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.