FORMAL METHODS IN PHILOSOPHY
SHOOTING RIGHT WITHOUT COLLATERAL DAMAGE

Pascal Engel
University of Geneva
Pascal.engel@unige.ch

Summary :

The problem with formal methods in philosophy is not whether to apply them or not. For most analytic philosophers they are basic and unquestionably legitimate and fruitful. The problem is their adequacy. Sometimes there is too much use of formal methods in a field, and their richness and the diversity of the formalisms overkills the part of the field and the questions that they are supposed to illuminate. At other moments, they miss the target and fail to capture what is essential. So the important question is when to use them so that they can shoot correctly at their target without too much collateral damage.

Introduction : no drama

Once upon a time, the conflict between those among analytic philosophers who were friends of formal methods and those who were hostile to them raged. But the days of logical positivism have gone, as well as those of the opposition between Carnapian methods and ordinary language philosophy. Today the atmosphere is more peaceful. Almost nobody, I gather, believes that all philosophical problems can lend themselves to a formal treatment, and that formalisms are all encompassing. But there is still an opposition between those who think that one can go a long way with them, and those who are sceptical (very often for quite oppositie reasons: for instance Wittgensteinians dislike them because they believe that there is a conceptual core in philosophy which cannot be represented by any formalism, and "experimental philosophers" dislike them because they take them to be typical of the armchair method).

The question is: how far can we go, and what are the limits of scepticism about the use of formal methods? The problem is not to throw away formalism. The formalisms of logic and probability theory are part of the language and the natural equipment of today's philosophers: quantified modal logic, probability theory are the common language (they are almost Kuhnian paradigms). The problem is to assess their scope. I would like here to suggest that one of the main difficulties of the formalist method lies not with the use of formalism *per se*, but with the frequent lack of preliminary elucidation of the philosophical layout on which it is supposed to be applied.

1. Virtues and vices of formalisation

Formalism in philosophy has its virtues and its vices. According to its friends, formalism helps us to make philosophical problems and their possible answers precise: they improve its formulation, make more visible the possible assumptions, the possible solutions and their

consequences, and help see clearly what the options are. In this sense they have a preparatory and heuristic role.They have also a negative and refutational role: if a thesis is not formalisable, it is very likely to be obscure and false, and we are led to remove ambiguities. But formalisms have a positive role as well: sometimes they can be used as heuristics. Formalism has also a number of vices. It has been objected that it misses the force of some philosophical problems by reducing them to their expression into a formalism, that it mistakes properties of the models for properties of the things to be modelled, that sometimes it is like using a sledgehammer to crack a walnut: as La Fontaine says in the fable *The bear and the amateur gardener*: "A foolish friend may cause more woe than could indeed the wisest foe".

In order to assess these criticisms, we need first to understand how formalisation works. Formalisation is not a mere regimentation of natural language into a symbolic language, such as predicate calculus, probabilistic models, or modal logic. It involves a preliminary regimentation of philosophical language into formal language, which has primitives, axioms, and metatheoretical properties, such as soundness and compleness – if possible. It involves a choice of formalisms and the possibility to compare them. It then involves the use of the properties of the formal language to clarify the assumptions of a fragment. Among the most familiar formalisms are quantified modal logic, model theory, probability theory, decision theory, utility theory and their many subranches. As a number of proponents of the formalist approach have remarked[1] , these formalisms can be considered as models, in the same sense as the one with which we are familiar in the philosophy of science. Among their virtues are, in the first place, their heuristic importance: they suggest something that we might later be able to explain in a model-independent way. In the second place, models help us to explain features of science, ie to deal with more realistic (and, as it happens, more complicated) situations. As Hartmann says, "when different intuitions pull in different directions the philosophical model tells us which intuition wins" in which part of the parameter space. The virtues of models are thus described by Sven Ove Hansson:

"Philosophy can never be reduced to mathematics. But we can often produce mathematical models of fragments of philosophy and, when we can, we should. No doubt the models usually involve wild idealizations. It is still progress if we can agree what consequences an idea has in one very simple case. Many ideas in philosophy do not withstand even that very elementary scrutiny, because the attempt to construct a non trivial model reveals a hidden structural incoherence in the idea itself. By the same token, an idea that does not collapse in a toy model has at least something going for it. Once we have an unrealistic model, we can start worrying how to construct less unrealistic models."[2]

At the same time, Hansson is very lucid about the possible drawbacks of this method. They are the risks of oversimplification (reduction of primitive notions to a minimum), of false unification of concepts (for instance in deontic logic putting together all predicates of prescription (*should*, *must*, *ought*) under the same operator O may be a good idea in some contexts, not in others) false conceptual primitivity (selecting one concept because of its simplicity and elegance), invitation to ad hoc constructions, which are mere artefacts of the model, formalisation involves implicit ontological assumptions that are not innocuous, enigmatic style where important

[1] I rely here on the very useful treatments of these issues by Douven / Horsten 2008
Hansson 2000, Hartmann 2008

[2] See also the preface of his book Hansson 2001.

philosophical choices are made without explanation. Igor Douven and Leon Horsten (2008) are also critical when they say:

"If formal methods function, in some sense, as paradigms in the philosophy of science, then it should not come as a surprise that for every formal method there comes a point of diminishing returns. When a formal method has been applied in one area of the philosophy of science, it is very natural to try to apply the same technique to other branches of the philosophy of science. But at some point the new applications begin to look forced and somehow unnatural: the formal method does not succeed in shedding (new) light on the conceptual problems at hand."

I quite agree, and just want here to confirm these diagnoses, without ignoring the fact that they come from researchers who illustrate the virtues of the method of formalisation.

3 Good shots and bad shots of formalisation

Everyone agrees that often formalisation hits its target, and that it is succesful. Thus philosophy has benefited from the following examples of successful formalisation:

- Anselm's ontological proof (Gödel, Plantinga, Lewis, Oppy)
- Truth in L by Tarski
- Lewis's trivialisation result for conditionals
- Lewis on belief as desire
- Skyrms' formalisation of the social contract
- Gardenförs impossibility result about the Ramsey test
- Bayesianism and Hume's argument on miracles (Salmon, Earman)

Why are such results successful? Because, it seems to me, they all combine a careful formulation of a philosophical problem in the first place, then its formulation in formal terms, and then a result which shows that one thesis is wrong or flawed or that a certain approach is fruitful. Of course, none of these formalisation solves the problem at issue, but they all give us clearer insight into the nature of the problem at hand, and help us in testing the coherence of a given thesis. Take for instance Lewis (1998) treatment of desire-as-belief. Lewis considers the view in moral psychology, defended by cognitivists against Humeans, that some normative beliefs – and not desires as the Humean has it - might motivate us to act, which we might call "besires" since they would both be beliefs and states capable of having a motivating role. To that effect Lewis formulates in decision theoretic terms the notion of a desire-as-belief or "besire", desire-like states that are reducible to belief-like states: the desirability of X is the probability that X is good - and shows that if there were such states, decision theory would be crippled. This is not a knockdown argument against cognitivism about moral motivation, but it sets limits to it. It shows that if one wants to defend cognitivism, decision theory has to be revised. It also clarifies its link with the Humean theory of motivation. Like Lewis' trivialisation result for conditionals, it is a very interesting formal results which tels where a philosophical thesis leads, and where it cannot lead.

But not all formalisations are so successful. The problems raised by other philosophical claims based on formalisation is that either they draw too bold consequences from a formal result,

or that they misconceive the preliminary interpretative step, which consists in setting clearly what the philosophical problem is, and what one can expect from a formal treatment of it.

As an example of the former difficulty, overgeneralisation, we can consider Hans Rott 's comparison between belief change theory and decision theory. Hans Rott , in his impressive book, *Change, choice and inference* (2001) shows important representation theorems, and in particular that all operations of belief change that are generated by rational choice functions, with the choices satisfying certain coherence constraints, satisfy corresponding rationality postulates for belief change. He shows that conversely all operations of belief change that satisfy certain rationality postulates can be represented as operations that are generated by rational choice functions, with the choices satisfying corresponding coherence constraints. What does it show?

The formalisms were devised independently from each other, with different objectives. the correspondence results show that theoretical and practical reason obey the same structure, and since the logic belief change is a special case of the logic of rational choice, the latter has primacy over the former. To simply a lot, take the familiar AGM postulates:

(*1) $K^* \alpha = \mathrm{Cn}(K^* \alpha)$;
(*2) $\alpha \in K^* \alpha$;
(*3) $K^* \alpha \in \mathrm{Cn}(K \cup \{\alpha\})$
(*4) if $\neg \alpha / K$, *then* $Cn(K \cup \{\alpha\}) \subseteq K^* \alpha$
*(*5) if* $Cn(\alpha) = L$, *then* $K^* \alpha = L$
*(*6) if* $Cn(\alpha) = Cn(\beta)$, *then* $K^* \alpha = K^* \beta$.
(*7) $K^* (\alpha \wedge \beta) \subseteq \mathrm{Cn}(K^* \alpha \cup \{\beta\})$;
(*8) if $\neg \beta \notin K^* \alpha$, *then* $K^* \alpha \; K \subseteq (\alpha \wedge \beta)$.
+ Non monotonic principles (where inf (α) *stands from the set of sentences which can be inferred non monotonicall from* α
(Or) $\mathrm{Inf}(\alpha) \cap \mathrm{Inf}(\beta) \subseteq \mathrm{Inf}(\alpha \vee \beta)$
(RMon) if $\beta \notin \mathrm{Inf}(\alpha)$, then $\mathrm{Inf}(\alpha) \subseteq \mathrm{Inf}(\alpha \wedge \beta)$.

Consider now the rational choice postulates:

(I) If $S \subseteq S$, then $S \cap \delta(S) \subseteq \delta(S)$; (Property α)
(II) $\delta(S) \cap \delta(S) \subseteq \delta(S \cup S)$; (Property σ)
(III) if $S \subseteq S$ and $\delta(S) \subseteq S$, then $\delta(S) \subseteq \delta(S)$; (Aizerman's Axiom)
(IV) if $S \subseteq S$ and $\delta(S) \cap S \neq \varnothing$, then $\delta(S) \subseteq \delta(S)$.(Property β +)

Rott's results are that if we are given a choice function on sets of sentences, we can, for define a revision operation by letting $\beta \notin K^* \alpha$ hold if and only if $\alpha \rightarrow \beta \notin \delta(\mathrm{Cn}(\neg \alpha))$, and similarly for contraction and inference. Given such conceptual bridges there is a 1–1 correspondence between postulates for belief change and inference, on the one hand, and postulates for rational choice, on the other hand. Rott interprets his results thus:

"The central and most surprising result [.. .] is that all postulates of theoretical reason are derivable from more general, practical principles of rational choice" (Rott , 2001, p.5).

… "Philosophically, I take this to be a strong indication of the unity of practical and theoretical reason" (Rott 1999:5, 214)

The problem with this interpretation of the results , as it was first noted by Hansson, is that Gardenförs' original axioms of belief change were modelled after Lewis' models for counterfactuals, and as Olsson (2003) has remarked "the revision postulates were, in effect, partly motivated by the fact that in the context of belief models they jointly generate the same logic of conditionals as do prominent choice principles in the context of possible worlds models". If this is the case, there is less than it meets the eye in these results, which show partial overlaps between principles of rational choice and principles of belief change. They do not show that the *whole structure* of theoretical reason is dependent upon that of practical reason. It does not follow that there no such parallels and that Rott's work has not illuminated them. But the parallels have to be motivated by a certain theory of belief change or of inquiry. Isaac Levi has long insisted that we should conceive of inquiry in a decision theoretic way. A large body of work in philosophy is devoted today to bringing together epistemology and the theory of practical rationality, but it is not clear that we can derive the unity of practical and theoretical reason from the structures of their formalisms.

For an example of the second difficulty, insufficient elucidation of the initial formulation of a philosophical problem and overhasty formalisation, let us consider Heinrich Wansing (2006) recent work on doxastic voluntarism. In a very interesting article, Wansing considers how a formalisation within deontic logic can vindicate a certain version of doxastic voluntarism. The paper, however, starts with an unpromising note:

« In this paper, I will not discuss what is maybe the most fundamental question concerning doxastic voluntarism, namely: What exactly does the doxastic voluntarist claim? … Doxastic voluntarism has been characterized by philosophers in many ways, not all being equivalent and some being quite nonspecific or even unclear. In terms of belief formation (understood as an action of belief acquisition), the thesis would be that a doxastic subject may sometimes form a belief (as a result of deciding to form it) » (Wansing 2006: 201)

Wansing is right that the claim is, in much of the recent philosophical literature, unclear. His objective is to show that one can give to the claim a more precise formulation. He starts from the formulation of a kind of deontologism about belief:

a is justified in believing that p if and only if a is permitted to form (or voluntarily acquire) the belief that p.

He then proposes to use the "*Dstit* " modal logic of agency, which is based on the operator :

Dstit p= *deliberately sees to it that* p

and claims that the Dstit language allows us to evaluate arguments against voluntarism , in particular the well known "anti-voluntarist" formulated by Feldman (2000):

(1) if deontological judgements about belief are true, then belief is under the control of the will
(2) but belief is not under the control of the will

(3) therefore deontological judgements about beliefs are false.

Wansing , however, does not deal with this argument, but notes that there can be unintentional actions, hence that we do not need to consider the concept of intention in deal with agency. According to him the Dstit formalisation meets this requirement. After having consider briefly the phenomenological argument against doxastic voluntarism (it is phenomenologically implauisble that we could acquire beliefs at will) and the possibility of indirect doxastic voluntarism (one can bring oneself to believe that p wy indirec means), Wansing proceeds to formaulate a semantics of belief formation ascription, by which ne combines the semantics of the dstit-operator and the sematics of standard doxastic logic . Wansing then uses the Prior - Thomason branching time structures

Each doxastic subject is supposed to be an agent who by her or his actions can influence the future course of the world. In stit-theory this idea is accounted for by assuming that for every individual agent, the histories passing through a moment are partitioned into sets of histories choice-equivalent for the agent. If two histories h and h' are choice-equivalent for an agent a at moment m, then a cannot discriminate by her or his actions at m between h or h '. The sets of histories choice-equivalentfor an agent at a moment m represent the ''choice-cells'' of the agent at m. So we are given the following definitions

α deliberately sees to it that A
[α *dstit*: A] is true in <T, ≥ , Agent, Choice,v >
iff
(i) $\forall$h' choice $^{m}\alpha$ (h) A is true at (m,h')
(ii) $\exists$h'$\in$ H_m such that A is not true at (m,h').

Then we are given an interpretation of ''agent α voluntarily acquires the (implicit) belief (knowledge) that A'' (or ''a forms the (implicit) belief (knowledge) that A'') as [a dstit: $B_{\alpha}A$] ([a dstit: K_{α} A)]
Wansing then claims that the dstit models must be augmented by a doxastic (epistemic) accessibility relation between moment/history-pairs.

At this stage we may ask : where does Wansing takes his notion of forming the knowledge that P *voluntarily*? How can he assimilate in his definitions the notion of an implicit belief with that of an implicit knowledge ? This is puzzling, for the problem of doxastic voluntarism, at least as it has been examined since Williams (1970) seminal paper, is the problem of whether it is possible *consciously* to decide to belief at will. The definition proposed here of a state of implicit belief seems to beg this question.
Wansing then proposes "doxastic distit models"

DEFINITION 2. [α *dstit* :B_{α} A] (α [dstit : K_{α}] is true in the doxastic (epistemic) *dstit* model <T,≤ Agent, Choice, R, *v*> at (m,h)
iff
$\forall$h' $\notin$ 2 Choice$^{m}\alpha$ (h) $\forall$h'' H_m,
if (m,h') R^{m}_{a}(m,h'') then A is true at (m,h'')
(ii) ($\exists$h', h'' $\notin$ Hm such that (m, h') R^{m}_{a}
and A is not true at (m,h'').

We are told that his allows us to make sense of a notion of "deciding to know implicitly". But what is the meaning of this notion ? the assimilation of the concept of belief to that of knowledge, which seems required to fit in the problem with epistemic logic, begs the question, at least because even though it migh make sense to talk of voluntary belief (in a sense to be determined), it makes little sense to talk about *knowing* at will. Knowledge is factive: how could one bring about knowledge of truths at will? Wansing is conscious of the difficulty and says here:

"Initially, it might sound highly implausible that an agent may decide not only to believe implicitly but even to know implicitly. What could this possibly mean? If epistemic voluntarism is a species of doxastic voluntarism, since knowledge is a kind of belief, how should an
agent decide to know something?" (ibid : 220)

Wansing's apparent answer to this question is that there can be voluntary knowledge *formation*:

"An agent α decides to know that p at a moment/ history-pair (m, h) iff at ever moment/history pair (m, h') such that h' is choice-equivalent with h for α at m, a knows that p at (m, h'). It is clear that ''α sees to it that α knows that p'' is satisfiable in some epistemic dstit model if p is neither a tautology nor a contradiction. Thus, if it is true at (m, h) that a sees to it that she knows that p, by her actions at (m, h) the agent can make sure that the future course of events comprises only histories h' such that p is true at every moment/history-pair (m, h'') compatible with what a knows at (m, h'). It is quite conceivable that agent α has the latter capacity…There are unintentional actions. In an unintentional performance of a generic action, the agent need not be aware of performing that action. When voluntarily acquiring the implicit knowledge that p, the agent need not be aware of forming this implicit knowledge." (Wansing 2006 : 219-220)

Wansing also argues that "an agent α can see to it that another agent β forms a certain belief only if α is not independent of α »

The model also allows us to make sense of the following possibility. Thus, if α and β are distinct and mutually independent agents, it is logically impossible that a sees to it that b sees to something. And therefore, under this assumption about α and β, α cannot see to it that b forms a belief. Whereas [α dstit: BβA] is satisfiable, [αdstit: [β dstit: Bβ A]] is not. It follows that α is never responsible for β's acts of belief formation, but may be responsible for β's beliefs. Agent a might, for example see to it that b believes something by applying hypnosis or, perhaps, brain surgery. (Wansing 2006: 222)

If I understand correctly this means that α can make it the case that β believes that p, through indoctrination, hypnosis, or whatever. But this possibility of indirect belief formation , through influencing an other agent's beleifs is granted by all parties within the doxastic voluntarism debate. It was never in question. That the *distit* model allows us to formulate it deos not seem to me to count in favour of the model.

Wansing 's conclusion is that the *dstit* model allows to make sense of, and in this sense of partially vindicating, a kind of doxastic voluntarism. But I do not see how this can be the case without begging a number of questions. In spite of its interest, this attempted formulation of the problem of doxastic voluntarism seems to me to beg too many questions. The very fact that a model allows the kind of "implicitly" knowing voluntarily that p involves a confusion between

the indirect acquisition of a belief, which is indeed quite often voluntary, and its direct acquisition, which is the difficult issue. But this issue is never raised by Wansing's formalisation. It therefore seems to function like an idle wheel.

What has gone wrong here is not the kind of conclusions which one can draw from the model, which may be correct although it is not clear that they show anything about the possibility of believing at will, but the kind of preliminary layout needed to *formulate* the problem. In order to be able to formalise properly, it is necessary in the first place to be clear about the truth conditions of given claims such as : *the can be voluntary belief formation*. And this syndrome is quite widespread. I am not claiming that all formalisations fall into such difficulties, but that it is a permanent temptation of the method that formalism is applied without enough taking into account the nature of the problem at hand. [3]

Conclusion: uses and misuses of formal models

To conclude, I would like to suggest some differences between philosophy and the exercise of modellingin science, since the comparison is often made between the two. Philosophy can use models, but its not modelling. Models, including formal models of various philosophical notions are enormoulsy useful in philosophy. But philosophy is not modelling. In the first place models in philosophy are not like models in science, e.g the Bohr model of the atom: they do not help visualising or give an intuitive summary of a theory which is otherwise developped fully: they *are* theories. For instance bayesianism is a theory of belief, not part of a larger and more comprehensive theory which could be formulated at a later stage. It involves strong assumptions which do not enter innocuously within the attempts, for instance, to model belief change. In the second place, models in philosophy cannot be tested against experiments. They can be tested only against intuitions (this claim is notoriously problematic, since it is often unclear what counts as an intuition and how they are tested, but in any sense of the notion of intuition here, we cannot equate it to experiments in science, in particular answers to questionnaires about a given notion, *pace* "experimental" philosophers, does not count as hypothesis testing in the scientific sense). In the third place, the activity of modelling is not neutral. Models modelise philosophical concepts and theses. Formal models in philosophy do not model uninterpreted concepts; they model already existing philosophical *theories*. But in gneral there is not only one theory in the offing; there are several rival ones. The model must be strong enough to compare several philosophical theories; but this is very rare. In the fourth place, as the example of doaxtic voluntarims above shows, modelisation must not be question begging. This was Hansson's worry about the choice of primitives. A model must not beg the question against a certain philosophical theory in being used to model another one. Certainly, the formaliser is not alone in philosophy in being guilty of this: even an ordinary language philosopher, or a philosopher hostile to logical or mathmetical modelling can beg questions! But the formalist is all the more in danger of doing so that his models have to be interpreted. Philosophy needs to identify the terms of a problem and to explore various ways to solve it. Among these ways, models can play a role. But only in so far as the specific problem is *identified*. If it is not identified, the modelling is idle or misleading. [4]

[3] As it was suggested to me by Igal Kvart, who has an unpublished paper on these issues.

[4] I thank Hannes Leitgeb and Stephan Hartmann for having invited me at the workshop on formal models held in the ECAP meeting in Krakow, Woldek Rabinowicz, Thomas Placek, and the participants in the workshop for their reactions, and Katarzyna Kijania Placek for her editorial work and her patience.

References

Douven E.& Horsten, L. 2008 Formal models in the philosophy of science ", *Studia Logica* (2008) 89: 151–162

Hansson, S.O, 2000 "Formalisation in philosophy", *Bulletin of Symbolic Logic* 6 (2): 162–75, 2000

Hannsson, S.O. 2001 *The structures of Values and norms*, Cambridge: Cambridge university Press

Hartmann, S. 2008 "Modeling in philosophy of science ", in M. Frauchiger and W.K. Essler (eds.), *Representation, Evidence, and Justification: Themes from Suppes* (Lauener Library of Analytical Philosophy; vol. 1). Frankfurt: ontos Verlag 2008

Lewis , D. 1988 "Desire as Belief", *Mind* 97 (418):323-32

Olsson E. 2003 Belief Revision, Rational Choice and the unity of Reason, *Studia Logica* **73**: 219–240

Rott, H. 2001 *Change, choice and inference* , Oxford: Oxford University Press

Wansing, H. 2006 "Doxastic decisions, epistemic justification and the logic of agency" *Philosophical Studies*, 128:201-227