La rationalité est-elle incodifiable ?

Pascal Engel
Université de Paris IV Sorbonne

Résumé : Cet article discute une thèse sous-jacente à la philosophie de Davidson et à sa méthodologie de l'interprétation : le caractère incodifiable de la rationalité. On montre qu'elle est à l'oeuvre dans l'usage du principe de charité par Davidson. A la fin de l'article, je critique cette thèse : les principes de la rationalité de l'interprète ne tombent pas du ciel. Au moins un noyau fort de rationalité psychologique doit pouvoir être codifiable.

Abstract: This paper discusses a thesis which underlies Davidson's philosophy of interpretation: the incodifiability of rationality. It is shown that this thesis is involved in Davidson's use of the principle of charity. At the end of the paper, the thesis is criticized: the principles of rationality do not come from nowhere. At least a strong nucleus of psychological rationality must be codifiable.

Philosophia Scientiæ, 6 (2), 2002, 91–107.

George Eliot écrit, dans *The Mill on the Floss* :

> Toutes les personnes à l'esprit large et fort se méfient instinctivement des gens à maximes ; parce que de telles personnes discernent vite que la complexité mystérieuse de notre vie ne doit pas être embrassée par des maximes, et que nous ficeler dans des formules de cette sorte revient à réprimer toutes les inclinations divines et inspirations qui naissent du progrès de nos intuitions et de notre sympathie. Et l'homme de maximes est pour le sens commun le représentant des esprits qui sont guidés dans leurs jugements moraux seulement par des lois générales, et qui pensent que cela les conduira à la justice par l'intermédiaire d'une méthode évidente et toute faite, sans avoir à s'inquiéter d'exercer la patience, la discrimination et l'impartialité, sans se soucier de s'assurer que leur jugement vient d'une estimation durement acquise de la tentation, ou d'une vie suffisamment riche et intense pour avoir créé une compassion pour tout ce qui est humain. (ed. *Penguin Books*, 7, 2, p. 628) [1]

Il s'agit ici des règles et des maximes morales. Mais ne peut-on appliquer les mêmes conseils de sagesse et de prudence aux règles de rationalité en général ? Existe-t-il un ensemble de formules générales applicables par une « méthode toute faite et évidente » à tous les cas particuliers dans lesquels nous avons à juger un comportement ou une croyance comme rationnels ? Le sentiment contraire, ici fortement exprimé par la romancière anglaise, conduit à soutenir que les règles et les normes de rationalité sont *incodifiables*. Mais si elles le sont, comment notre comportement et nos croyances peuvent-ils obéir à ces normes, et comment celles-ci peuvent-elles avoir un caractère prescriptif ? Comment peut-on soutenir qu'il y a des normes de rationalité qui sont constitutives de toute pensée et de toute action, et qui sont par conséquent douées d'une forme d'objectivité, si l'on soutient en même temps que ces normes ne peuvent pas être énoncées précisément, ou que si elles le sont, elles sont vouées à être vaines ? Il y a là une sorte de paradoxe, que la position qu'un auteur comme Davidson défend au sujet de la rationalité me paraît illustrer également. Mais est-ce vraiment un paradoxe ? Je voudrais essayer de montrer que ce n'en est pas un, et que ce constat permet de résoudre certaines des apories des conceptions contemporaines de la rationalité, en même temps que les difficultés de la position spécifique de Davidson.

I.

Formulée de manière passablement dogmatique, la position de Davidson sur la nature de la rationalité dérive directement de sa position sur la

1. Je dois la référence à W. Child [1994, 58 n].

nature de l'esprit et du langage[2]. On peut l'appeler, comme on le fait souvent, un *interprétationnisme*. Il s'agit d'abord d'une conception de l'interprétation ou de la nature des explications de l'action, de la pensée et du langage, selon laquelle comprendre les actions, les croyances ou les énoncés d'un individu, c'est expliquer ceux-ci en termes de ses raisons. Une thèse célèbre de Davidson est que ces explications en termes de raisons sont une espèce d'explication causale. Mais une thèse non moins célèbre de Davidson est que les explications causales en question sont aussi des explications rationnelles, à la fois au sens où expliquer une action par des raisons c'est être capable de trouver chez l'agent une trame d'attitudes propositionnelles, et au sens où toute formulation d'une telle trame présuppose, de la part de l'interprète, que soient suivies certaines normes de rationalité, qui ont, par rapport à toute attribution d'attitudes de ce genre, ce que Davidson appelle un « rôle constitutif ». En d'autres termes un comportement, un énoncé, ou une croyance qui ne pourraient pas donner lieu à une interprétation conforme à ces normes de rationalité ne seraient pas l'expression d'une action, d'une croyance ou d'une signification. En ce sens, une explication *purement* causale du comportement d'un agent, qui ne révèlerait pas ce comportement comme le produit de raisons rationnelles, mais comme, par exemple, le produit de dispositions physiques ou de régularités seulement naturelles, ne serait pas l'explication de quelque chose de mental. John McDowell a parfaitement exprimé ce trait de la position de Davidson :

> Reconnaître le statut idéal du concept constitutif, c'est se rendre compte que les concepts des attitudes propositionnelles ont leur lieu propre dans des explications de type spécial dans lesquelles les choses sont rendues intelligibles en se révélant être, au moins approximativement, telles qu'elles doivent rationnellement être. Ces explications s'opposent à un style d'explication dans lequel on rend les choses intelligibles en représentant leur exercice comme un cas particulier de ce qui se produit généralement. [Mc Dowell 1986, 389]

Il s'ensuit que tout comportement, toute pensée, et toute signification *doivent* être rationnels, même quand ils n'apparaissent pas comme parfaitement rationnels, et même quand il apparaissent irrationnels. Un sujet qui serait *totalement* irrationnel ne serait pas intelligible, parce que l'explication de son comportement en termes de raisons serait inapplicable. La thèse interprétationniste est donc très forte : elle ne dit pas seulement que nous pouvons interpréter un agent ainsi, mais que nous le devons, sous peine de ne pas le comprendre. Et elle ne dit pas seulement

2. Dans [Engel 1996] j'ai donné une sorte de résumé dogmatique des principales positions de Davidson.

que toute interprétation présuppose des normes de rationalité ; elle dit aussi que toute pensée, tout épisode mental, doit être le produit d'une interprétation.

Cette position soulève toutes sortes de questions. Par exemple, elle présuppose qu'il existe un lien nécessaire, ou « constitutif », entre l'interprétation de la pensée et l'interprétation du langage, et par conséquent qu'une pensée qui ne pourrait pas être exprimable linguistiquement ni être communicable ne serait pas réellement une pensée. Or c'est loin d'être évident. De plus elle présuppose que toute forme de compréhension, par exemple la compréhension d'un langage, est une interprétation. Cela aussi est loin d'être évident. Davidson soutient que l'interprétation est nécessaire à la pensée, et il soutient aussi qu'elle est suffisante pour elle : mais est-ce bien le cas ? On peut très bien admettre, par exemple que *croire que p* implique que l'on soit *interprétable comme croyant que p* (condition nécessaire) sans pour autant admettre qu'être *interprétable comme croyant que p* implique que l'on *croit que p* (condition suffisante). Je ne m'intéresserai pas, néanmoins, ici à ces questions, du moins pas directement. Les questions qui nous intéressent sont les suivantes. Quelle est la nature des normes de rationalité qui sont, selon Davidson, impliquées dans toute interprétation ? En quel sens sont-elles nécessaires à l'interprétation ? Et en quel sens le caractère nécessaire ou constitutif de ces normes interdit-il toute analyse de la rationalité, et par conséquent, si l'on admet l'équation davidsonienne, de toute analyse de la pensée et de l'action en termes empiriques et en termes de dispositions causales ? Autrement dit en quel sens la rationalité et la pensée sont-elles irréductibles à des trames naturelles ?

II.

Intéressons-nous d'abord à la nature et au statut des normes de rationalité en question. En quel sens sont-elles des *normes*, et en quel sens sont-elles des normes *de rationalité* ? Nous pouvons formuler ces questions ainsi : est-ce que des jugements quant à ce qu'un individu croit, fait ou signifie par des expressions qu'il utilise sont des jugements normatifs ?

Un jugement normatif est un jugement au sujet de ce qu'un individu *doit* ou *devrait* faire ou penser, à propos de ce qu'il est *rationnel* pour lui de faire ou de croire, ou au sujet de ce qu'il est *correct* ou incorrect*, juste* ou *injuste*, *bon* ou *mauvais* de faire ou de croire. L'occurrence de termes de ce genre dans des jugements signale l'existence de certains traits qui ne sont pas descriptifs ou factuels, mais évaluatifs ou prescriptifs. Nous appelons couramment de tels traits des *normes*. Mais ces normes ne sont

pas toutes de la même nature. On les range en général dans deux classes : des normes *pratiques* d'une part, quant à ce qu'un individu doit faire, et *épistémiques*, quant à ce qu'un individu doit croire ou penser. Ceci soulève diverses questions traditionnelles. Tout d'abord quelle est la relation entre nos jugements factuels ou descriptifs et nos jugements normatifs ? Quelle est la relation des normes aux faits ? Les normes appartiennent-elles à un domaine ontologique autonome, ou dépendent-elles du domaine des faits naturels ? En d'autres termes, les jugements normatifs sont-ils de véritables jugements, des assertions susceptibles d'être vraies ou fausses, et portant sur un domaine d'entités particulières, ou sont-ils de pseudo-jugements ou de pseudo-assertions qui ne sont que l'expression de nos attitudes ? Ensuite quelles sont les relations entre les différentes classes de normes, épistémiques et pratiques ? Sont-elles exclusives ou bien y-a-t-il des traits communs et des recoupements au moins partiels entre les normes épistémiques et les normes pratiques ? Sont-elles exhaustives, ou bien y-a-t-il d'autres sortes de normes à l'œuvre dans nos jugements normatifs ?

Commençons par le second groupe de questions. Davidson, quand il analyse l'interprétation du langage et des croyances, semble bien soutenir que cette interprétation est régie par des normes épistémiques distinctes des normes qui valent pour l'interprétation des actions, qui sont spécifiquement pratiques. Au nombre des premières il compte en particulier le fameux principe de charité, selon lequel un interprète doit présupposer que les croyances de celui qu'il interprète sont, dans leur majorité, *vraies* et *cohérentes*. On tient souvent que ce principe porte sur le nombre et la nature des croyances de l'interprété, et sur la proportion des croyances vraies et rationnelles par rapport à celles qui seraient fausses ou irrationnelles. C'est correct ; mais c'est oublier que le principe de charité vise d'abord à énoncer certains principes qui sont constitutifs de la *notion* même de croyance, que l'on peut expliciter de la manière suivante. Un sujet ne peut croire le contenu d'une proposition que s'il croit en même temps que cette proposition est *vraie*, s'il est capable de donner son assentiment à sa vérité, ou autrement dit, pour employer son vocabulaire, que s'il la tient-pour-vraie. Ce lien est mis particulièrement en valeur par le fameux « paradoxe de Moore » : il est bizarre d'asserter des phrases de la forme « p mais je crois que *non* p » ou « p mais je ne crois pas que p », parce que c'est le *but* même d'une assertion que d'exprimer une croyance, et une croyance *vraie*. En ce sens même un menteur fait une assertion et doit se représenter lui-même comme disant la vérité, même s'il sait qu'il ne la dit pas. Ce point, souvent noté, au sujet de l'acte de langage d'assertion n'est pas propre à l'assertion lin-

guistique, ou à l'énonciation ; il est propre aussi à la pensée elle-même : un individu qui ne se représenterait pas comme croyant que p est vrai ne pourrait pas non plus se représenter comme croyant que p. Davidson refuse de dire que c'est là un lien conventionnel, parce que c'est pour lui un lien constitutif ou nécessaire. Sans l'attitude propositionnelle de « tenir-pour-vrai » un énoncé potentiel, l'interprétation ne peut pas commencer [3]. C'est en ce sens que l'on peut dire que la vérité est présente comme une norme de la croyance : elle est ce que vise la croyance. En second lieu, croire que p ce n'est pas seulement croire que p est vrai, mais aussi croire que p est *justifié* ; il est essentiel à une croyance, et au fait de se représenter comme ayant une croyance, qu'on soit en mesure de la justifier, c'est-à-dire d'en donner les raisons. Or donner des raisons d'une croyance, ou la justifier sont des notions intrinsèquement normatives. Ce ne sont pas des notions factuelles. On est libre, ensuite, de donner une certaine analyse de la notion de justification, et Davidson semble ici user de plusieurs sortes d'analyses de cette notion, y compris une analyse causaliste. Quand il dit que la plupart des croyances d'un sujet doivent être vraies, il semble s'appuyer sur un critère fiabiliste de la justification des croyances : un système de croyance est fiable s'il est capable de produire un ensemble de croyances vraies suffisamment important, et ceci peut dépendre de propriétés causales et factuelles de l'organisme et de ses relations à l'environnement. A d'autres moments, Davidson semble s'appuyer sur un critère typiquement « évidentialiste » (au sens de l'anglais *evidence* désignant les données qui peuvent être utilisées comme preuves) de la justification des croyances, par exemple quand il défend le principe d'information totale de Carnap [4], et à d'autre moments encore, il défend nettement une analyse cohérentiste [5]. Quelle que soit en définitive son analyse de la justification des croyances, la notion même de justification fait partie du concept de croyance, qui en ce sens est un concept normatif.

Quand il traite des notions d'action, de désir, ou de raisonnement pratique, Davidson soutient parallèlement qu'il existe des normes pratiques qui s'attachent à toute notion de ce type. Les normes en question ne sont pas tant des normes morales, ou déontiques, au sens usuel de ce terme, que des normes du genre suivant : « Si un agent accepte une raison (ou un ensemble de raisons) r, et s'il tient qu'il est préférable de faire a plutôt que b, étant donné ces raisons r, alors, toute choses égales par ailleurs, il jugera qu'il vaut mieux pour lui faire a que b » (c'est le

3. Cf. « Communication and convention » *in* [Davidson 1984, 274 sq].
4. Dans le contexte de son analyse de *self deception* cf. [Davidson 1985].
5. « A Coherence Theory of Truth and knowledge », 1985a.

« principe de continence » pour le raisonnement pratique, que Davidson met en parallèle avec le principe de l'information totale de Carnap pour le raisonnement inductif) [6]. D'autres principes propres aux désirs et aux préférences seront ceux de la théorie de la décision bayésienne (comme celui de la transitivité des préférences, et celui de la maximisation de l'utilité espérée).

Qu'il s'agisse, par conséquent, des croyances ou des désirs, Davidson soutient qu'il existe des normes intrinsèquement liées à ces concepts, à la fois épistémiques et pratiques, qui font partie de leurs conditions d'intelligibilité. Ces normes sont aussi des normes de rationalité : elles indiquent à quelles conditions un agent est susceptible d'être rationnel quand il a des désirs et des croyances, en même temps qu'elles nous disent ce que c'est, pour une croyance ou un désir, que d'être rationnel. Cela ne répond pas encore à la question que nous posions plus haut, celle de savoir si ces normes épistémiques et pratiques sont exclusives et exhaustives. Pour répondre à cette question, il nous faut d'abord répondre au premier groupe de questions, qui portent sur la relation des normes aux faits.

III.

On remarque souvent qu'il est bien trop simple, et également erroné, de diviser nettement les propriétés en deux classes, normatives d'une part et naturelles ou factuelles de l'autre, et de diviser parallèlement les jugements en deux classes, les jugements normatifs d'une part et les jugements factuels ou descriptifs de l'autre. Car nombre de propriétés « normatives » le sont en vertu de traits descriptifs ou factuels, et nombre de traits factuels contiennent des éléments normatifs. L'adjectif « bon », par exemple, est de toute évidence un prédicat évaluatif ; mais le fait qu'un couteau soit un bon couteau dépend aussi, de toute évidence, de propriétés naturelles du couteau, descriptibles en termes factuels— par exemple que sa lame soit aiguisée de telle ou telle façon. De même le fait que ce soit la norme, par exemple dans un certain pays, que les écartements des rails de chemin de fer soient tels ou tels, dépend clairement de propriétés descriptives quant à cet écartement, ou encore le fait que ce soit la norme pour un joueur de basket d'avoir telle taille, par exemple plus de deux mètres, dépend clairement de propriétés naturelles des joueurs de basket. En ce sens, il est clair qu'il existe des liens de dépendance entre les traits normatifs et les traits factuels, et qu'il y a

6. C'est l'un des principes que Davidson utilise dans son analyse de l'akrasia. Cf. « How is Weakness of the Will possible ? », [Davidson 1980, 37-38], et qu'il appelle « le principe de continence ».

des passages du « est » au « doit » [7]. Il en est de même des descriptions du comportement ou des croyances d'un agent en termes de normes de rationalité. Ces descriptions nous disent ce qu'*idéalement* un agent *devrait* croire ou faire, étant donné ses croyances et ses actions. Mais il est très difficile de dire ce qu'un agent devrait croire ou faire sans considérer ce qu'il croit ou fait effectivement. Mais ce qu'il croit ou fait ne peut être caractérisé qu'à la lumière de principes normatifs quant à ce qu'il devrait croire ou faire. Davidson soutient précisément qu'il y a une forme d'échange entre le descriptif et le normatif au sujet de la théorie bayésienne de la décision :

> It is common to make a strong distinction between Bayesian theories as normative and as descriptive. As a picture of how a perfectly rational agent should act, such theories are sometimes allowed to be essentially correct, though perhaps oversimplified. (through failure to consider the cost of information or computation, for example). As descriptive theories, however, they are considered to be at best limited in application and absurdly idealised. Many experiments seem to bear out this view.
>
> I doubt that there is an interesting way of understanding the purported distinction. Until a detailed empirical interpretation is given to a theory, it is impossible to tell whether an agent satisfies its norms ; indeed, without a clear interpretation it is hard to say what content the theory, whether normative or descriptive, has. Nor does it make sense to say that in a normative application one is not interested in the truth of the theory, since the question whether someone acts, or preferences and beliefs, are in accord with the theory is just the question whether or not the theory is true of him. On the other hand (and more important) is the fact that the concepts of thought, choice, and intentional action are so laden with normative considerations that the theories that employ these concepts cannot be tested for empirical validity without the use of normative standards. Decision theory deals with intentional choices, and supplies a rationale for such choices. But the general idea that intentional actions can be rationalized is no invention of decision theory ; it is part of the everyday concept of intentional action [8].

On peut tirer deux idées importantes, et complémentaires, d'un tel passage. La première porte sur l'application de modèles d'action ou de croyance rationnelle, tels que peuvent nous en fournir la théorie de la décision et la logique, à des individus dont nous désirons interpréter le comportement. Ces modèles sont « normatifs » au sens où ils nous disent, idéalement ce qu'un agent devrait faire ou croire. Si ces descriptions idéales ne peuvent pas s'appliquer à ces agents, alors il devrait s'ensuivre qu'ils ne sont pas rationnels au sens où la théorie normative le

7. Cf. Searle « How to derive *ought* from an *is* » in [Searle 1979].

8. « A New basis for decision theory », [1985b, 89]

prescrit. Par exemple un agent qui croirait que *p*, et qui croirait que *si p alors q,* mais qui ne croirait pas, sur la base de ces croyances, que *q*, serait en ce sens irrationnel, ou en tous cas moins que rationnel. Ou encore un agent qui préfèrerait *a* à *b*, *b* à *c*, mais qui ne préfèrerait pas *a* à *c*, ou encore qui ne maximiserait pas son utilité espérée, serait irrationnel, ou moins que rationnel. Mais les expériences de psychologie du choix et des jugements de probabilité, comme celles qu'ont menées depuis les années 1950 des auteurs comme Suppes et Davidson lui-même, et à leur suite des auteurs comme Kahneman et Tversky, de même que les expériences des psychologues dans le domaine du raisonnement semblent montrer que les agents dévient fortement par rapport à ces normes de rationalité [9]. Doit-on en conclure que les sujets humains soumis à ces expériences sont irrationnels ? Si l'on suit ce que dit Davidson dans le passage cité, non. Car il n'est pas possible de distinguer nettement les principes de rationalité normative incorporés dans la théorie de la décision et dans la logique déductive en tant que ceux-ci sont vrais d'agents idéalement rationnels et les descriptions de ces agents comme ayant certaines croyances et certains désirs. Cela ne veut pas dire que l'on ne puisse pas, localement, déterminer certains de leurs comportements comme irrationnels ou comme déviants par rapport à ces principes. Mais l'idée qu'ils puissent être totalement irrationnels est elle-même inintelligible. Car le fait même que l'on puisse leur attribuer les croyances correspondantes, déductives ou probabilistes, suppose déjà que l'on soit en mesure de leur appliquer les normes de rationalité en question, autrement dit qu'ils soient en général, rationnels. Il en est de même pour les cas d'irrationalité bien connus étudiés par Davidson, tels que l'*akrasia* ou la duperie de soi. L'un des paradoxes de l'irrationalité, mais qui est aussi un paradoxe de la rationalité, est que tout jugement quant à l'irrationalité d'un comportement ou d'une croyance doit s'effectuer sur fond d'une présomption générale de rationalité, selon laquelle un agent doit être, dans l'ensemble, rationnel, même si sa rationalité est moins que parfaite. Il s'ensuit qu'il n'est pas possible, selon Davidson, de séparer nettement ce qui est vrai, empiriquement, d'un agent humain, et ce qui est vrai, idéalement, de cet agent. La condition même de l'interprétabilité du comportement présuppose sa rationalité, et sa conformité à des normes de rationalité. Dans la mesure où ces normes sont nécessaires, elles sont sans exception : tout agent ou sujet rationnel doit les exemplifier. Mais en même temps, on ne peut pas dire qu'elles sont vraies d'un agent donné, ni dire en quoi elles s'appliquent. Comment est-ce possible ?

9. Cf. en particulier les remarques de Davidson dans « Hempel on Explaining Action », in [Davidson 1980, 272 sq].

IV.

C'est ici que l'on touche au point essentiel : tout comportement, et toute croyance, selon Davidson, doivent répondre à ces normes, mais celles-ci ne sont pas codifiables. Si elles étaient codifiables, cela voudrait dire qu'il est possible de partir d'un ensemble de règles et de principes fixes, et d'un ensemble de circonstances, pour en déduire une spécification complète de ce que l'on doit faire ou croire dans ces circonstances. Et dans la mesure où les normes de rationalité sont les normes de l'interprétation, cela veut dire qu'il n'y a pas d'ensemble défini de principes permettant de parvenir à la meilleure interprétation possible d'un agent. Quand nous interprétons quelqu'un d'autre, nous essayons de donner un sens à son comportement et à ses croyances. Mais en faisant cela, nous nous appuyons sur notre propre conception de la rationalité en même temps que nous nous appuyons sur les canons généraux de rationalité que codifie la logique ou la théorie de la décision. Et nous pouvons faire appel à ces ressources sans limite. Dire que la rationalité n'est pas codifiable ne veut pas dire qu'elle n'est pas codifiée : elle l'est, manifestement, et c'est en ce sens qu'il existe des manuels de logique et des manuels de théorie de la décision. Mais même s'il existe de tels manuels, qui énoncent des principes supposés être vrais de tout agent idéal, cela n'implique pas qu'il existera, pour toute question de la forme : « Que devrais-je faire ou croire dans ces circonstances ? » un principe universel de rationalité théorique ou pratique. Par exemple des principes normatifs épistémiques tels que le suivant : « Croyez toujours ce que vous avez les meilleures raisons possibles de croire » semblent bien être universellement vrais et sans exception, de même que des règles du genre : « Si vous croyez que p et que *si* p *alors* q, alors vous devriez croire que q ». Mais ils n'indiquent pas ce que, dans une occasion donnée, on doit croire. Le second principe est vide, si l'on ne dit pas en quoi consistent nos « meilleures raisons ». Le premier est faux dans certains cas : un sujet, par exemple, qui croit que p et que *si* p *alors* q, mais ne croit pas que q n'est pas nécessairement irrationnel ; il peut être en fait rationnel pour lui, s'il a de bonnes raisons de croire que *non* q, de ne pas inférer q de ses autres croyances ; il peut être, dans ces circonstances, rationnel pour lui d'abandonner sa croyance que p ou sa croyance que *si* p *alors* q, ou même les deux. Mais il n'y a pas de recette générale qui nous permette de dire *a priori* laquelle des bonnes stratégies est la bonne. La logique épistémique étudie bien un certain nombre de règles de changement et de révisions de croyances de ce genre, et elle peut spécifier des révisions qui sont plus rationnelles que d'autres, par exemple qu'on ne doit pas ajouter une croyance à son stock

initial si on obtient une contradiction explicite. Mais face à une contradiction donnée, la logique épistémique ne nous donne pas de formule générale pour savoir laquelle de nos croyances on doit abandonner. Ce trait est directement lié à une caractéristique des croyances sur laquelle Davidson, à la suite de Quine, met l'accent : leur caractère holistique. Face au choix entre réviser l'une de ses croyances particulières ou réviser la théorie générale dont elles dépendent, on a toujours en principe le choix. Et le holisme des croyances est l'une des raisons de l'indétermination de l'interprétation, qui est une autre version de l'incodifiabilité de la rationalité.

Une autre manière de considérer le même point consiste à s'interroger sur la nature du principe de charité. On dit souvent que ce principe ne peut pas être une bonne maxime d'interprétation, parce que les croyances des agents sont souvent moins que rationnelles, et souvent également fausses, alors que le principe en question nous prescrit de croire le contraire. L'usage immodéré de la charité n'aboutira-t-il pas à la conclusion paradoxale qu'ils doivent être rationnels et véridiques même là où nous avons toutes les bonnes raisons de ne pas les trouver rationnels ou véridiques ? On en appelle alors souvent à un principe d' « humanité » selon lequel nous devons attribuer aux agents des croyances qui sont rationnelles et véridiques à proportion de la rationalité et de la véridicité de *nos propres* croyances. Mais Davidson lui-même nous enjoint à ne pas considérer le principe de charité comme un principe de *maximisation* de l'accord, mais comme un principe d'*optimisation* de la compréhension. En ce sens, il n'y a pas de différence entre le principe de charité et le principe d'humanité. Et dire que nous devons optimiser l'accord, étant donné la manière dont nous comprenons une situation revient précisément à dire que la charité et la rationalité ne sont pas codifiables.

On peut, enfin, parvenir à la même idée en considérant une stratégie usuelle pour tester la rationalité d'un comportement en théorie de la décision. Les théoriciens bayésiens supposent que nous avons des degrés de croyance qui sont des degrés de probabilité. Il s'ensuit que si les croyances d'un agent violent certaines normes du calcul des probabilités, ils deviennent incohérents. Cette incohérence, selon eux, se manifeste par le fait qu'on peut alors faire contre eux des « paris hollandais » (*Dutch books*) tels que, pour tout pari qui leur est présenté sur leurs croyances, ils seront perdants. Le prix de l'incohérence est alors clairement une perte d'argent : rien n'est plus irrationnel, en apparence, qu'être une « pompe à fric ». Supposez, par exemple, que votre degré de croyance en une proposition p soit le double de votre degré de croyance en sa négation, *non p,* et que votre degré de croyance en une autre proposition q soit le même

que votre degré de croyance en sa négation. Supposons aussi que p implique q. Comme une proposition ne peut pas être plus probable que la proposition qu'elle implique, vos degrés de croyance violent le calcul des probabilités. Si vous acceptez des enjeux dans un pari qui reflètent ces degrés incohérents de croyance — par exemple si vous jouez à 2 contre 1 la vérité de p et à 1 contre 1 la vérité de q — vous serez à la merci d'un pari hollandais, dans lequel votre adversaire pourra faire un ensemble de paris contre vous que vous perdrez systématiquement. Par exemple votre adversaire parie 15 € sur la vérité de q et 10 € sur la fausseté de p. Si q est vrai, vous perdrez 15 € sur q. Vous pourrez toujours gagner le pari de 10 € sur p, mais même si c'est le cas, vous perdrez toujours 5 €. D'un autre côté, si q est faux, vous gagnerez 15 € sur votre adversaire qui a parié sur q. Mais puisque p implique q, p est faux. Donc vous perdrez 20 € sur votre adversaire sur p. En ce cas aussi vous perdrez de l'argent.

L'argument dit du « pari hollandais » vise à rendre équivalentes l'incohérence et la perte d'argent, et à soutenir qu'il est toujours irrationnel d'avoir des degrés de croyances incohérents. Mais il n'y a aucune démonstration générale de ce résultat. Est-il toujours irrationnel d'avoir des degrés incohérents de croyance ? Ce n'est pas plus irrationnel que d'avoir des croyances entières (i.e. des croyances sans degrés, ou de degré 0 ou 1) contradictoires. Car on peut avoir les croyances en question, mais ne pas reconnaître qu'elles le sont. L'argument du Dutch Book est en ce sens aussi peu concluant que l'argument selon lequel un agent rationnel devrait inférer toutes les conséquences logiques de ses croyances. Car il n'est pas un agent parfait, et certaines des conséquences de ses croyances peuvent lui échapper. Tout ce que l'argument du pari hollandais montre est qu'il est irrationnel d'accepter des paris sur des enjeux qui reflètent vos degrés de croyance quand ces degrés de croyance sont incohérents. Mais vos degrés de croyance sont une chose, et les paris que vous acceptez sur eux en sont une autre. Il se pourrait parfaitement que vous ayez des degrés de croyance incohérents, mais que les raisons que vous avez d'accepter des paris sur eux ne le soient pas. Le pari de Pascal illustre une possibilité de ce genre. Il est incohérent de parier sur la vérité de la proposition « Dieu existe » alors même que cette proposition a toutes les chances d'être fausse. Mais ce n'est pas incohérent, étant donné toute l'éternité de bonheur que ne manquerait pas de vous procurer la vérité de cette proposition, à supposer, évidemment que vous ayez aussi de bonnes raisons de croire aux vérités de la religion chrétienne. Mais n'auriez vous pas aussi de bonnes raisons d'y croire, simplement en tant que maximisateur de votre utilité espérée ? Une irrationalité épistémique peut bien correspondre à une rationalité pratique. De même qu'il peut

être rationnel d'avoir des croyances incohérentes, il peut être rationnel de perdre de l'argent. Pourquoi le fait de perdre de l'argent serait-il toujours une marque d'irrationalité ?

Ceci montre, au passage, que les critères de la rationalité pratique et ceux de la rationalité épistémique ne sont pas totalement imperméables les uns par rapport aux autres, et que la rationalité d'un comportement ou d'une croyance dépend largement du type de but que l'on poursuit. Cela implique-t-il que toute rationalité soit instrumentale au sens large, c'est-à-dire au sens de la relation moyen-fin en général, comme le soutiennent les avocats du modèle du choix rationnel en sciences sociales ? En un sens oui, mais à condition d'étendre le critère de la rationalité au-delà des règles de la simple cohérence. Un exemple simple, emprunté au raisonnement économique, peut le suggérer. Les économistes nous prescrivent d'adopter une règle dite « de profits et pertes » (*sunk cost*) selon laquelle, une fois que l'on a effectué un choix qui se révèle défavorable, il faut en faire passer le coût aux profits et pertes, et ne pas assumer les conséquences de ce choix. Par exemple, si vous vous installez dans un restaurant et payez d'avance votre addition, et que le repas se révèle exécrable, il est rationnel de quitter la table et de ne pas manger le repas, ou si vous avez payé un ticket de cinéma pour découvrir un navet, il vaut mieux quitter la salle. Ajouter au prix payé le désagrément d'une indigestion ou d'un mauvais moment serait irrationnel. Mais combien d'entre nous n'ont pas quitté la salle de restaurant même après avoir constaté que le repas était mauvais ou la salle de cinéma après avoir réalisé que le film était nul ? Notre comportement est-il nécessairement irrationnel ? De même un couple âgé peut vouloir rester dans une maison devenue trop grande et trop coûteuse après le départ de ses enfants devenus adultes. « On a toujours habité là » est la raison qu'ils donnent. Sont-ils nécessairement irrationnels ? Le point n'est pas seulement qu'un comportement qui apparaît irrationnel au regard de certains critères (économiques) ne l'est pas nécessairement au regard d'autres (le désagrément d'avoir à vivre dans un autre environnement que celui où l'on a toujours vécu), mais qu'il ne semble pas y avoir de limite nette dans la recherche d'objectifs qui rendraient rationnel le comportement suivi [10]. C'est là un autre exemple de l'incodifiabilité de la rationalité.

10. Je ne veux pas dire par là qu'il faut s'interroger, outre sur la rationalité instrumentale des agents, sur la rationalité de leurs valeurs (pour reprendre la distinction weberienne traditionnelle), et que l'incodifiabilité de la rationalité est un autre nom de ce que Weber appelle « la guerre des dieux » (le pluralisme intrinsèque des valeurs). Ce que je veux dire est que, étant donné une interprétation d'un comportement comme rationnel en vertu d'un objectif ou d'une valeur V, il est toujours possible de le considérer également comme rationnel (ou irrationnel) en vertu d'un objectif ou

L'incodifiabilité de la rationalité n'est en fait qu'un autre nom du principe de rationalité lui-même. Elle est aussi une conséquence directe de l'usage étendu du principe de charité. Quand on objecte, comme on le fait souvent, à ce principe, en disant qu'il nous oblige à traiter les individus comme idéalement rationnels, alors que manifestement ils ne le sont pas, on confond le rôle normatif du principe avec un rôle descriptif. Quand j'envisage de traiter une autre personne comme ayant des croyances contradictoires, je ne peux pas faire autre chose qu'établir cette incohérence en fonction de mes propres critères de détection des incohérences (de mes propres normes). On ne peut pas espérer que l'autre personne fasse autre chose : elle interprète d'après ses propres normes. Et il y a peu de chances pour que nous divergions. En d'autres termes, même si les agents ne satisfont pas *en fait* à ces normes, les interpréter suppose que nous nous les représentions comme *tendant* à y satisfaire. Mais le comportement des agents réels ne « correspondra » jamais qu'imparfaitement à ces idéaux, de même que la manière dont leurs comportements entrent en rupture avec ces normes ne sera jamais établie strictement, ce qui signifie aussi que la question de savoir en quoi ce sont des idéaux n'est jamais claire non plus. Il est de l'essence du principe de charité de ne pas s'appliquer — ou de manquer de s'appliquer — de manières qui soient codifiables ou explicitables complètement.

V.

La thèse de l'incodifiabilité de la rationalité n'est pas propre à Davidson. Elle a été également défendue par des auteurs comme McDowell à partir d'une lecture de l'éthique d'Aristote et de Wittgenstein. Elle va de pair, comme on l'a vu avec la citation de George Eliot, avec une forme de ce que l'on a appelé un particularisme en éthique, qui rejette l'emploi des règles morales et toute codification en morale [11]. C'est une thèse attrayante, qui satisfait à la fois nos intuitions quand il s'agit d'examiner le statut des normes épistémiques et celui des normes pratiques : même si l'on admet qu'il y a des règles de rationalité, il ne semble tout simplement pas y avoir de possibilité de déduire quelque comportement conforme à ces règles. J'ai longtemps cru que cette thèse était correcte [12]. Mais j'en suis moins sûr. La stratégie idéalisatrice impliquée par le principe de charité suppose, comme on l'a vu, que quand je détecte des incohérences

d'une valeur V', ou d'une autre valeur V", et qu'il n'y a pas de fait probant qui autorise une interprétation plutôt qu'une autre.

11. Dans son *Mind, value and reality* (1998) McDowell indique p. 399 dans son index plus de dix occurrences de cette notion, bien qu'il n'emploie pas le terme.

12. Cf. Engel 1990, 1992. Contra, cf. Engel 1997, 1999.

chez autrui selon *ma* perspective et quand il détecte des incohérences chez moi selon *sa* perspective, il finira par y avoir une convergence entre nous, parce que nous nous interprétons tous les deux comme tendant vers les idéaux de cohérence rationnelle et logique. Le problème, avec ce raisonnement n'est pas dans la conclusion — nous convergerons — mais dans la prémisse : on suppose que mes capacités et les siennes pour détecter ses/mes incohérences dans son/mon discours et ses/mes actions sont précisément les *mêmes* capacités, et on suppose justement l'existence de ces capacités. En d'autres termes, l'interprétationnisme suppose que nous avons les aptitudes nécessaires pour distinguer, dans une trame de comportement ou de discours, ce qui est rationnel de ce qui ne l'est pas. Mais d'où viennent ces aptitudes ? Nous sont-elles échues par magie ou par don divin ? Quand nous détectons que deux croyances sont contradictoires, par exemple, nous obéissons bien à certains critères. Il en est de même quand nous détectons une incohérence dans un comportement. Nos jugements quant à la rationalité d'une croyance ou d'une action doivent bien pouvoir s'expliquer. Ce n'est pas parce que nous ne pouvons pas tous les spécifier qu'il n'y a aucune théorie possible des critères par lesquels nous reconnaissons quelque chose comme rationnel. L'interprétationnisme de Davidson suppose que l'interprète a des critères de rationalité, mais ne dit pas d'où il les tient. Il est plus plausible de supposer que ces critères proviennent de certaines aptitudes, qui sont basées dans des compétences des agents. La même conclusion peut être atteinte si l'on considère les échecs de la rationalité inférentielle manifestés par les nombreuses expériences de psychologie du raisonnement. Si la plupart des humains font des erreurs élémentaires de raisonnement, ce n'est pas parce qu'on ne peut les interpréter du point de vue de normes de rationalité idéale, mais parce qu'on a de bonnes raisons de penser qu'ils ne suivent même pas ces normes. Ces échecs peuvent être expliqués par une *psychologie* du raisonnement humain. La même difficulté affecte la théorie davidsonienne de l'interprétation du langage. Davidson soutient que comprendre un langage suppose toujours le point de vue d'un interprète, et que cet interprète a accès aux significations des expressions d'autrui parce qu'il les confronte à *ses* propres interprétations. Mais pour éviter une régression à l'infini, Davidson suppose que l'interprète a directement (et non par le biais d'une interprétation) accès à *ses* propres significations (un accès privilégié). Mais comment expliquer cet accès ? Est-ce un pur et simple donné ? Ou bien quelque chose qu'on peut expliquer ? En d'autres termes, comment l'interprète acquiert-il les significations qu'il va ensuite confronter, par la procédure décrite par Davidson, à celles d'autrui. Il a bien fallu qu'il apprenne son langage. Mais cette capacité n'a rien de

mystérieux. Elle est sans doute explicable. Mais si elle l'est, alors cela signifie que la connaissance tacite qu'il a des règles sémantiques de sa propre langue est une capacité cognitive qui ne peut pas simplement être expliquée par l'interaction des agents dans la communication [13] Elle peut l'être par une théorie psychologique de la compétence linguistique. La source ultime du scepticisme davidsonien quant à la codifiabilité de la rationalité provient évidemment de son « monisme anomal », dont une des composantes est le déni qu'il existe des lois psychologiques strictes. Ce que je viens de suggérer est qu'il existe au contraire des lois psychologiques spécifiables des compétences des agents à partir desquelles ont peut évaluer leur comportement rationnel, et par conséquent que l'anomie du psychologique n'est pas aussi radicale que le soutient Davidson [14]. Cela veut dire que l'irrationalité, comme la rationalité, ne sont sans doute pas des propriétés aussi incodifiables et inexplicables que des auteurs comme Davidson ou McDowell le soutiennent. Ou en tout cas, qu'il leur faut bien expliciter quelles sont les compétences des interprètes *indépendamment* de la méthode d'interprétation, sous peine de présupposer ce qui est en question. Et je ne vois pas comment cela peut être fait sans supposer que la rationalité, ou l'irrationalité est, au moins jusqu'à un important degré, le produit de certaines capacités psychologiques naturelles des agents. Cela veut dire qu'elle ne peut pas être absolument incodifiable.

Bibliographie

Child, B.

1994 *Causality, Interpretation, and the Mind* , Oxford : Oxford University Press.

Davidson, D.

1980 *Essays on Actions and Events*, Oxford University Press, tr. fr *Actions et événements*, Paris : PUF 1993.

1984 *Inquiries into Truth and Interpretation*, Oxford University Press, tr. fr. *Enquêtes sur la signification et la vérité*, Nîmes : J. Chambon.

1985 *« Deception and division »*, in Le Pore, ed. *Actions and Events*, Oxford : Blackwell, tr. fr. in *Paradoxes de l'irrationalité*, Combas : L'Eclat 1991

1985 a *« A Coherence Theory of Truth and Knowledge »*, in E. Le Pore, ed *Truth and Interpretation*, Oxford : Blackwell.

1985 b *« A New Basis for Decision Theory »*, *Theory and Decision*

13. C'est ce que j'ai soutenu, contre Davidson, dans le dernier chapitre de [Engel 1994a]. La difficulté est particulièrement bien analysée dans [Smith 1998].

14. Cf. notamment [Engel 1996a] (notamment p. 203) sur ce point.

Engel P.

1994 *« Trois formes de normativité »*, in Engel, ed. *Lire Davidson*, Combas : L'Eclat

1994 a *Davidson et la philosophie du langage*, Paris : PUF

1996 *« Le petit Davidson portatif »*, in Engel, ed. *Davidson analysé*, Cahiers de philosophie de l'Université de Caen, 31.

1996 a *Philosophie et psychologie*, Paris : Gallimard, Folio.

1997 *« Davidson on meaning, understanding, and normativity »*, in Cardenos de Philosophia, Universidade Novella de Lisboa, Lisbonne, 3.

1999 *«The norms of the Mental »*, in L. Hahn, ed. *The Philosophy of Donald Davidson*, Open court, La salle, Ill.

Mc Dowell, J.

1997 *Mind, Value and Reality*, collected papers, Harvard : Harvard University Press.

Smith, B.

1988 *« On knowing one's own Language »*, in C. Wright, B.C. Smith, and C. McDonald, eds, *Knowing one's own Minds*, Oxford, Oxford University Press, 1998, 391-428.