Soyons charitables, mais pas trop !

Ruwen Ogien
CNRS

Résumé : Dans cet article, j'essaie de mettre en évidence deux limites du principe de charité. 1) Le principe est censé nous aider à résoudre le problème de la sous-détermination observationnelle des états mentaux. Mais il échoue de ce point de vue. Le principe permet d'exclure les attributions de croyances qui ne respectent pas le principe, mais il ne permet pas de départager les attributions de croyances en conflit qui respectent également le principe ; 2) Dans ses versions fortes, le principe de charité exclut la possibilité de l'irrationalité motivée. Ceux qui croient à la possibilité de l'irrationalité motivée ne peuvent accepter raisonnablement les versions fortes du principe de charité.

Abstract: In this paper, I insist on two limits of the principle of charity. 1) The principle is supposed to help us solve the problem of the observational under-determination of our attributions of beliefs and desires. But it fails on that ground. The principle rules out attributions of beliefs and desires that are incompatible with the principle, but it cannot help us decide between incompatible attributions of beliefs and desires that are compatible with the principle; 2) Strong versions of the principle rule out the possibility of motivated irrationality. If one believes in the possibility of motivated irrationality, one cannot reasonably endorse strong versions of the principle of charity.

Philosophia Scientiæ, 6 (2), 2002, 109–125.

A première vue, le principe de charité interprétative n'est rien d'autre qu'une explication théorique d'un fait brut concernant notre psychologie ordinaire : notre tendance à chercher (et à trouver, bien sûr) des raisons aux comportements humains, même à ceux qui nous paraissent les plus gratuits ou les plus irrationnels. C'est cette tendance, je crois, qui nous rend tellement sensibles aux aphorismes des moralistes du XVIIe siècle, à leur façon d'expliquer par des motifs purement égoïstes, c'est-à-dire rationnels au sens instrumental, des comportements apparemment sublimes ou désintéressés, c'est-à-dire apparemment absurdes ou irrationnels. Mais nous sommes également attirés par des explications rationalistes qui ne sont pas étroitement égoïstes ou instrumentales. Celles qui insistent sur l'importance de la recherche d'une certaine forme de cohérence dans nos conduites ou celles qui soulignent nos engagements à ne pas nier les évidences ne nous semblent pas moins séduisantes. En fait, nous estimons de façon très générale que, si un agent fait quelque chose, c'est parce qu'il a une *bonne raison* de le faire, qu'elle soit égoïste ou pas, qu'elle soit évidente ou cachée, qu'elle nous paraisse bizarre ou tout à fait naturelle [Boudon 1995].

Cependant, c'est une chose de mettre en évidence ce fait brut de notre psychologie ordinaire, c'en est une autre de proposer une explication théorique de ce fait. Selon Davidson, c'est un ensemble de contraintes *logiques* liées à la possibilité même de comprendre autrui qui explique notre tendance à donner des raisons aux comportements humains, même à ceux qui nous paraissent les plus gratuits ou les plus irrationnels. Pour Daniel Dennett, cette tendance est plutôt l'effet d'une sorte de calcul *instrumental* [Dennett 1987] : c'est un bon moyen de prédire les conduites d'autrui. On peut avoir aussi des raisons de penser que ce sont finalement des contraintes *morales* liées à la volonté de traiter autrui comme un agent autonome susceptible d'être loué ou blâmé qui expliquent cette tendance [1]. En réalité, il existe différentes versions du principe de charité, qui correspondent aux différentes façons d'expliquer notre tendance à interpréter de façon rationnelle nos comportements apparemment irrationnels.

Dans ce qui suit, je n'essaie pas d'apporter d'arguments en faveur de telle ou telle explication. Je veux seulement montrer qu'il ne faut pas exagérer la portée ou les pouvoirs du principe de charité en général. Certains philosophes semblent penser que ce principe pourrait bien être une sorte de remède universel aux maux de la philosophie de l'esprit ou de l'action. Le principe de charité permettrait, entre autres, de résoudre

1. C'est le point de vue que j'essaie de défendre dans *Les Causes et les raisons*, Nîmes, J. Chambon, 1995.

le problème de la sous-détermination observationnelle des attributions d'états mentaux (c'est-à-dire, de répondre à la question : qu'est-ce qui nous autorise à attribuer tels ou tels désirs ou telles ou telles croyances à nous-mêmes ou à autrui ?) ou de donner une signification aux cas centraux d'irrationalité motivée (prendre ses désirs pour des réalités, nier l'évidence, se mentir à soi-même, agir à l'encontre de son meilleur jugement etc.). A mon avis, c'est une erreur. En réalité,

1) Aucune des versions du principe ne permet de résoudre le problème de la sous-détermination observationnelle des attributions d'états mentaux. Le principe permet d'exclure toutes les attributions de croyances qui ne respectent pas le principe, mais il ne permet pas de départager deux attributions de croyances en conflit qui respectent également le principe.

2) Le principe de charité dans ses versions fortes exclut la possibilité de l'irrationalité motivée. Par conséquent, si on croit à la possibilité de l'irrationalité motivée, on ne peut pas endosser une version forte du principe de charité.

C'est, du moins, ce que je vais essayer de montrer.

Sous-détermination et charité

A partir de données vagues, incomplètes, souvent inadéquates, comment faisons-nous pour attribuer raisonnablement (c'est-à-dire de façon suffisamment *justifiée*) des pensées, des croyances, des désirs et, par conséquent, des actions à autrui et à nous-mêmes ? Ce genre de question évoque irrésistiblement la tradition herméneutique. L'herméneute essaie de reconstruire à partir de données fragmentaires le sens d'un texte. Mais il ne faut pas confondre interprétation et herméneutique [Engel,1991]. L'interprétation ne fait pas nécessairement appel à la notion très vague de sens. Elle ne porte pas sur des textes mais sur des actions. Son objectif est plus simple (on pourrait dire, plus naïf) : justifier l'attribution d'états inobservables à partir de données observables. Son principal problème, c'est celui de la *sous-détermination* de ces attributions par les données, c'est-à-dire, plus précisément, de la possibilité qu'existent plusieurs attributions incompatibles entre elles mais toutes parfaitement compatibles avec les données.

Dans toutes les sciences empiriques, différentes théories sont compatibles avec des données d'expérience et incompatibles entre elles. Cependant, les sciences les moins controversées sont celles où existent des moyens de limiter leur prolifération. A leur *périphérie*, nos théories scientifiques sont tout de même contrôlées par la *perception* (tactile, visuelle,

etc.). D'autre part, les capacités de *prédiction* restent un assez bon critère de sélection de ces théories. En revanche, dans le domaine de l'attribution d'états mentaux, il n'existe aucune contrainte de ce genre. Ce sont des états inobservables (si l'on ne donne pas un sens trop large, et creux finalement, au terme "observation"). Et ce qui caractérise nos prédictions de comportements à partir d'états mentaux, c'est ou bien leur vacuité totale (en réalité, ce ne sont pas des prédictions, mais des reconstructions rétrospectives des raisons d'agir à partir des actions effectuées) ou bien leur faillite massive (lorsqu'elles sont d'authentiques prédictions). Tel est, du moins, en gros le point de vue de Quine [Quine 1960]. Naturellement, Quine recommande de renoncer au projet désespéré de construire des sciences à propos de ces états internes que sont les croyances, les désirs, les intentions, etc. Cela ne signifie pas qu'il ne puisse pas exister, pour lui, de sciences humaines ou sociales mais seulement que ces sciences doivent éliminer les moindres résidus de vocabulaire mental si elles veulent sortir de leur état déplorable. Ainsi, d'après Quine, la psychologie béhavioriste va dans le bon sens. En dépit du rejet à peu près général du béhaviorisme au sens strict, les leçons de Quine ont porté. Toutes les théories raisonnables essaient de proposer une solution au problème de la sous-détermination observationnelle des attributions d'états mentaux. Certaines d'entre elles pratiquent encore la chirurgie radicale prescrite par Quine. Le behaviorisme, qui n'a jamais vraiment disparu des laboratoires, retrouve une certaine respectabilité théorique sous la forme de certaines variétés de fonctionnalisme, d'après lesquelles ce qui compte dans la caractérisation de certains états mentaux (tels que la douleur ou la faim), ce sont ses causes et effets typiques plutôt que leurs propriétés qualitatives, telles qu'elles sont éprouvées par l'agent. L'idée que les bonnes sciences sociales ne peuvent se faire qu'en isolant des phénomènes supra-individuels est loin d'être morte et enterrée.

D'un autre côté, si l'espoir de fonder les explications des phénomènes sociaux sur les actions individuelles a résisté, c'est parce que ses promoteurs ont montré (contre Quine) que les attributions d'états mentaux n'étaient peut-être pas radicalement sous-déterminées. Il existerait, dans les sciences intentionnelles, c'est-à-dire celles qui admettent le vocabulaire mental des croyances, des désirs, des pensées, etc., des contraintes aussi sérieuses que les contraintes de prédiction ou de perception dans les sciences naturelles.

On peut les ramener à trois.

1. *Contraintes universelles de type normatif*, relatives au caractère supposé cohérent, rationnel des conduites humaines.

2. *Contraintes universelles de type naturel*, relatives à ce que nous

apprennent nos meilleures théories de l'architecture cognitive de l'espèce humaine.

3. *Contraintes locales de type social ou culturel*, relatives à ce que nous apprennent nos meilleures théories de l'organisation des sociétés humaines.

Prises ensemble, ces contraintes sont-elles suffisantes ? Sont-elles de nature à justifier l'idée que des sciences humaines et sociales non béhavioristes, plus attentives aux mécanismes fins d'explication que les macro-théories, sont possibles ?

Ce sont des questions auxquelles je n'essaierai pas de répondre ici. Je m'intéresserai seulement aux limites du premier ensemble de contraintes, normatives et universalistes, relatives au caractère supposé cohérent, rationnel des conduites humaines. C'est à cet ensemble de contraintes que peut s'appliquer l'expression générique : « principe de charité interprétative ».

Quelques versions du principe de charité

Comme il arrive souvent en philosophie, la discussion autour de l'idée du principe de charité n'a pas abouti à un accord, mais à une sorte de désaccord raisonnable relatif aux multiples possibilités d'interpréter ce principe [Stein 1996], [Delpla 2001]. Ses versions se sont accumulées. Il existe aujourd'hui des versions *a priori* et *a posteriori*, faibles et fortes, étroites et larges, socratiques et aristotéliciennes.

1) *A priori, a posteriori*

Lorsque le principe de charité interprétative est *a priori*, il affirme l'existence d'un lien *conceptuel* entre, d'une part, la possibilité de comprendre les croyances, les désirs et les actions verbales et non verbales de soi-même et d'autrui et, d'autre part, l'attribution à soi-même ou autrui du maximum de rationalité ou du minimum d'irrationalité. Lorsque le principe de charité interprétative est empirique ou *a posteriori*, il se présente plus modestement comme une méthode qui s'est révélée efficace dans la prédiction et la compréhension des actions humaines. Un évolutionniste pourrait suggérer, par exemple, que notre tendance à interpréter les comportements apparemment irrationnels en termes rationnels a toutes sortes de conséquences intéressantes pour la survie de notre espèce. C'est pour cela qu'elle a été sélectionnée. Un pragmatiste pourrait dire, de son côté, que l'attribution du maximum de rationalité ou du minimum d'irrationalité, nous permet de comprendre, expliquer et prédire les actions humaines mieux que tout autre principe connu. Mais ni

l'évolutionniste ni le pragmatiste n'accorderont une valeur absolue à ce principe.

2) *Forte, faible*

La version forte dit qu'il faut toujour*s* adopter ce principe. Ce qui signifie qu'il faut toujours considérer que votre interlocuteur comprend ce qu'il dit, respecte le principe de contradiction, essaie de mettre ses actions en harmonie avec ses désirs et ses croyances, que ses croyances sont vraies en grande partie, qu'elles correspondent aux vôtres et ainsi de suite. Dans la version faible, il est recommandé d'adopter ce principe dans les cas seulement où il n'y a pas de preuves flagrantes du fait qu'il soit inapproprié. Lorsqu'on essaie d'évaluer des théories du type de celles qui attribuent une mentalité prélogique à certains interlocuteurs, le choix entre la version forte ou faible n'est pas sans conséquences. Si on adopte la version forte du principe, on dira que l'hypothèse de la stupidité de l'interlocuteur doit être exclue dans tous les cas. Si on adopte la version faible seulement, on dira que la probabilité d'une interprétation défectueuse est seulement plus élevée que celle de la stupidité de l'interlocuteur.

3) *Étroite, large*

La version étroite nous demande de ne jamais attribuer à notre interlocuteur des énoncés illogiques. La version large nous demande de ne pas attribuer à notre interlocuteur trop de croyances irrationnelles (contradictoires, allant à l'encontre des évidences, etc.).

4) *Socratiques, aristotéliciennes*

Dans la version qu'on peut appeler « socratique » du principe de charité, toutes les formes d'irrationalité sont des manifestations d'ignorance ou d'impuissance. C'est évidemment en référence au fameux adage socratique « Nul n'est méchant intentionnellement, volontairement ou en toute connaissance de cause », que cette version peut être appelée ainsi. Dans la version qu'on peut appeler « aristotélicienne », certains comportements irrationnels, au moins, peuvent être intentionnels, volontaires, accomplis en toute connaissance de cause. C'est, bien sûr, en référence aux objections d'Aristote à l'adage socratique, que cette version peut être appelée ainsi.

Je voudrais insister sur le fait qu'aucune de ces versions ne permet de résoudre le problème de la sous-détermination observationnelle des états mentaux. Certes, le principe de charité interprétative impose des contraintes cruciales à l'interprétation, mais elles sont trop faibles. Le principe permet d'*exclure* toutes les attributions de croyances qui ne respectent pas le principe, mais il ne permet pas de *départager* deux

attributions conflictuelles de croyances qui le respectent également. Rien ne nous interdit de penser que deux interprétations incompatibles entre elles soient compatibles avec le principe de charité. Dans des cas de ce genre, le principe de charité ne peut pas nous dire quelle est la meilleure interprétation [2].

En fait, quelles que soient ses versions (conceptuelle ou empirique, forte ou faible, étroite ou large, socratique ou aristotélicienne), le principe de charité interprétative présente la même exigence, qui est plutôt négative. Il nous demande de mettre en doute (au moins) ou de rejeter (au plus) les interprétations qui ne respectent pas l'idée que nos croyances, désirs, actions sont globalement rationnels, dans différents sens du mot « rationnel » (cohérence, respect des évidences, choix des moyens appropriés etc.). L'utilité de ce principe a été démontrée dans l'examen des thèses de Levy-Bruhl, ou plus exactement, de ce qu'elles sont devenues dans l'histoire des idées [Delpla 2001, 22-27]. Il permet de rejeter la théorie de la mentalité prélogique selon des principes assez clairs.

Les principaux promoteurs du principe de charité semblent parfois endosser des attitudes dépourvues d'ambiguïté par rapport à ses différentes versions [Quine 1960], [Davidson 1991], [Dennett 1990]. Quine paraît favorable à la version étroite (purement logique) du principe de charité et Davidson à la version large (celle qui est relative à l'ensemble des croyances). Dennett semble plutôt instrumentaliste et favorable à une version *a posteriori*, alors que Davidson penche plutôt pour une version *a priori*. Mais, en réalité, tous oscillent entre ces différentes positions. Dans le cas de Davidson, qui est celui auquel je m'intéresse plus particulièrement, on pourrait dire que sa personnalité philosophique est double. Il y a, d'une part, le Davidson qu'on pourrait appeler « maximaliste », adepte d'une version *a priori*, large, forte et socratique du principe de charité et, d'autre part, le Davidson « minimaliste », qui semble plutôt défendre une version *a posteriori*, étroite, faible et aristotélicienne du principe. Le Davidson maximaliste est celui qui devrait, en principe, exclure la possibilité de la faiblesse de la volonté et de l'irrationalité motivée en général. Le Davidson minimaliste est celui qui réussit à montrer comment la faiblesse de la volonté et l'irrationalité motivée en général sont possibles.

2. A ce sujet, je me permets de renvoyer à mon « Philosophie des sciences sociales », dans Jean-Michel Berthelot (éd.), *Épistémologie des sciences sociales*, Paris : PUF, 2000.

Charité et irrationalité motivée

L'application du principe de charité pourrait avoir certaines conséquences indésirables, dans la mesure où elle pourrait aboutir à rendre l'irrationalité inconcevable. En effet, si nous adhérons très strictement au principe, nous aurons le choix entre deux possibilités, lorsque nous serons confrontés à un comportement irrationnel.

Le principe nous demande de vérifier si le comportement apparemment irrationnel n'obéit pas, tout bien considéré, à certaines raisons.

1) Si nous trouvons ces raisons, nous éliminons l'irrationalité. En dépit des apparences, le comportement est rationnel.

2) Si nous échouons, nous sommes contraints de dire que le comportement est privé de raisons, non rationnel, un cas de bêtise ou de folie.

L'application du principe de charité peut donc nous conduire à porter des jugements extrêmes, dont certains (des jugements de bêtise ou de folie) ne sont pas du tout charitables, ce qui est plutôt paradoxal. D'autre part, en ne laissant ouvertes que deux possibilités, l'absolument rationnel et l'absolument non rationnel, le principe de charité semble nous obliger à endosser une attitude réductionniste. Il nous interdit de préserver la spécificité d'un ensemble de comportement bien connus, ni tout à fait rationnels ni tout à fait non rationnels. Je veux parler de nos hésitations exagérées, de nos paresses injustifiées, de nos défaillances inexplicables, de nos infidélités gratuites à nos décisions ou à nos résolutions initiales, de nos facilités à succomber devant les plus faibles tentations, de nos manières curieuses de remettre au lendemain, de retarder, d'atermoyer, d'ajourner ce qui n'est pas forcément inquiétant ou encore de l'infinité des façons que nous avons de nous mentir à nous-mêmes, de prendre nos désirs pour des réalités, de rejeter les évidences ou d'oublier plus ou moins volontairement ce qui ne nous embarrasse même pas.

La question se pose, en fait, de savoir si Davidson veut proposer une justification théorique de notre tendance à trouver des raisons aux conduites humaines quelles qu'elles soient, ou s'il souhaite plutôt montrer que cette tendance nous rend aveugles à cet ensemble de phénomènes. Il me semble que Davidson estime qu'il faut prendre les comportements irrationnels au sérieux, c'est-à-dire, ne pas les réduire, comme nous le faisons habituellement, à des comportements rationnels déguisés ou à des comportements purement pathologiques ou absolument non rationnels. Ce programme est-il compatible avec l'adoption d'une version maximaliste du principe de charité, *a priori*, large, forte et socratique ? Je ne le crois pas. Et Davidson ne le croit probablement pas non plus. Son

analyse de l'irrationalité motivée montre plutôt une tendance à endosser une version minimaliste de ce principe. Mais qu'est-ce que l'irrationalité motivée ?

Il est probable qu'un grand nombre de nos croyances, ou de nos raisonnements spontanés, comme on dit, devraient être déclarés « irrationnels » s'ils étaient strictement évalués selon les critères de la rationalité tels qu'on les trouve exposés dans les ouvrages de méthodologie des sciences causales ou empiriques ou dans les livres de logique élémentaire. Cependant, tant que nos raisonnements équivoques, biaisés, circulaires, inappropriés, nos croyances fausses, contradictoires, injustifiées peuvent être expliqués par des causes contingentes telles que le manque d'attention, l'ignorance, la fatigue, les troubles émotionnels, personne ne semble penser qu'un problème philosophique sérieux puisse se poser à ce propos.

Les difficultés proprement philosophiques semblent apparaître lorsque ces infractions aux normes supposées du raisonnement correct ou des croyances justifiées semblent massives, récurrentes, systématiques et, surtout, lorsque personne ne semble disposé à y renoncer. Comme le dit très bien Amélie Rorty, lorsque des êtres rationnels montrent une résistance tout à fait inattendue à la correction des erreurs, nous avons besoin d'une explication d'un genre tout à fait particulier [Rorty 1988, 216]. Nous voulons savoir non seulement comment cela se produit, mais nous voulons aussi savoir pourquoi cette résistance prend une forme systématique et en quoi *telle* ou *telle* forme particulière d'erreur systématique peut exercer une certaine attraction sur nous.

Ce qu'on peut appeler l'« irrationalité motivée », ce sont ces formes d'irrationalité qui possèdent une structure qui les distingue des erreurs dans les raisonnements ou dans les croyances dues à des causes contingentes ou involontaires [Pears 1982], [Pears 1984]. Elles semblent conscientes, voulues, systématiques. Elles sont soutenues par certains désirs et certaines croyances raisonnables. A la suite de Nisbett et Ross, on peut parler de *théorie chaude* de l'irrationalité motivée lorsque l'irrationalité est soutenue par des désirs et de *théorie froide* de l'irrationalité motivée lorsqu'elle est soutenue par des croyances [Nisbett & Ross 1980].

Dans l'ensemble des formes d'irrationalité motivée, qui sont peut-être innombrables, les philosophes ont pris l'habitude d'examiner quatre cas qu'ils ont fini par juger particulièrement significatifs. Il s'agit de :

— Prendre ses désirs pour des réalités
— Nier l'évidence
— Se mentir à soi-même
— Agir à l'encontre de son meilleur jugement (cas dits d' « incontinence », d'« akrasia » ou de « faiblesse de la volonté »).

La possibilité même de ces formes d'irrationalité motivée a été souvent contestée. Comme Davidson et quelques autres avant lui l'ont souligné, la question n'est pas de savoir si le mensonge à soi-même et la faiblesse de la volonté sont rationnels ou irrationnels mais s'ils peuvent exister, s'ils sont possibles ou concevables [Davidson 1970]. Lorsqu'à la suite de Socrate, certains philosophes prétendent que la faiblesse de la volonté n'existe pas, c'est parce qu'ils jugent qu'il est impossible *pratiquement* que quelqu'un choisisse le pire alors que le meilleur est à sa portée. D'autres philosophes invoquent un argument conceptuel pour nier la possibilité de la faiblesse de la volonté. Ils pensent que, de même qu'il serait absurde d'affirmer « Je sais que c'est vrai mais je ne le crois pas », il est absurde de dire « Je sais que c'est bien mais je ne le fais pas ». C'est en raison du lien *conceptuel* ou de la connexion *interne* entre juger bon, vouloir et faire que la faiblesse de la volonté est inconcevable. Quoi qu'il en soit, les philosophes qui nient la possibilité de la faiblesse de la volonté ne manquent pas.

La possibilité du mensonge à soi-même est, elle aussi, loin d'être unanimement reconnue. Ce qui semble caractériser le mensonge interpersonnel, le mensonge à autrui, c'est qu'à son aboutissement, le menteur croit que p et la personne à qui le menteur a menti croit que non p. En fait, lorsque le mensonge a réussi, si on peut dire, les croyances du menteur et de la personne qui a été trompée sont contradictoires. Transposé au cas du mensonge intrapersonnel (le mensonge à soi-même), c'est la *même* personne qui entretiendrait ces deux croyances contradictoires. Évidemment, dans le cas du mensonge interpersonnel, le fait que la personne à qui l'on ment ne sache pas qu'on lui ment est d'une grande importance. Mentir à quelqu'un qui connaît la vérité, lorsqu'on sait qu'il la connaît, c'est faire toutes sortes de choses : se couvrir de ridicule, accomplir un rituel, tenter sa chance en espérant qu'autrui a oublié la vérité, se moquer d'autrui, compter sur la politesse, la faiblesse, les désirs inavoués, les intérêts d'autrui. Mais ce n'est pas « mentir » à proprement parler. Or, dans le cas du mensonge intrapersonnel, il semble, au début de l'examen du moins, que la personne à qui on ment connaît en principe la vérité puisque c'est la *même* personne. Ainsi, le mensonge à soi-même serait une forme d'irrationalité motivée en ce sens que la personne qui se mentirait à elle-même entretiendrait des croyances contradictoires et que, placée dans la situation hypothétique où elle se rendrait compte de cet état de choses, elle persisterait dans ses croyances sans chercher à éliminer l'une ou l'autre d'entre elles [Davidson 1991, 45-61]. Mais le mensonge à soi-même caractérisé de cette façon risque de paraître aussi paradoxal, aussi inconcevable, que la faiblesse de la volonté. Il semble en effet qu'on

ne puisse pas attribuer de croyances contradictoires conscientes à une personne en raison de la signification même du terme « croyance » et de sa relation intrinsèque à l'idée de vérité, ou en raison des postulats très généraux qui gouvernent l'attribution des croyances à une personne.

Il semble toutefois que ces réserves habituelles ne nous empêchent nullement de considérer que l'idée d'un mensonge à soi-même, aussi paradoxale soit-elle, n'est pas entièrement farfelue et que l'ambition de donner une signification à l'irrationalité en général, sans la *réduire* au rationnel ou au non rationnel est parfaitement légitime. Habituellement, les philosophes optent pour la réduction. Dans le cas du mensonge à soi-même, ils proposent deux stratégies réductionnistes.

1) La première consiste à nier qu'il y ait un *authentique* conflit de croyances. Il s'agit plutôt d'un conflit entre des pensées passagères et des croyances robustes ou entre des croyances irréfléchies, formées spontanément, et des croyances réfléchies, acceptées etc. [Bach 1981], [Cohen 1995].

2) La seconde admet qu'il y a bel et bien un conflit de croyances, mais nie qu'il s'agisse d'un conflit de croyances entretenues par la *même* personne ou par une personne intégrée, unifiée, ou d'un conflit de croyances entretenues *simultanément.* Cette seconde stratégie aboutit ou bien à faire l'hypothèse de la *division de l'esprit* ou bien celle de la *distribution temporelle des croyances*[3].

Ces stratégies pourraient être justifiées par la version maximaliste du principe de charité, celle qui dit : pour toute action ou toute croyance apparemment irrationnelle, il existe, *a priori*, une interprétation acceptable qui nous autorise à placer cette action ou cette croyance dans un tableau rationnel. Et de fait, il semble bien qu'il soit toujours possible de soutenir que l'agent qui se ment à lui même n'entretient pas *vraiment* des croyances contradictoires en toute conscience, et qu'il n'est donc pas *vraiment* irrationnel.

Mais ce n'est pas du tout la stratégie qu'adopte Davidson. Ce qui la distingue de toutes celles que je viens d'évoquer, c'est qu'elle a pour ambition *de ne pas affaiblir le paradoxe.* En réalité, les stratégies qui consistent à diviser l'esprit purement et simplement ou à distribuer temporellement les croyances aboutissent à éliminer l'irrationalité du mensonge à soi-même, ce qui devrait suffire à les disqualifier, comme le sou-

3. Jon Elster, *Ulysses and the Sirens*, Cambridge : Cambridge University Press, 1984, et introduction à *The Multiple Self*, Cambridge : Cambridge University Press, 1986, qui contient aussi l'essai de Jon Elster « Deception and Self-deception in Stendhal », p. 93-113 ainsi qu'un autre essai de David Pears, « The Goals and Strategies of Self-Deception », p. 59-77.

ligne Davidson. Il me semble d'ailleurs qu'on n'appréciera jamais assez la contribution de Davidson à la clarification du problème de l'irrationalité, pour des raisons qu'il a lui-même très bien exposées.

Contrairement à ce qu'on pourrait avoir tendance à penser, la justification philosophique de l'idée d'irrationalité n'est pas donnée d'avance. Chaque fois que, dans la vie courante, nous sommes confrontés à un cas de prétendue irrationalité, nous avons tendance à en proposer une description réductrice : l'irrationalité supposée n'était qu'un cas de rationalité déguisée, ou bien d'absence totale de rationalité. Cette tendance affecte certaines théories philosophiques ou psychologiques. Davidson pense que c'est cette tendance qui nous rend tellement sensibles à cet aspect de la méthodologie freudienne qui consiste à élargir l'éventail des phénomènes accessibles à l'explication 21579641rationnelle [Davidson 1991, 24-25]21579641. Cependant, nous faisons peut-être preuve de faiblesse en tant que philosophes si nous ne reconnaissons pas que la véritable difficulté, lorsque nous examinons l'idée d'irrationalité, consiste précisément à ne pas conclure qu'elle n'est, au fond, qu'une forme déguisée de rationalité ou d'absence totale de rationalité. La véritable difficulté, c'est d'essayer de *penser l'irrationalité en tant que telle*, sans la réduire au rationnel ou au non rationnel. Je ne crois pas qu'il soit possible d'exposer cette objection plus clairement ou plus brièvement que Davidson lui-même lorsqu'il dit :

> Le paradoxe sous-jacent de l'irrationalité, qu'aucune théorie ne peut vraiment éviter, est le suivant : si nous l'expliquons trop bien, nous en faisons une forme déguisée de rationalité ; alors que si nous attribuons l'incohérence avec trop de désinvolture, nous ne faisons que compromettre notre capacité à diagnostiquer l'irrationalité en retirant l'arrière-plan de rationalité requis pour justifier un diagnostic quelconque [Davidson 1991, 40].

Davidson semble nous dire : « Soyez charitables, mais pas trop, car sinon vous ne pourrez pas sauver le phénomène de l'irrationalité motivée ».

C'est la thèse socratique affirmant la souveraineté de la raison qui semble exclure la possibilité du phénomène. En effet, dire qu'une action est irrationnelle (par opposition à non rationnelle), c'est affirmer que la raison a fait défaut alors qu'elle était présente. Ce qui est difficilement concevable, on ne peut pas le nier. Voici comment Davidson présente les choses.

> La notion d'action, de croyance, d'intention, d'inférence ou d'émotion irrationnelle est paradoxale. Car ce qui est irrationnel n'est pas simplement ce qui est rationnel ou ce qui se tient hors des limites du rationnel : l'irrationalité est un échec au domicile de la raison elle-même. Quand Hobbes dit que

> seul « l'homme a le privilège de l'absurdité », il suggère que seul un être rationnel peut être irrationnel [...]. Le paradoxe de l'irrationalité n'est pas aussi simple que le paradoxe apparent contenu dans la notion d'une plaisanterie ratée ou celle d'un objet d'art de mauvaise qualité. Le paradoxe de l'irrationalité a sa source dans ce qui fait partie de nos manières les plus fondamentales de décrire, de comprendre et d'expliquer les états et les événements psychologiques [Davidson 1991, 21].

On pourrait dire que, dans le raisonnement de Davidson, la thèse socratique n'est rien d'autre qu'une version très forte du principe de charité.

A première vue, Davidson soutient que cette version forte n'exclut pas l'irrationalité motivée. Il semble même affirmer que c'est elle qui permet de la penser. Pourquoi ? Eh bien, pour Davidson, de même qu'il ne peut pas y avoir de malentendus ou d'incompréhension si nous déchirons complètement la toile de fond de la compréhension, de même il ne peut y avoir d'agent irrationnel si nous l'avons entièrement dépouillé de ses attributs rationnels. On ne dit jamais des tulipes ou des chevaux qu'ils sont irrationnels pour la bonne raison que nous ne pensons pas qu'il soit intéressant de juger qu'ils sont rationnels. Ce n'est que pour autant que nous traitons autrui comme être rationnel que nous pouvons envisager son irrationalité. Si nous cessons de le percevoir comme un être rationnel, nous perdons, *en même temps*, la capacité de le percevoir comme irrationnel. Et c'est pourquoi Davidson estime qu'il existe un authentique *paradoxe de l'irrationalité.*

Peut-on être logique sans être rationnel ?

Davidson examine quatre formes d'irrationalité. Sa stratégie est identique dans chaque cas : identifier le noyau inaltérable de rationalité, sur le fond duquel l'irrationalité peut nous apparaître.

1. *Prendre ses désirs pour des réalités.* Pour saisir ce qu'il y a d'irrationnel dans le fait de prendre ses désirs pour des réalités, il faut distinguer l'*explication causale* des croyances et leur *justification.* Supposons qu'étant chevelu, je croie néanmoins que je suis chauve. On me demande : « Mais pourquoi crois-tu que tu es chauve ? ». Et je réponds : « Parce que je *désire* sincèrement être chauve ». En tant qu'explication causale de ma croyance, la réponse est satisfaisante, même si elle n'est pas tout à fait conventionnelle (les cas d'individus qui désirent sincèrement être chauves étant, à ma connaissance, plutôt rares). Mais en tant que *justification rationnelle* de ma croyance, ma réponse est inappropriée. On peut donc dire que, du point de vue épistémologique, prendre ses désirs pour des réalités est irrationnel parce que c'est une tentative de justifier une

croyance par ce qui ne peut pas la justifier. Ce qui pourrait justifier ma croyance que je suis chauve, ce n'est pas le désir que j'ai de l'être mais certains témoignages des sens (qui, en l'occurrence, devraient m'inciter à une toute autre conclusion).

2. *Se mentir à soi-même.* Ce qui veut dire, du point de vue épistémologique, expliquer la formation d'une croyance à partir d'une autre croyance qui, en fait, la contredit : « Je crois que je suis chauve *parce que* je ne suis pas chauve ». L'irrationalité est aussi évidente que dans le premier cas et sa source est identique. Ce qui explique causalement ma croyance que je suis chauve (le fait que je croie que je ne suis pas chauve) ne peut pas servir à *justifier* rationnellement ma croyance.

3. *Nier l'évidence.* C'est adopter une conclusion contraire à toutes les données, en dépit du fait que rien ne nous interdit d'avoir accès à ces données ou même en dépit du fait que nous avons eu accès à ces données. Davidson évoque une mésaventure qui lui est arrivée au cours d'un safari. Son guide portait une jupe, avait une voix haut placée et s'appelait Hélène, mais il s'obstinait à en parler comme s'il s'agissait d'un homme ou, plutôt, il s'étonnait qu'un homme puisse s'appeler Hélène, etc. Ce cas s'apparente formellement au suivant comme nous allons le constater.

4. *Agir à l'encontre de son meilleur jugement.* C'est faire quelque chose pour une raison, tout en ayant une meilleure raison de ne pas le faire. C'est, en quelque sorte, une définition formelle de l'*akrasia* ou de la faiblesse de la volonté, qu'on peut illustrer de la façon suivante. J'ai une bonne raison de boire un litre de whisky (me calmer avant de me présenter à l'épreuve de conduite pour obtenir mon permis) tout en ayant une meilleure raison de m'abstenir de boire (ne pas être en état d'ivresse au moment de l'examen). Et je choisis, sans y être contraint par une compulsion interne, de boire un litre de whisky avant d'aller à l'examen. Si on exclut les objections habituelles (en réalité, je ne voulais pas réussir l'examen ou mon angoisse était telle qu'il est absurde de dire que j'ai *choisi* de boire, etc.), c'est un cas complexe d'irrationalité et non d'absence pathologique de rationalité. En effet, faire quelque chose pour une raison, tout en ayant une autre raison *meilleure* de ne pas le faire, n'est pas plus grave que ne pas tenir compte de toute l'information dont on dispose lorsqu'on essaie d'expliquer un phénomène. C'est ce qui fait, d'ailleurs, que ce genre d'irrationalité s'apparente formellement à la négation de l'évidence. Bien qu'il y ait des différences entre les deux cas (dans le second, c'est une action qui est mise en relation avec des croyances, dans le premier, ce sont des perceptions), on peut dire qu'à chaque fois, certaines *règles d'induction* sont violées.

Chacun de ces comportements apparemment irrationnels met en relation des ensembles de phénomènes différents. Quand on prend ses désirs pour des réalités, le trouble vient de la relation défectueuse entre croyances et désirs ; lorsqu'on se ment à soi-même, ce sont les croyances entre elles qui sont curieusement associées. Lorsqu'on nie l'évidence, le problème semble se situer au plan de la relation entre croyance et perception. Lorsqu'on agit à l'encontre de son meilleur jugement, c'est entre les croyances, les désirs et les actions qu'un problème semble se poser. Davidson éclaire les quatre cas de la même façon. Prendre ses désirs pour des réalités, se duper soi-même, nier l'évidence, agir à l'encontre de son meilleur jugement, c'est être *déraisonnable sans être illogique.* Plus précisément, il montre que, dans chaque cas, on viole des règles *d'induction* ou des règles de justification des croyances mais qu'en aucun cas on ne transgresse le principe logique de non contradiction. Mais en adoptant cette stratégie d'explication, qui se contente de sauver la *logique* de l'agent apparemment irrationnel, Davidson passe de la version forte, large, socratique du principe de charité à la version faible, étroite, aristotélicienne, défendue par Quine.

Comment sauver la zone grise des comportements humains ?

La version forte du principe de charité. nous laisse devant l'alternative suivante : ou bien absolument rationnel ou bien absolument non rationnel. Le défaut de cette version, c'est qu'elle exclut la possibilité de l'existence de cette zone grise des comportements humains, qui comprend ce qui n'est ni tout à fait rationnel ni tout à fait irrationnel : ce qu'on appelle l'« irrationalité motivée ». Par conséquent, si on croit en la possibilité de l'irrationalité motivée, on ne peut pas endosser une version forte du principe de charité.

Les philosophes qui, à la manière de Davidson, croient à la fois au principe de charité et à la possibilité de l'irrationalité motivée, risquent d'être prisonniers d'une contradiction s'ils endossent une version forte du principe ou de la croyance. Pour éviter la contradiction, ils doivent ou bien affaiblir le principe de charité ou bien leur croyance en la possibilité de l'irrationalité motivée.

Davidson propose une version forte de l'irrationalité motivée (elle admet, par exemple, qu'il est possible d'agir intentionnellement, en toute conscience, contre son meilleur jugement). Il n'est donc pas très étonnant qu'il soit conduit à endosser une version faible du principe de charité.

Mais la question se pose de savoir si endosser une version faible du principe de charité ne revient pas à reconnaître les limites du principe ou, ce qui est plus embarrassant encore, à le priver de tout pouvoir explicatif.

Bibliographie

Bach, Kent

1981 *An Analysis of Self-deception*, Philosophy and Phenomenological Research, 41, 1981 : 351-370.

Boudon, Raymond

1995 *Le Juste et le vrai,* Paris : Fayard.

Cohen, L. Jonathan

1995 *An Essay on Belief and Acceptance*, Oxford : Clarendon Press.

Davidson, Donald

1970 *How is Weakness of the Will Possible ?*, in Joel Feinberg (éd.), Moral Concepts, Oxford Readings in Philosophy, Oxford : Oxford University Press, 1970 ; 93-113. Cité d'après la traduction française de Pascal Engel : *Comment la faiblesse de la volonté est-elle possible ?*, in Actions et événements, Paris : PUF, 1993 ; 37-65.

1982 *Paradoxes of Irrationality*, in Richard Wollheim & James Hopkins (éds.), Philosophical Essays on Freud, Cambridge : Cambridge University Press, 1982 ; 289-305. Cité d'après la traduction française par Pascal Engel : *Paradoxes de l'irrationalité*, [Davidson 1991] ; 21-43.

1982 *Rational Animals*, Dialectica, 36 : 318-327. Cité d'après la traduction française par Pascal Engel : Animaux rationnels, *in* [Davidson 1991] ; 63-75.

1985 *Deception and Division*, in Ernest Le Pore & Brian McLaughlin (éds.), *Actions and Events. Perspectives on the Philosophy of Donald Davidson*, Oxford : Blackwell, 1985 ; 138-148. Cité d'après la traduction française par Pascal Engel : Duperie et division, *in* [Davidson 1991] ; 45-61.

1991 *Paradoxes de l'irrationalité*, traduction française par Pascal Engel de trois articles de D. Davidson, Combas : L'éclat, 1991.

Delpla, Isabelle

2001 *Quine, Davidson. Le principe de charité*, Paris : PUF, 2001.

Dennett, Daniel

1987 *The Intentional Stance*, Cambridge, Massahusetts : M.I.T. Press, 1987. Cité d'après la traduction de Pascal Engel : La *Stratégie de l'interprète*, Paris : Gallimard, 1990.

Elster, John

1984 *Ulysses and the Sirens*, Cambridge : Cambridge University Press.

1986a *Introduction*, in J. Elster (éd.), *The Multiple Self*, Cambridge : Cambridge University Press.

1986b *Deception and Self-deception in Stendhal*, in J. Elster (éd.), *The Multiple Self*, Cambridge : Cambridge University Press, 93-113.

Engel, Pascal

1991 *Interpretation without Hermeneutics. A Plea Against Ecumenism" ?* Topoï, 10, 1991, 137-146.

Nisbett, R. & Ross, L.

1980 *Human Inference, Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, New Jersey : Prentice Hall.

Ogien, Ruwen

1995 *Les Causes et les raisons*, Nîmes : J. Chambon.

2001 *Philosophie des sciences sociales*, in Jean-Michel Berthelot (éd.), Épistémologie des sciences sociales, Paris : PUF.

Pears, David.

1982 *Motivated Irrationality, Freudian Theory and Cognitive Dissonance*, in R. Wollheim et J. Hopkins (éds.), Philosophical Essays on Freud, Cambridge : Cambridge University Press, 264-288.

1984 *Motivated Irrationality*, Oxford : Clarendon Press.

1986 *The Goals and Strategies of Self-Deception*, in J. Elster (éd.), The Multiple Self, Cambridge : Cambridge University Press, 59-77.

Quine, W.V.O.

1960 *Word and Object*, Cambridge, Massahusetts : M.I.T. Press. Cité d'après la traduction de Paul Gochet : *Le Mot et la chose*, Paris : Flammarion, 1977.

Rorty, Amélie Oksenberg

1988 *Mind in Action : Essays in the Philosophy of Mind*, Boston : Beacon Press.

Stein, Edward

1996 *Without Good Reason*, Oxford : Clarendon Press.