un contenu actuel. Wittgenstein s'était dit "choqué" (le 14 mai 2013) par cette invitation à un réalisme phénoménologique du contenu psychique. Russell a bien traité du solipsisme comme résumant une expérience *totale* (*all embracing*) (*op.cit.* pp.10-11, 13), perceptuelle et aussi relevant de notre compréhension des données mathématiques et abstraites, tout en mettant en doute que l'on puisse prouver que

[2] : Bertrand RUSSELL, *Theory of knowledge*, The 1913 Manuscript, Routledge, 1984, p.21. Russell est beaucoup plus clair que Wittgenstein dans son exposition et sa critique, qui le conduit à s'opposer aux nouveaux réalistes américains.

la négation que l'énoncé solipsiste soit vraie.

Wittgenstein résume cette question que Russell a laissée irrésolue par une interrogation plus forte qui interdit selon nous et selon P. Sullivan qu'on puisse penser à une forme d'idéalisme transcendantal. Puisque ce sujet dont il parle dans le *Tractatus* n'est pas vraiment un sujet de la connaissance ou d'acquaintance, quelle place occupera-t-il face au monde des sciences de la nature ? — *Wo in der Welt ist ein metaphysisches Subjekt zu merken ?* (T. 5.663). ("Où pourrait-on dénicher dans le monde un sujet métaphysique ?"). Impossible de le savoir en première intention : nous savons uniquement que ce n'est pas un sujet psychologique qui est à trouver (5.641). Pourtant le fait qu'il y ait un sujet métaphysique présupposé par le langage n'implique pas par lui-même que la thèse du solipsisme soit une vérité métaphysique.

Considérons donc ci-dessous, une nouvelle fois, les énoncés tels qu'ils apparaissent dans le *Tractatus Logico-philosophicus*. Dans la mesure où c'est un peu la profession de foi du philosophe professionnel qui est rendue caduque par l'énoncé emphatique n° 5. 6 — il n'est pas sans intérêt, par le biais de cet exercice, de considérer justement le genre de profession de foi grand style qu'un philosophe de profession ne peut pas faire. En quoi et pourquoi devient-elle inopérante, c'est en fait ce qu'il nous explique. Cette exigence n'est nullement, bien sûr, une préconisation technique. La question posée pourrait plutôt se formuler ainsi : où se fait la démarcation entre ce qu'on a encore "envie de dire", pour soutenir un *point de vue réaliste* par exemple, et le constat amer qu'il n'y aurait pas de connaissance philosophique qui puisse transgresser les limites de la logique et du langage : *Wir können in der Logik nicht sagen : das und das gibt es in der Welt, jenes nicht* (T: 5.61) ("Nous ne pourrions pas dire en logique : dans le monde il y a cette chose-ci et cette chose-là, et non pas telle autre"). En résumé si l'on soutient qu'aucune théorie de la vérité et qu'aucune logique ne permettent de transcender notre expérience, comment pourrions nous ne pas en conclure que l'énoncé solipsiste et l'énoncé qui le nie sont, tous les deux, privés de sens (*unsinnig*)[3].

Etudier *de parti-pris* l'affirmation solipsiste et son *retournement*, offre ainsi apparemment un angle d'attaque assez mal choisi sur le plan épistémologique. Certes, ce parti-pris nous permet de comprendre la version sophistiquée de ce double mouvement qu'a étudiée Saul Kripke, dans les *Recherches Philosophiques*, pour conclure justement comme il l'a fait au scepticisme de son auteur (1982). Mais quoique l'argument postérieur enregistré sous le label de "l'impossibilité

[3] : Telle est l'hypothèse du solipsisme *sémantique* qu'a développée Frascolla, et sur la quelle nous revenons plus loin.

du langage privé", paraisse constituer le complément nécessaire de l'affirmation soutenue dans le *Tractatus* — et à la différence de J. Hintikka et de B. Williams [4], qui plaident en faveur d'une résolution définitive de ce problème dans la chronologie de l'œuvre —, il me semble qu'il n'y a pas vraiment d' *opinion* philosophique déclarée que nous aurait léguée Wittgenstein à ce propos. Il n'y a guère plus de démission philosophique intégrale, comme quelques-uns de ses interprètes l'ont soutenu. Les remarques abondantes de l'été 1916 dans les *Carnets*, qui tournent autour du même problème, ne sont pas ironiques. La situation a changé entre l'examen que fait Wittgenstein de l'idéalisme solipsiste tel que défendu par Russell en 1913, et ce qu'il en dit à partir de 1915. A y regarder de près, c'est tout l'inverse en effet : on se trouve devant une affirmation hyperbolique, mais qui — si elle est accréditée par son lecteur — entraîne le désengagement de toute opinion exprimable comme une opinion personnelle. S'il y a donc effectivement un lien avec la critique du "langage privé", ce lien reste indirect : Wittgenstein serait en effet parvenu à montrer, tout en le disant sous la forme presque outrée d'une affirmation solipsiste, que la déviation coupable de la philosophie consiste à tenter de sublimer un moment d'égoïté qui serait dévolu à la pensée en acte ; mais il n'y parvient qu'à défaut d'une immersion mystique ou empathique dans le réel[5]. D'où cette affirmation complémentaire : *Das denkende, vorstellende, Subjekt, gibt es nicht* (5.631). *Il n'y pas de sujet pensant, de sujet de la représentation*. Est-ce pour se défausser de l'influence de Schopenhauer qui est notée par beaucoup de commentateurs, comme dejà Hacker (1972) après E. Anscombe, et plus récemment A. Nordmann (2005) ?

L'idée qui préside à cette conception des choses peut être résumée (très grossièrement) ainsi : que toute croyance soit exprimable *comme une opinion privée* rend du même coup cette opinion rigoureusement injustifiable dans son contenu. Rien n'est communicable de ce que je sais seul savoir. Cette même opinion, tout la justifie et rien ne lui rend justice. Par exemple si je comprenais : "Je suis mon monde" (5.63), comme voulant dire : "Je ne sais

[4]: J. HINTIKKA, "On Wittgenstein's Solipsism", in *Mind* (67), 1958 ; B. WILLIAMS, "Wittgenstein and Idealism", in *Understanding Wittgenstein* (ed. par G. Vesey), Londres 1974.

[5] : On pourrait ainsi juger aujourd'hui, en défense de la lecture qu'a faite Bouveresse, que *L'Homme probable* — ce roman citationnel unique en son genre, qui est aussi un traité dramatique sur la situation de l'intellectuel contemporain — illustre également la position que nous voudrions définir.

Cf. *L'Homme probable, Robert Musil, le hasard, la moyenne et l'escargot de l'histoire*, L'Eclat, 1993, réed. 2013 (voir, en particulier, sa dénonciation de l'individualité héroïque dans le chapitre V). Que ce moment d'égoïté soit fichtéen ou phénoménologique importe peu dans le cas qui nous occupe.

rien dont je ne fasse l'expérience par moi-même". La tentation serait alors de vouloir passer de l'intentionnalité phénoménale au contenu mental, quand c'est le second qui justifie et fonde la première [6]. Son noyau théorique ainsi désamorcé, la phrase ne dit plus rien, puisque sa prétention se veut indépendante du medium conceptuel qui permet de l'asserter : elle ne serait qu' *une justification sans raison*. A la lettre toutefois, pourvu que cela ait encore un sens de le dire de cette manière, on ne peut pas faire comme si, chez Wittgenstein, n'étaient pas clivées *l'assertion* d'une proposition et *l'expression* de son contenu propositionnel ; il y a là un dilemme hautement sensible que le *Tractatus* ne veut surtout pas trancher. On sait qu'il dénonce la manière même d'écrire le signe de l'assertion. D'où aussi la difficulté d'une reconstruction orthodoxe de l'argument tel qu'il nous est présenté. C'est probablement que le "réalisme interne du langage" (comme on s'en explique plus loin) ne coïncide pas avec la notion d'un *langage de la pensée* : ce qui rendrait prophétique la stratégie qu'a retenue Wittgenstein au sein des discussions les plus récentes.

1. *L' acceptabilité de l'argument solipsiste.*

En tant que telle, la structure de l'argument est platement circulaire, et on ne l'expose pas sans peine. La proposition que Wittgenstein consigne en 5.6, est d'abord là aussi celle du génitif epexégétique, souligné par lui en italiques : "*Die Grenzen meiner Sprache* bedeuten die Grenzen meiner Welt" (*Les limites de mon langage* signifient < ou dénotent > les limites de mon monde). Mais ce *n'est pas un énoncé d'identité*. La phrase n'énonce pas que les limites de l'un "sont" les limites de l'autre, ni qu'il s'agisse peut-être des mêmes limites [7]. De cet énoncé dérivent, dans le *Tractatus*, onze numéros qui commentent celui-là [8]. Parmi les éclaircissements importants de 5.6, on trouve des expressions comme celle que nous avons citée ci-dessus : "Je suis mon monde (Le microcosme)" (5.63) ; ou bien, "Le Je fait son entrée en philosophie grâce à ceci que "le monde est mon monde" (5.641). Mais, à dire vrai, le concept de "limite" est l'objet principal sur lequel porte la réflexion de Wittgenstein, comme en témoignent 5.61 et 5.62 — ou encore 5.632 : "Le sujet n'appartient pas au monde,

[6] : Sur cette thèse, voir Adam PAUTZ, "Does Phenomenology Ground Mental Content ?", in Uriah KRIEGEL, *Phenomenal Intentionaliry*, Oxford UP 2013, pp. 194-234.

[7] : David FAVRHOLDT, *An interpretation and critique of Wiuttgenstein's Tractatus*, Munksgaard, Copenhague, 1967, pp. 145-189, a offert le premier exposé détaillé de cette question qui reste selon moi l'un des plus pertinents, bien qu'il ait paru après le livre de Stenius.

[8]: Cf. l'édition critique, *Tractatus Logico-Philosophicus / Logische-Philosophische Ab-handlung*, procurée par B. MCGUINNESS et J. SCHULTE, Suhrkamp, 1989, pp. 135-139

mais il est une limite du monde". On observe une conclusion semblable à la fin de 5.641 : "Le Je philosophique n'est pas l'être humain, ni le corps humain, ou l'âme humaine, dont s'occupe la psychologie, mais le sujet métaphysique, la limite — non une partie du monde".

Il est ainsi manifeste que seule la signification conceptuelle de la limite nous permette de reconnaître ce qu'on doit entendre par *solipsisme*. Bouveresse remarque (MI, pp.152-153) que cette notion d'une démarcation, si la limite pouvait jamais être tracée entre la réalité physique *et* le monde psychologique, comme entre le langage *et* le monde (donc si l'on adoptait la traduction de *Grenze* en un sens obvie : comme une *frontière*), ruinerait l '"identification surprenante du solipsisme avec ce qui semble en être l'antithèse absolue" (à savoir le *réalisme*). Wittgenstein affirme, effectivement, que l'argument, "s'il est conduit strictement", coïncide avec le réalisme *pur* (5.64). L'artifice de cette contiguïté géométrique entre le point et ce qui n'est pas lui (le monde, *dont il ne fait pas partie*), comme celui de la visuabilité de la scène mentale qui manifestement n'est pas déterminée dans un sens optique, vident en effet le sujet d'aucun "point de vue" véritable. C'est pourquoi, ajoute Wittgenstein : "rien dans le champ visuel ne permet d'inférer qu'il est vu par un œil". En quoi donc justifier encore cette *myness* du point de vue, s'il est considéré comme métaphysiquement impersonnel ?

La métaphore spatiale de la subjectivité que Wittgenstein dessine dans le texte (5.6331) — l'œil étant attaché comme une prothèse à l'apex de la boucle que délimite l'espace du champ de vision —, s'il elle ne doit pas être déchiffrée, ni même "vue" sur la page, nous indique-t-il (alors que nous ne voyons précisément que ce dessin), réduit l'argument à quelque chose que nous ne pouvons pas dire. *Das Gesichtsfeld hat nämlich nicht etwa eine solche Form.* Mais il y a une finesse de l'argument à cet endroit, puisque la prémisse est que je ne vois pas mon œil, et que je ne peut donc considérer que le rapport de mon œil au champ visuel soit une *image* de la situation métaphysique du sujet par rapport au monde. Ce qui se montre *montre* ce qui ne se comprend pas comme on le voit. D'où le cercle argumentatif dans l'emploi réfléchi du verbe *zeigen*. Nous avons bien l'impression que l'argument se réfute soi-même, ou bien qu'il n'est pas un argument : la lecture qu'en a faite D. Pears, pour influente qu'elle ait été, est hélas dévastatrice [9]. On peut naturellement soutenir que deux affirmations équivalentes — mais non pas homophones — se rencontrent en ce point : "Le sujet est complètement absent du monde", et

[9]: Cf. *The False Prison*, Clarendon Press, Oxford, 1987, Tome 1, pp. 153-190, trad. fr. *La Pensée-Wittgenstein*, Aubier, 1993.

"le sujet n'est *rien* en dehors du monde". La difficulté serait de décider alors en quoi ce qui se montre ne peut pas non plus être vu dans le texte (ou être lu dans la phrase), ni délimité hors du texte dans le champ visuel. Le fait est qu'on ne sait plus en l'occurrence de quel voir il s'agit, ni ce qu'on nous montre, qui ne se visuabilise pas vraiment sous la forme du dessin que nous voyons.

La structure de l'argument maintenant illustré est-elle plus recevable ? Il y a en réalité deux manières de penser ce qu'énonce à ce propos Wittgenstein dans le *Tractatus* : (i) démontrer que l'argument n'est pas réfutable, parce qu'il n'en est pas un ; (ii) démontrer que la validité de l'argument dépend de l'interprétation générale que l'on donne des autres énoncés du *Tractatus*. Il est très clair que le solipsisme est difficilement compréhensible dans l'acception classique de l'idéalisme subjectif ("il n'existe pas d'autres esprits que le mien", ou "il n'y a pas de réalité extérieure », selon les cas) ; il rentre assez malaisément aussi sous la figure d'un argument kantien qu'a reprise Stenius (le "je pense" accompagne nécessairement toutes mes représentations). On peut le présenter comme un "non-sens philosophique" (ou logique, on l'a déjà dit), et pour d'autres il n'est pas autre chose qu'une fadaise de plus, une extravagance métaphysique non moins répréhensible que tant d'autres. En outre, il est clair que Wittgenstein ne fournit aucun argument concluant *contre* le solipsisme ; il n'en existe guère plus qu'il n'y en a classiquement contre le scepticisme. On ne s'étonnera pas que les commentateurs aient préféré la plupart du temps (depuis R. Rhees jusque David Pears [10]) adopter une posture interprétative franchement anti-solipsiste : il leur suffit de se fonder sur les déclarations de Wittgenstein dans les années 30 pour les projeter dans le traité. Seul Favrholdt lui accorde une place privilégiée, mais son interprétation *phénoméniste*, en réalité très subtile ne dit pas seulement : "il n'y a que mes sensations" ; elle est assez proche du particularisme de Chisholm. On pourrait le comprendre enfin au sens de Reid qui imagine une bulle phénoménale, une sphère semblable à celle des bandes dessinées qui n'enferme plus un monde euclidien. Pourtant il faut noter que l'emploi du terme de limite sert un peu plus tôt (en 5.561), pour statuer sur ce qu'il en est du monde *avant* qu'il soit dit que ce monde soit le *mien* : "La réalité empirique est délimitée (*begrenzt*) par la totalité des objets. Cette limite (*Grenze*) se montre encore dans la totalité

[10]: David PEARS (op. cit, trad. fr. p. 171), récuse que Wittgenstein soit "acculé" au solipsisme, ou "se déclare" solipsiste : "Wittgenstein ne souscrit pas au solipsisme dans le *Tractatus*, car il serait obligé de le traiter comme une théorie susceptible d'être vraie ou fausse. Il considère simplement que le solipsiste a compris une vérité importante qui ne peut pas être énoncée dans le discours factuel, mais seulement être montrée"

des propositions élémentaires". De nouveau le *zeigen sich* revient, et nous interroge de façon lancinante. Nous paraissons forcés, en pareil cas, de prendre "monde" et "réalité empirique" pour deux dénominations distinctes : ce qui serait justement en contradiction avec l'argument solipsiste (c'est même du reste ce que D. Pears propose : le solipsiste est un prête-nom). Il y a donc effectivement quelques difficultés à maintenir une cohérence réelle entre les autres énoncés du *Tractatus* et l'énoncé d'après lequel le solipsisme dirait quelque chose de "distinct" : car en fait il semble que je ne dise rien "du" monde, en disant que le monde est *mon* monde.

Bouveresse en conclut qu'on se trouve apparemment en présence d'une tautologie exprimable en deux formules du type : 1) "le champ visuel est mon champ visuel" (énoncé solipsiste pur) ; et 2) "Le monde existe" (énoncé réaliste pur). Ni l'un ni l'autre n'ont cependant de négation douée de sens et par conséquent ne sont des énoncés nécessaires. Je ne puis pas certain que ce soient deux vraies tautologies. On me pardonnera, j'espère, de ramener l'exposé développé de Bouveresse (une centaine de pages) a quelque chose d'aussi réducteur qu'une citation comme celle-ci, qui commente 5.6331 :

> Rien dans le champ visuel ne permet d'inférer qu'il est vu par un œil physique, non pas parce que je ne vois pas cet œil, mais parce que, même si je le voyais, je verrais simplement mon œil à côté d'autres choses dans le champ visuel, et non pas que ce qui est vu est vu par lui. Même si la corrélation qui existe entre l'œil physique et le champ visuel est purement empirique, il n'empêche qu'elle ne peut, *pour des raisons logiques*, faire l'objet d'une expérience visuelle (il faudrait pour cela en quelque sorte "voir" de l'extérieur le champ visuel, l'œil qui le voit et qu'il le voit). Donc ce qui est important, c'est que je ne puis en aucun cas voir *que* ce qui est vu est vu par quelque chose, que ce quelque chose puisse ou non se trouver aussi éventuellement dans le champ visuel. Rien dans l'évidence visuelle proprement dite n'autorisera jamais à dire que le visible est vu par quelque chose qui est ou n'est pas visible en principe ou en pratique. Et, dans ces conditions, il est tout à fait secondaire pour ce que Wittgenstein veut montrer que ce qui, dans la réalité, permet de voir ne soit pas lui-même visible (MI, p.160)

De telles remarques nous incitent à penser qu'il faut en effet d'abord se demander à quelles conditions l'argument est compréhensible. Bouveresse décompose et sépare *ce qui est visible* et *ce qui est vu* à l'aune de ce qui se montre, et qui par conséquent ne se démontre pas. En d'autres termes, il n'y a

pas d'évidence visuelle ; donc il n'y a pas de proposition énonçable d'après laquelle ce qui est vu aurait besoin d'un sujet oculaire (à savoir l'œil). On ne peut rien avancer à l'encontre. Un œil ne peut en effet rien nous dire pour justifier *que* ce qui est vu est vu parce qu'il serait objectivement visible. *Alles was wird sehen könnte auch anders sein* (T: 5.634). On pourrait toujours voir aussi les choses autrement, confirme Wittgenstein. Mais constater que ce qui est visible pourrait être autrement visible qu'il n'est vu, c'est aussi constater qu'il pourrait être descriptible d'une autre façon et par un autre dispositif propositionnel. Soutenir que l'argument : "Je suis mon monde" serait ainsi "autoréfutable" (dans sa forme) n'en est pas moins parfaitement injustifié. Il n'est pas énonçable comme tel — comme un argument —, parce que nous ne pouvons ni lui donner un sens ; ni (si nous lui conférons un sens) démontrer qu'il est faux que "je sois mon monde" pour des motifs contingents. David Bell a poursuivi cette voie de recherche en partant de la supposition que la critique de la *centralité* du point de vue ou du "moi" psychologique, ne préjugeait pas d'un nouveau statut de la subjectivité qu'on devrait prendre à la lettre, sans du tout conclure pour cela que la critique du "langage privé" s'appliquerait rétrospectivement au *Tractatus* [11].

On peut, selon Bell, déduire de la transitivité de l'identité que de

1/ "*Die Welt ist meine Welt*", et

2/ "*Ich bin meine Welt*", suivrait

3/ "*Ich die Welt bin*".

Et l'on pourrait selon lui, lire de droite à gauche 3/ — de sorte que "je suis le monde" (ce que Wittgenstein n'écrit pas), dirait expressément *que le monde et le Je sont une seule et même chose*. La disparition de l'adjectif possessif devient en l'occurrence ce qui est signifiant ; il indique la place vide de ce qui logiquement ne peut pas être "dit". Comme Bouveresse, Bell montre que si nous admettons un *solipsisme sans sujet* — ce qui est plus ou moins la version orthodoxe — *alors nous n'avons plus de solipsisme* : il ne reste plus que le monde, qui de ce fait n'est justement pas "mon" monde. Tandis que si nous prenons les énoncés du *Tractatus* pour recevables, chacun séparément, et en toute rigueur, ce qu'ils disent est plutôt que si la subjectivité est bannie *de l'intérieur* du monde, elle n'en est pas moins la propriété caractéristique du monde saisi comme un tout. Il n'y a pas de faits "subjectifs" ; le monde empirique est "sans propriétaire", (etc.) ; ce sont bien là des énoncés corrects

[11]: David BELL, "Solipsismus, Subjektivität und öffentliche Welt", in *Von Wittgenstein Lernen*, ed. par Wilhelm Vossenkuhl, Akademie Verlag, Berlin, 1992, pp. 29-52.

— et nonobstant la logique est vis-à-vis du monde dans une situation telle qu'elle ne "peut en parler en aucun sens de *l'extérieur*". C'est aussi pourquoi Bouveresse a maintenu avec raison qu'il y a sans nul doute, chez Wittgenstein, la postulation d' un *sujet métaphysique* — "présupposé" logiquement — tandis que Bell s'oriente quant à lui vers l'idée curieuse d'une assomption du monde public dans la subjectivité. Bouveresse défend clairement, à l'opposé, le caractère a priorique :

> En fait, ce n'est pas simplement parce qu'on ne rencontre nulle part dans le monde un sujet, qu'il ne peut y avoir de sujet au sens où on l'entend habituellement. Le sujet ne *doit* pas pouvoir se rencontrer dans le monde, s'il est réellement le sujet. Car ce qui est une condition de possibilité de toute expérience ne doit pas pouvoir être expérimenté, ce qui est donné dans l'expérience ne peut être en même temps *a priori*. Ce qui est donc déterminant n'est pas que le Je ne soit pas un objet de l'expérience, mais qu'il ne *puisse* pas (au sens logique, non pas empirique, du terme) l'être. L'idée de Wittgenstein est que, si le sujet devait apparaître d'une manière quelconque dans le contenu de l'expérience et intervenir dans la description que nous en donnons, il constituerait nécessairement une caractéristique *contingente* de cette expérience (tout ce que nous pouvons observer et décrire est contingent) ; et cela signifierait qu'il peut aussi bien, le cas échéant, ne pas y apparaître et ne pas faire partie de la description, c'est-à-dire qu'il n'est pas le sujet. Aussi étrange que cela puisse paraître, c'est bel et bien pour préserver le caractère *a priori* et nécessaire du Je comme centre du monde que Wittgenstein est amené à lui dénier toute épaisseur et toute consistance "mondaine" (MI, p.128).

Les conditions de recevabilité de l'argument solipsiste sont suspendues à l'interprétation, qui reste très difficile, pour ne pas dire chantournée, de *ce qui se montre* — et c'est la part d'ombre du solipsisme — tandis que la seule chose qui s'exhibe dans un tel argument, c'est qu'il n'est pas réfutable formellement comme un argument. De la proposition *le monde existe*, je ne peux pas inférer que *j'existe*. De même, si le "Je" existe dans la langue, il n'existe pas de fait en-dehors de sa situation dans la langue. Nous comprenons que dans la mesure où le sujet n'est pas un nom d'objet, il n'est pas le sujet d'une expérience empirique ; il ne l'est pas, parce que (selon Bouveresse) *a priori* il ne pourrait pas l'être au sens expérimental. Il y aurait une sorte d'inconsistance du *sujet*, sans que

nous comprenions pourquoi celui-ci *reste alors placé au centre du monde* comme une garantie *a priori* du monde "tel que nous le trouvons".

Bell invoque, à l'appui de sa thèse, un double principe de *littéralité,* et de *bienveillance* dans l'interprétation : l'un veut que "tout texte qui contient l'affirmation que *p*, comporte toutes choses égales d'ailleurs, une justification que l'auteur souscrit à la conviction que *p* " ; l'autre, que nous négligions "chaque fois que possible, de présumer que les théories et les thèses d'un auteur sont fausses, manifestement incohérentes ou insensées". D'après lui, le non-respect de ces deux principes a laissé accréditer l'idée que le solipsisme était une curiosité, et qu'à *proprement parler* Wittgenstein ne pouvait pas avoir été solipsiste, ce que *Le Cahier bleu* et d'autres textes de la période intermédiaire manifesteraient avec éclat. Les nombreuses déclarations en faveur du solipsisme, contenues dans les *Carnets* justement, ne seraient plus que des approximations équivoques. Contre cette version des choses, D. Bell demande qu'on lui accorde trois attendus : 1) que pour être acceptable, le solipsisme n'implique aucun énoncé empirique faux ; 2) que le solipsisme soit libre de toute contradiction ; 3) et enfin, que le solipsisme soit philosophiquement intéressant.

En ce qui concerne le premier d'entre eux, il revient sur un élément piquant et anecdotique, qu'il présente comme une forme de confirmation "factuelle" du solipsisme que Bouveresse avait, nous l'avons vu, proscrit

sous toute acception empirique que ce soit. Cette vérification très particulière relève de l'idée que si j'argumente devant autrui, il doit exister quelqu'un pour qui ce que prononce le solipsiste serait juste, or si cette personne existe, le solipsisme est faux. Russell avait argumenté en ce sens dans sa *Theory of Knowledge* : il ramenait l'énoncé à une croyance du type : "je suis le seul qui existe", ou "il n'y a que mon esprit qui existe". Russell concluait que le solipsiste ne pouvait pas croire raisonnablement lui-même que cette conviction fût irréfutable. Miss Anscombe y faisait allusion dans son *Introduction* :

> On raconte que Mrs Ladd Franklin aurait écrit à Bertrand Russell en lui disant qu'elle était solipsiste et qu'elle ne pouvait pas comprendre pourquoi personne d'autre ne l'était pas aussi ! Il est possible que l'effet comique ait été intentionnel, et qu'il se soit agi d'une plaisanterie de sa part. La nécessité du solipsisme est éminemment discutable ; pourquoi le solipsiste n'en discuterait pas avec quiconque serait capable de discuter ? Même si deux solipsistes échangeaient leurs vues en se congratulant mutuellement, il ne sortirait rien de leur discussion : aucun des deux ne

> cèderait sur l'autre, quant à la position unique qu'il conçoit pour lui-même. Si deux personnes discutent du *cogito* de Descartes, ils peuvent se mettre d'accord : "il s'agit d'un argument que je peux m'administrer à moi-même", et chacun peut soutenir que l'autre est dans l'erreur s'il le met en discussion (...)
>
> D'autre part, il est très difficile de penser à un moyen de sortir du solipsisme. En effet, on tient souvent le solipsisme pour irréfutable, mais aussi pour trop absurde chez qui l'adopte personnellement. Dans la version de Wittgenstein, il est clair que le "je" ne sert pas à référer à quelque chose comme le corps ou l'âme, car sous ce rapport tous les hommes sont semblables. Le "Je" réfère au centre de la vie ou au point à partir duquel toutes choses sont vues[12].

Cet argument de l'interlocuteur est contesté par David Bell, qui ne croit pas que le solipsisme soit une profession de foi en elle-même extravagante ; il n'impute pas à Christine Ladd Franklin la "faute" majeure, dénoncée par David Pears, qui serait de prétendre que le solipsisme s'incorpore au *discours factuel*. Il n'y aurait, selon lui, qu'une discordance logique entre ce que le solipsiste nous "dit" et ce qu'il "fait". Cependant, s'il faut donner une portée à l'argument de Wittgenstein, nous devons précisément supposer, d'après Bell, que "tous les aspects du comportement le plus ordinaire sont compatibles avec le solipsisme". Tel est bien ce que Christine Ladd Franklin objectait d'ailleurs par avance à Russell dans sa lettre du 21 août 1912 :

> Le solipsisme n'est rien d'autre qu'une description du caractère incontestable des faits de l'expérience ... Je suis moi-même (pour autant que je le voie) la seule solipsiste, c'est-à-dire en effet que je représente un réalisme hypothétique. Ne voyez-vous pas que c'est bien là l'unique position logique [13] ?

S'il y a une possibilité d'accepter le solipsisme autrement qu'en le rapportant à l'existence des "autres esprits", ce ne peut l'être que de cette façon, sans y voir une pathologie philosophique ou un argument vide et sans portée. Bouveresse serait sans doute beaucoup plus d'accord avec David Pears sur ceci que

[12]: G.E.M. ANSCOMBE, *An Introduction to Wittgenstein's Tractatus*, Hutchinson University Library, Londres, 1959, p. 168

[13]: cité in David Bell, *op. cit.* , p. 37

les formes d' "individuation" de mon monde ne permettent nullement de rendre l'énoncé "vrai" (quelle que soit la façon dont on le tourne) ; mais il serait peut-être en désaccord avec lui sur l'idée que c'est la conviction qu'on peut acquérir une *connaissance directe de l'ego* qui serait renversée par l'argument. Wittgenstein ne pense pas que le "Je" soit un nom logique, et n'en fait pas l'hypothèse. L'origine du raisonnement qui pousse Wittgenstein à nier que le moi ne puisse être "une partie du monde", ne devrait pas être imputable à la simple réfutation de Russell.

Reste une dernière version compréhensive, qui est due à Brian Mc Guinness, lequel nous demande de lui conférer un contenu autobiographique : c'est-à-dire de penser que l'énoncé du solipsisme reflète "le désarroi d'un officier autrichien dans une guerre perdue d'avance". On ne peut pas l'écarter en principe : le double héritage de Mach et de Schopenhauer ayant fait du "sujet voulant" l'exacte *contrepartie* de la négation du sujet de la pensée et de la représentation. Les *Carnets secrets* contiennent une brève remarque qui semble indiquer que le "point de vue solipsiste" a primitivement été, non pas comme le pense David Pears une conséquence logique de la dépersonnalisation de l'ego, mais une sorte de *réaction conservatoire* de l'intellect jeté dans une situation de précarité extrême, s'affrontant au monde extérieur dans une sorte de réclusion spirituelle. Le 8 décembre 1914, Wittgenstein blessé à la jambe, et lisant le tome VIII des *Œuvres* de Nietzsche, consigne ceci :

> Je suis fortement touché par son hostilité contre le Christianisme. Dans ses écrits, sans doute, une part de vérité est contenue. Certes, le Christianisme est la seule voie *sûre* menant au bonheur. Mais, que se passerait-il, si l'on se détournait de ce bonheur ? ! Ne vaudrait-il pas mieux, malheureux, faire naufrage dans cette lutte sans espoir contre le monde extérieur ? Une telle vie, pourtant, est dénuée de sens. Mais pourquoi ne pas mener une vie dénuée de sens ? Est-ce indigne ? Comment est-elle conciliable avec le point de vue solipsiste fort ? Que dois-je faire, pour que ma vie ne glisse pas hors de moi ? Je dois toujours être conscient de ce qu'elle est mienne – de mon esprit toujours - - - [14]

Ce passage très troublant n'a peut-être pas grande valeur sous l'angle d'une meilleure intelligibilité, mais il corrige sur le principe l'erreur qui veut que l'éthique de Wittgenstein soit incompatible avec l'argument solipsiste. On

[14]: Ludwig WITTGENSTEIN, *Geheime Tagebücher, 1914-1916*, édités par Wilhelm Baum, Turia & Kant, Wien, 1991, p.47

peut supposer, comme le fait David Pears, que le solipsisme implique un détachement du corps vivant : le "flottement" de l'ego, qu'évoque *The False Prison*. Ici, toutefois, la dissolution d'une vie "dénuée de sens" est concomitante d'un combat contre le monde extérieur. Ce combat justifie pleinement le "point de vue" solipsiste de ne pas laisser la vie s'échapper de son *focus* principal qu'est la conscience d'être en vie ; il permet de comprendre la phrase où Wittgenstein dit "le monde et la vie sont un" (5.621). Les *Carnets* de l'année 1916 ne sont qu'une transposition diachronique de cet état existentiel. Il reste par conséquent une grande différence entre le fait que l'énoncé solipsiste soit manifestement "dénué de sens", et la perception du cours d'une existence qui serait dénuée de sens si le "point de vue" solipsiste n'était qu'un point de vue absurde[15].

2. *La vérité du solipsisme peut-elle être montrée ?*

Il y a maintenant vingt ans (en 1993) un quotidien viennois, *Die Presse*, publiait sur six colonnes les confessions d'un chauffeur de taxi, étudiant en philosophie au chômage, qui sous l'effet d'une lecture du *Monde comme volonté et comme représentation*, décrivait assez exactement, mais à partir de l'habitacle de son véhicule, la "bulle phénoménale" que David Pears présente telle la version courante que le sens commun se fait du solipsisme. Wittgenstein contesterait évidemment cette forme d' *inhabitation* du sujet psychologique, pour reprendre un terme scolastique remanié par Brentano, qui se ferait dans son environnement visuel immédiat (*mon champ visuel*). On sait qu'il la traduit longtemps après par l'image de la mouche, prisonnière de sa cloche de verre, dans les *Remarques sur l'expérience privée et les sense-data* ; mais ce n'est là encore qu'une ruse pour stigmatiser la mauvaise interprétation que se donne à lui-même le sujet de l'introspection, condamnée par Auguste Comte et par Brentano, bien qu'elle continue d'avoir ici ou là de chauds partisans.

Telle n'est sûrement pas la "vérité" du solipsisme — du moins celle qu'il intéresse le "sujet métaphysique" : nous avons vu que par elle il n'est que *présupposé,* occupant la limite grâce à laquelle est assurée l'intelligibilité du monde pour la logique. Afin d'en saisir la portée, il faut revenir à nouveau sur la position qui consiste à contrer la tentation "phénoméniste" par une élucidation progressive de l'énoncé le plus énigmatique du *Tractatus :* "ce que

[15] : Christian Paul BERGER a tenté de montrer dans *Erstaunte Vorwegnahmen*, Bölhau Verlag, Wien, 1992, pp. 207-233, que le solipsisme de Wittgenstein était peut-être compréhensible du seul point de vue "littéraire", par exemple celui de Trakl, et il n'y a aucune raison de le mettre en doute.

nous ne pouvons pas penser, nous ne pouvons pas le dire ; nous ne pouvons donc pas non plus *dire* ce que nous ne pouvons pas penser" (5.61) ...

> Cette remarque nous donne la clef nous permettant de trancher la question de savoir dans quelle mesure le solipsisme est une vérité.
>
> Ce que le solipsisme effectivement *veut dire* (*meint*) est tout à fait correct, seulement cela ne se laisse pas *dire*, mais se montre.
>
> Que le monde soit *mon* monde, cela se montre dans le fait que les limites du langage (le seul langage que je comprenne) signifient les limites de *mon* monde (T: 5.62).

"Ce qui se montre" de la vérité du solipsisme, et qui nous paraît être une énormité : *die Welt* meine *Welt ist*, constitue toujours la difficulté irréductible qu'il importe d'appréhender — mais nous percevons d'emblée qu'elle ne saurait être comprise (ni dissoute) par la seule paraphrase ordonnée qu'on peut en fournir. On ne va pas récrire *Thomas Graindorge* d'Hyppolite Taine dans le langage de Maurice Blanchot. Il est plutôt impératif de comprendre ce que l'argument du solipsisme *ne montre pas*. L'un concerne le statut de la *proposition-pensée* ; l'autre questionne la *forme reproductive* qui sert à définir le statut de la proposition-image. Dans les deux cas, "ce qui se montre" paraitrait échapper complètement à la délimitation métaphysique du "Je". Il existe deux versions canoniques pour envisager un tel problème : le solipsisme *linguistique*, d'une part ; le solipsisme *épistémologique*, de l'autre, quoique nous donnions au second des deux un sens plus élargi que de coutume.

Le sens du *solipsisme linguistique* que Stenius avait défendu, consiste à dire que le monde n'est rien d'autre que celui de "l'utilisateur "du langage, et qu'il est tout ce que le langage permet de penser de façon douée de sens : Stenius y voit une adaptation raffinée de la réfutation kantienne de l'idéalisme. Le "*Je pense*" transcendantal devient l'autre nom du solipsisme en ceci que la vérité logique dépendrait d'une sorte de contrainte structurale qui dépasse le sujet de l'intérieur et surplombe la réalité mouvante des phénomènes. C'est une option encore très forte, que l'on retrouve aussi chez Strawson dans *The Bounds of Sense*. Mais le défaut de cette lecture c'est qu'il n'y a pas de fondement nouménal de la signification que nous puissions alléguer ; dans cette option disparaît l'idée que la *réalité empirique* puisse exister indépendamment du langage et de la pensée — or Wittgenstein est certainement convaincu que, si nous ne pouvons pas sortir du langage, nous ne pouvons pas inférer de là que le langage ne se rapporte pas à une réalité dont il nous dit qu'elle se "comporte de telle ou telle manière" (*es verhält sich so und so*) ; c'est-à-dire

qu'elle *se projette* dans notre langage, en fonction de ce que ce dernier autorise et bien qu'elle pourrait par ailleurs se comporter tout autrement, si la structure de notre langage l'y autorisait [16]. — En disant que la réalité *se projette dans notre langage*, qu'est-ce que nous entendons vouloir dire ? — N'est-ce pas purement métaphorique ? Et n'est-ce pas l'inverse qui se produit comme l'a défendu Maria Cerezo ?

> Pourquoi devrions-nous soutenir que le sujet métaphysique est celui qui opère la projection ? Parce que la connexion due à la dépiction, la corrélation de deux faits par le moyen de la corrélation de leurs objets, n'est pas un phénomène qui a lieu *à l'intérieur du monde*. Le monde est tout ce qui est le cas, l'existence et la non-existence des états de choses. Tout ce qu'il y a dans le monde ce sont des faits, c'est-à-dire des connexions entre objets qui existent. Mais *il n'y a pas de connexions entre les faits du monde*. La connexion de deux faits n'est pas elle-même un fait. Pour assurer qu'il y en ait une, il est nécessaire de faire appel à quelque chose qui se tient *à l'extérieur du monde*, et le seul élément qui ait cette caractéristique nous dit le *Tractatus* est le sujet métaphysique. En étant le sujet qui assure la projection, le sujet métaphysique brise l'apparente symétrie de la relation dépictive et constitue proprement le langage : grâce à lui, des faits deviennent des faits linguistiques, des faits par où se déclare *ce qui est le cas*.
>
> D'un côté, Wittgenstein établit qu'il n'y a pas de sujet dans le monde. Le sujet pensant *à l'intérieur du monde* (l'esprit, l'âme au sens psychologique) n'existe pas. D'un autre côté, il considère le sujet tel un point sans extension, comme une limite du monde, mais *avec lequel la réalité reste coordonnée* [17].

Cette lecture du solipsisme linguistique, comme source de la coordination avec le monde, ne s'oppose pas frontalement à l'autre version développée dans l'interprétation scandinave. La thèse du *solipsisme épistémologique*, ou gnoséologique, défendue par Favrholdt stipule que nous devrions partir de l'énoncé déjà cité 5.631 : *Das denkende, vorstellende, Subjekt gibt es nicht*, pour en

[16]: Wittgenstein emploie "réalité empirique" autrement, semble-t-il, qu'il ne fait pour *Wirklichkeit*. Sur ce point, voir l' étude de Markus AENISHÄNSLIN, *Le Tractatus de Wittgenstein et l'Ethique de Spinoza, études de comparaison structurale*, Birhäuser, Bâle, 1993,

[17] : Maria CEREZO, *The Possibility of Language, Internal Tensions in Wittgenstein's Tractatus*, CSLI, Stanford, 2005, pp. 130-131.

inférer que le monde est un ensemble d'objets — un "tout" d'objets — unifié de l'extérieur du sujet, mais sur le compte de l'*ego empirique*, comme le rappelle avec insistance le commentateur danois. (Soit dit en passant, même si ce dernier interprète semble parfois plus disert et volubile qu'il ne faudrait, il a parfaitement compris les dessous de l'affaire : je ne peux me regarder de dos ; je vois de l'encre sur le papier, j'entends des sons, et le problème judicieusement reconstruit par Favrholdt, est qu'il y a une différence entre l'ego empirique et l'espace logique où ma pensée est inscrite). La preuve en est que le champ visuel n'est pas strictement délimité : il n'est pas un contexte analogique pertinent pour que mes pensées et mes perceptions se disent l'une par l'autre. Comment procéder alors ? Si l'on soutient, comme Maria Cerezo, que le sujet est *à l'extérieur du monde*, on contrevient à la thèse selon laquelle le solipsisme confine avec le réalisme *pur* (5.64). Nous sommes pris dans le cercle de l'argument et pourtant cet argument n'est pas contradictoire. Il n'y a pas de contradiction, parce que les limites de mon langage ne sont pas *les* limites *du* monde : elles ne font que les indiquer.

Si l'on veut sortir du dilemme déclaratif, nous devons comprendre que le "Je", qui n'est plus qu'un foncteur logique de l'énonciation, n'a pas d'autre raison d'exister. Il ne réfère pas à un objet, à un fait ou à un complexe de pensées auto-suffisant. Dans les deux versions du solipsisme que j'ai indiquées, la théorie qu'on dit être dépictive (et qu'on croyait indépendante d'eux) les rapproche nécessairement : cette théorie — si mal désignée de la sorte —, est finalement reçue comme une théorie "objectiviste", *se rapportant immédiatement à l'expérience.* Le réalisme interne du langage, en quoi il est incontestablement *mon* langage, se fonde sur une relation extrinsèque aux objets du sens et par le biais d'un ensemble de propositions qui les décrivent. L'assomption d'un sujet impersonnel n'en reste pas moins doublement équivoque : ce n'est pas un sujet logique et ce n'est guère plus une personne ; mais il reste le locuteur, et les quatre mots : "Je suis mon monde" sont pour lui des mots réels. Il est temps maintenant de dissiper cette équivoque entre l'illusion de voir dans le miroir un sujet métaphysique (le *Wahnsinn*) et l'*Unsinn*, qui consiste à sortir des limites du langage explicite. Tantôt chez Favrholdt, on y parvient par la critique de l'ontologie des entités objectives ; tantôt chez Hintikka, en assimilant l'ensemble des propositions possibles à des fonctions d'objets.

Il est possible d'admettre évidemment que Wittgenstein "ait encouragé le malentendu" dont serait victime Russell qui a cru le but que s'assigne le *Tractatus* est de "s'occuper des conditions qui rendent possibles un langage logiquement parfait". Dans la *Préface* du *Tractatus*, Wittgenstein déclare explicitement avoir voulu "tracer une limite à l'expression de la pensée, ou plutôt — non

pas à l'exercice de la pensée — mais à l'expression des pensées : car pour tracer une limite à l'exercice de la pensée, il faudrait pouvoir penser des deux côtés de cette limite (il nous faudrait donc pouvoir penser ce qui ne peut être pensé)". On est en pareil cas obligatoirement ramené à la question du solipsisme. A-t-il un "contenu de vérité", comme on le dit dans l'herméneutique philosophique, sous-entendu un *Gemeinte* ? Oui, si nous donnons une lecture compréhensible de 5.61 — source du malentendu qu'il s'agissait de dissiper. "Nous ne pouvons pas non plus dire ce que nous ne pouvons pas penser", affirme Wittgenstein. Il ne s'agit donc nullement de gloser *ad nauseam* sur ce qu'on pourrait dire qui ne peut être dit, et d'assimiler le solipsisme à la promotion de l'ineffable, comme l'a martelé James Conant avec une mauvaise foi patente. Ce serait croire que ce qu'on ne peut pas dire ne serait pas explicitement verbalisé dans ce que nous disons par une sorte de rature ou de réservation constante[18]. La délimitation concerne l'expression linguistique de la pensée. L'exclusion concerne ce qui est explicitement contraire aux faits. Sous ce rapport, comme on l'a vu, le solipsisme demeure parfaitement pensable de l'intérieur du sujet de l'énonciation, non pas dans son for intérieur, mais au regard de ce qui possiblement concevable, ou encore au sens de la *phénoménologie cognitive*. Sous ce dernier label, nous ne sommes plus dans la relation de référence des noms d'objet, desquels nous ne pouvons justement rien dire, comme le rappelle Wittgenstein. Deux lectures au final doivent être écartées pour ne pas prolonger le cauchemar.

(a) Il ne s'agit pas de "faire voir" dans une phrase que l'on *ne* saurait jamais *exprimer en elle* trivialement, ni littéralement, *ce que l'on dit*. Cette version est déjà une précaution philistine erronée — pseudo-litote ou ellipse — qu'on attribue à Wittgenstein : on pense de suite ici au *nihilisme logique* que lui impute Meyerson, son premier lecteur en France. Contre cette thèse, il est possible d'affirmer que le texte du *Tractatus* ne cesse de traiter de nos pensées comme autant de propositions que nous faisons au sujet de la réalité.

(b) Mais il ne s'agit pas non plus, à l'inverse, de proférer une opinion quelconque, pour ajouter de suite après *que* l'on dit bien ce que l'on dit, si et seulement si ce que l'on dit avoir dit est formellement recevable (c'est la tentation d'une lecture diplomatique). Contre

[18] : Voir entre autres textes, "The Method of the Tractatus", in *From Frege to Wittgenstein*, ed. by Erich H. RECK, Oxford UP, 2002. Sur la question de l'expression, p. 376, sur les limites du langage et la conviction de Conant qu'elles ne peuvent former qu'un *einfach Unsinn*, p. 424

> cette thèse, il suffit de rétorquer que la pensée est intégralement manifeste dans chacun des énoncés que nous formons. Les items du *Tractatus* ne sont pas les propositions possibles dont parle le *Tractatus*.

La forme d'expression surdéterminée négativement en (a), et la soustraction de l'assertion par une clause conditionnelle (en b), ne paraissent en rien commensurables ni comparables. Ce ne sont que deux formulations sophistiques différentes. En somme, il ne s'agit nullement de "délimiter" ce supplément rhétorique que nous devrions exclure pour montrer que nous ne le disons pas. On saisit dans cette hypothèse de quelle *vérité* du solipsisme il est parlé. Pour beaucoup, il est évident que l'*agnosticisme* prononcé à l'endroit des objets représente une contribution parallèle à l'abandon d'une visée transcendante au langage. Cette contribution élusive est aussi importante que le désenchantement du mysticisme syntaxique dont il fournit la démonstration. En résumé, Wittgenstein a besoin du sujet métaphysique de la même manière que la stipulation des objets du monde lui est indispensable. La discussion qu'ont poursuivie J. W.Cook et P. Carruthers sur la portée de la "doctrine" métaphysique du *Tractatus*, indique que la perspective exégétique ne pouvait pas s'émanciper, dans les années 1990, d'un couplage de cette sorte qui n'est pas vraiment anti-dualiste, mais qui reconduit à l'aporie telle que nous l'avons précédemment formulée [19].

L'intérêt de revenir sur cet argument, en dépit de son "incohérence", est qu'il enveloppe la critique d'une conception intra-psychologique du *Gedanke* (la pensée), telle que Frege l'a stigmatisée. Toute la question de fond étant de savoir si le *réalisme interne du langage* est compatible avec le solipsisme déclaré qui semble relever d'une justification *externe* dès le principe de son énoncé. Mes pensées ne sont-elles pas des faits du monde à part entière ? L'une des réponses probables à cette dernière question est celle de Frascolla (2000). Car dans cette question se devine aussi le genre de profession de foi qu'un philosophe de profession *ne peut pas faire.* Frascolla explique en effet correctement à notre avis la situation de l'assertion solipsiste au sein de l'économie énonciative du *Tractatus*.

> Dans la discussion sur les limites spéculaires du langage et du monde, la référence à un sujet est requise proprement du fait que

[19]: John W. COOK, *Wittgenstein's Metaphysics*, Cambridge University Press, 1994, pp. 55-68 (retour à l'interprétation phénoméniste) ; et Peter CARRUTHERS *The Metaphysics of the Tractatus*, Cam-bridge University Press, 1990, qui défend l'ineffable "myness" du point de vue contre cette inter-prétation, pp. 79-83.

> de telles limites renvoient, en dernière instance, aux objets, soit aux signifiés des noms : c'est seulement par l'œuvre du locuteur que de tels signifiés peuvent être établis. (...) Mais comment interpréter le mot "solipsisme" ? Très génériquement, le solipsiste soutient que tout ce qui existe n'existe qu'en rapport avec son expérience : pour tout sujet, le monde est *son* monde, et coïncide avec ce que *lui* fait l'expérience directe qu'il est. Mais l'un des traits distinctifs de la position de Wittgenstein consiste à adopter la perspective solipsiste non seulement au regard des faits, mais au regard des objets (non simplement par référence au monde, mais par référence à la substance du monde). Parce que les objets ne sont rien d'autre que les signifiés des noms, la conception que Wittgenstein retient comme acceptable peut être définie comme un *solipsisme sémantique*. Cette thèse émerge du lien qu'il noue entre le solipsisme et celui des limites du langage (et du monde), qui renvoie justement aux objets. Ce lien entraine l'idée que la thèse "le monde est mon monde", formulée par le solipsiste n'est compréhensible qu'en un sens plein : autrement dit non seulement le monde qui est de fait est mon monde, *mais tout monde concevable est mon monde*. Ce sont les objets, en tant que signifiés des noms, qui sont convoqués dans le champ de ce que le locuteur expérimente directement : et cela même, avec leur substantialité, garantit que les constituants de tout monde possible tombent dans ce domaine.
>
> (...) A la conclusion, à la rigueur inexprimable, que le monde soit mon monde, nous serions donc conduits par la constatation que les limites du monde réfractés dans celles du langage, duquel et pour lesquelles seul je suis compétent à le savoir, ne seraient pas les limites d'un monde sans propriétaire, mais proprement celles du monde mien, d'un monde dont les constituants ultimes — les objets — se retrouvent tous à l'intérieur du cercle de mon expérience directe.[20]

La justification problématique de l'assertion solipsiste comporte donc ceci d'étrange qu'elle n'est pas susceptible de provoquer la recherche d'un contre-exemple. Puisque Wittgenstein revendique la relativité de l'espace logique en regard du sujet, et puisque le solipsisme sémantique semble en découler, il suit ici que

[20] : Pasquale FRASCOLLA, *Tractatus Logico-Philosophicus, Introduzione alla lettura*,Carocci editore, Roma, 2000, p. 272-273

le locuteur ne peut pas être une entité empirique comme les autres — ainsi que le défendait Favrholdt. Nous savions que l'esprit avait été dissous dans son unité par une authentique réduction à l'absurde : le voici maintenant confronté au postulat ontologique de la contingence des faits. D'où cet énoncé proche de Musil que confesse 5.631 :

> Si j'écrivais un livre "le monde tel que je l'ai trouvé", il faudrait aussi y parler de mon corps et dire quels sont les membres soumis à ma volonté, quels autres ne le sont pas, etc ; c'est là en effet une méthode qui consiste à isoler le sujet ou plutôt à montrer qu'en un sens important, il n'y a pas de sujet : c'est la seule chose dont il *ne* saurait être question dans ce livre.

La vérité du solipsisme n'est donc pas la vérité romanesque du sujet frappé d'une sorte de bovarysme névrotique. Et néanmoins l'isolation du sujet — que réfute Wittgenstein tout en l'inscrivant dans le donné sensoriel — exige une justification métaphysique supérieure, condition indispensable si elle ne veut pas sombrer dans l'idéalisme. Le naufrage du *Tractatus*, en ce sens même, comme le soutient Simons à la suite de Moore, c'est le déni du professionnalisme académique. Ce qui distingue l'attitude professionnelle en philosophie est, à l'évidence, extrêmement difficile à déterminer. Wittgenstein n'a cessé de s'en écarter ; il concédait fort difficilement que le modèle de Russell, qui fut son mentor, aurait pu s'imposer à lui. Mais, si l'on *ne peut pas* professer le "point de vue" solipsiste, selon Russell, c'est bien d'abord parce que l'acquaintance avec mes *sense-data* rend cette théorie impénétrable. Colin McGinn et Thomas Nagel sont pourtant revenus aujourd'hui sur cette position, expliquant qu'elle est peut-être l'un de ces tournants que le sujet est forcé de prendre dans des conditions semblables, notamment quand survient l'écueil du brouillage des croyances ou celui du "floutage" des objets de notre environnement. La même suspicion à l'égard du solipsisme *réaliste* affecte Wittgenstein, on l'a vu, quand il traite des objets substantifs de la réalité du monde, puisqu'ils jouent pour lui le même rôle que jouent les *sense-data* russelliens (sans du tout se confondre avec eux). Tandis que T. Nagel étend l'objectivité à toute la sphère du mental sur la base même de l'argument de Wittgenstein, Mc Ginn avoue de son côté que pratiquement, "nous devrions tous être solipsistes".[21] — Une affirmation qui semble quand même assez déraisonnable et frise le non-sens, si nous devions la prendre à la lettre. Elle

[21]: Colin McGinn, *The Problem of Consciousness*, Blackwell, 1991, pp. 70-71 ; Thomas Nagel, *Le Point de vue de nulle part*, trad. S. Kronlund, L'éclat, 1993, pp. 42-48 et 126-132.

ne surprendra pas le béotien *dogmatique ;* ce dernier ne met pas en doute ce qui lui apparaît comme un corps de vérités dénuées de pertinence informative. Mais ce sont en fait d'*autres* vérités qu'il n'a justement pas besoin de justifier qui lui permettent d'éviter toute emphase sur le mode de l'argument solipsiste.