Principe de charité et sciences de l'homme[1]

Denis Bonnay

(Université Paris Ouest-Nanterre-La Défense, Ireph & IHPST – DEC, ENS Ulm)

&

Mikaël Cozic

(IJN & DEC, ENS Ulm)

Résumé : Selon le principe de charité, nous devons toujours faire l'hypothèse qu'autrui est rationnel lorsque nous cherchons à interpréter ses comportements. Le présent article est consacré à une discussion critique de ce principe et de ses conséquences épistémologiques, à la lumière des récents apports des sciences cognitives concernant les mécanismes de la mentalisation. Nous soutenons que la validité du principe de charité est limitée par l'existence probable de mécanismes d'interprétation non-théoriques (théorie de la simulation) et que, par voie de conséquence, les théories de la rationalité, les théories scientifiques de la cognition et du comportement et les produits de la mentalisation ne s'identifient ni en principe, ni en fait.

Abstract : According to the principle of charity, we always have to make the assumption that other people are rational when we try to interpret their behavior. This paper is a critical discussion of the principle and its methodological consequences in the light of recent advances in cognitive science concerning mentalization. We argue that the validity of the charity principle is limited by the likely existence of interpretative non-theoretical mechanisms (simulation theory). As a consequence, we claim that theories of rationality, scientific theories of cognition and behavior and products of mentalization can neither in fact nor in principle be identified with one another.

---

[1] A paraître dans Th. Martin (ed.), *La scientificité des sciences de l'homme*, Paris : Vuibert, 2009.

Mots-clés : rationalité, interprétation, principe de charité, simulation, logique, théorie de la décision

Key Words : rationality, interpretation, charity principle, simulation, logic, decision theory

1 Introduction

On peut étudier l'homme à travers les contingences de son inscription dans un temps, dans un espace, dans certaines formes d'organisation sociales. C'est ce que font chacune à leur manière l'histoire, la géographie humaine ou la sociologie. On peut également s'intéresser à l'homme en tant qu'animal rationnel pour étudier la nature de la rationalité elle-même. De ce point de vue, une évolution marquante des sciences de l'homme a été de voir différentes disciplines concourir à une élaboration systématique, en général formelle, de *modèles de rationalité*. La logique déductive a indéniablement ouvert la voie à la fin du XIXème siècle, mais l'élaboration des principes de rationalité va bien au-delà : elle inclut la théorie bayésienne des probabilités, la logique inductive, la théorie de la décision, la théorie de la révision des croyances, etc.

Ces modèles de rationalité articulent des *contraintes de rationalité*. Par « contraintes de rationalité », nous entendons des règles générales que doivent respecter les états mentaux et les comportements d'un être rationnel. Parmi les contraintes de rationalité figurent d'abord des contraintes pour la rationalité *théorique,* comme celles que l'on peut trouver en logique déductive ou dans la théorie bayésienne des probabilités. Considérons un individu rationnel, Kurt. Voici deux règles plausibles pour Kurt :

si Kurt croit que A et croit que si A alors B, alors il ne croit pas que non B[2]

si Kurt estime probable que A, alors il n'estime pas probable que non A

Il s'agit bien de contraintes de rationalité théorique, qui énoncent des règles que les croyances de Kurt doivent respecter pour que Kurt puisse être considéré comme un individu rationnel. A côté des règles pour la rationalité théorique, on trouve également des contraintes de rationalité *pratique* comme celles qui sont formulées en théorie de la décision. L'un des piliers de cette dernière est, par exemple, le principe de transitivité des préférences selon lequel

si Kurt préfère x à y et y à z, alors il préfère x à z

Il s'agit bien d'une contrainte de rationalité pratique, qui énonce une règle que les préférences de Kurt doivent respecter pour que Kurt puisse être considéré comme un agent rationnel.

Un autre fait marquant de la réflexion contemporaine sur la rationalité concerne le *statut* des contraintes de rationalité que nous venons de décrire. Selon une idée très répandue, au moins depuis les travaux de Quine, Davidson et Dennett, elles conditionnent la possibilité de comprendre des comportements humains, aussi bien linguistiques que non-linguistiques. Cette idée est connue sous le nom de *principe de charité* ; le principe de charité veut que les

---

[2] Nous simplifions évidemment : la logique déductive ne parle pas directement de croyances : il y est question de relations logiques entre énoncés. Pour aboutir à la contrainte que nous évoquons, qui découle de la validité du *modus ponens*, il faut supposer que les croyances d'une créature rationnelle doivent refléter les relations logiques qui existent entre leurs contenus.

contraintes de rationalité ne nous disent pas simplement quelles règles devraient idéalement respecter les croyances et les désirs des gens, mais qu'elles constituent également des hypothèses qu'il est nécessaire de faire pour pouvoir interpréter un individu comme ayant des croyances ou des désirs.

Parlons d'*hypothèse de rationalité* pour désigner l'hypothèse selon laquelle les états mentaux (croyances et désirs, en particulier) et les comportements des individus satisfont les contraintes de rationalité. Adopter le principe de charité, c'est considérer qu'il faut (toujours) faire l'hypothèse de rationalité, mais il y a bien sûr autant de versions de l'hypothèse de rationalité que de contenus donnés aux contraintes de rationalité. Il peut y avoir des versions plus faibles que d'autres – que l'on compare une version qui n'inclut que des règles tirées de la logique déductive à une autre qui inclut en outre des règles portant sur les probabilités subjectives. Il peut également y avoir des versions incompatibles avec d'autres parce que reposant sur des principes de rationalité divergents. Dans ce qui suit, nous ferons cependant comme s'il existait *une* version de l'hypothèse de rationalité qui inclurait principalement (1) les règles issues de la logique déductive classique, (2) les règles issues de la théorie bayésienne des probabilités subjectives et (3) les règles issues de la théorie bayésienne de la décision.

D'un point de vue épistémologique, on voit immédiatement l'intérêt heuristique de l'hypothèse de rationalité : elle permet d'asseoir les sciences de l'homme sur des modèles – les modèles de rationalité. L'économie contemporaine, qui repose essentiellement sur la théorie de la décision et la théorie des jeux, illustre ceci de manière spectaculaire, en faisant reposer ses explications des comportements individuels et collectifs sur des hypothèses fortes de rationalité. Bien que cette justification pragmatique de l'hypothèse de rationalité n'entre pas dans les justifications traditionnelles du principe de charité, Davidson notamment souligne que ce principe introduit une différence fondamentale entre les « sciences de la nature » et les « sciences de l'homme », en déterminant des conditions que doivent satisfaire en tant que tels les objets d'une science de l'homme, à savoir les états mentaux et les comportements intentionnels[3]. Inversement, le principe de charité met peut-être les disciplines qui construisent les modèles de rationalité dans une position trop confortable : que valent ces modèles, si leur portée n'est ni exactement normative, ni exactement descriptive, et s'ils ne peuvent être falsifiés ? Et surtout que penser des nombreuses données apportées par la psychologie expérimentale qui semblent mettre en évidence des déviations systématiques du comportement des individus par rapport aux normes de rationalité ?

Dans ce qui suit, nous nous proposons d'examiner les conditions d'une possible remise en cause du principe de charité et les conséquences épistémologiques de cette remise en cause. Nous allons commencer par exposer plus en détail le principe lui-même et ses justifications, en revenant sur la démarche quasi-transcendantale qui fait du principe de charité une *condition de possibilité* des sciences de l'homme. En examinant de manière critique ces justifications, nous verrons ensuite que le lien entre charité et compréhension n'est peut-être pas aussi indissoluble qu'il n'y paraît, et que les enseignements des sciences

[3] Ainsi la charité conduit à une forme d'anti-naturalisme : la charité ne vaut bien sûr pas pour les sciences de la nature, mais elle vaut pour les sciences de l'homme, donc celles-ci ne sont pas réductibles aux sciences de la nature.

cognitives – les théories empiriques de la mentalisation[4] – suggèrent d'autres voies pour la compréhension et l'interprétation des états mentaux et des comportements intentionnels. Au plan épistémologique, nous soutiendrons pour finir que théorie naïve, théorie descriptive et théorie normative sont trois types distincts de théories du mental qui ne doivent pas être confondus, contrairement à ce que suggèrent les tenants du principe de charité.

2 Le principe de charité

2.1 Charité, traduction et interprétation

Le principe de charité est devenu fameux en philosophie à travers son utilisation par Quine[5] dans *Word and Object* (1960). Quine l'invoque dans son analyse de la *traduction radicale*, c'est-à-dire de « la tâche du linguiste qui, sans pouvoir être aidé par un interprète, entreprend de pénétrer et de traduire un langage jusqu'alors inconnu » (1960, tr. fr. pp. 59-60). Voici ce que dit Quine dans le passage fameux consacré à la traduction des connecteurs logiques :

> *«…supposons qu'on prétende que certains indigènes sont disposés à accepter comme vraies certaines phrases traduisibles dans la forme* 'p *et non*-p'. *Cette supposition est absurde au regard de nos critères sémantiques… Une traduction malicieuse peut rendre les locutions indigènes aussi étranges que l'on veut. Une meilleure traduction leur imposera notre logique…*
>
> *La maxime de traduction qui est à la base de tout ceci, c'est qu'il est probable que les assertions manifestement fausses à simple vue fassent jouer des différences cachées de langage... La vérité de bon sens qu'il y a derrière cette maxime, c'est que la stupidité de notre interlocuteur, au-delà d'un certain point, est moins probable qu'une mauvaise traduction... »* (1960, §13)

La portée du principe de charité quinien ne se réduit ni à l'élimination des contradictions logiques ni à la situation du linguiste perdu dans la jungle. En effet, premièrement, la préservation de la logique par la traduction n'est qu'un cas particulier d'une maxime générale – « sauver ce qui est obvie » (Quine, 1970, tr. fr. p.123) – qui nous enjoint de choisir un manuel de traduction tel que les vérités évidentes de la langue étrangère soient traduites par des énoncés vrais, si possible également évidents, de notre langue. Deuxièmement, la traduction radicale commence à la maison, comme aime à le dire Quine : nous l'appliquons tout aussi bien lorsqu'il s'agit de « traduire » ce que dit un locuteur du français qui répond « oui et non » à une question à laquelle il était supposé répondre par oui ou par non.

Mais dans quelle mesure ce principe s'impose-t-il ? Avant tout, il faut bien comprendre qu'appliquer le principe de charité, ce n'est pas faire acte de charité : on ne

---

[4] Nous entendons par mentalisation l'attribution d'états mentaux à autrui, et utilisons le terme comme traduction de l'expression anglaise « *mind reading* ».

[5] Quine lui-même attribue le principe de charité à Wilson (1959).

suppose pas que les indigènes sont logiquement cohérents, comme on supposait que les sauvages sont naturellement bons. Comme le dit Quine, la stupidité est moins probable que la mauvaise traduction ; c'est une question de « *common sense* ». Et le principe de charité joue alors comme un critère négatif, en nous enjoignant de ne pas adopter un manuel de traduction qui rendrait les locuteurs indigènes trop évidemment irrationnels. On pourrait certes aller contre le sens commun, mais ce serait pure perversité. Davidson a tenté d'aller plus loin que cette justification par défaut[6], en mettant le principe de charité au centre de son traitement de l'*interprétation radicale* (Davidson, 1973 et 1974a). Davidson fait du principe de charité un principe positif qui guide toute tentative de solution au problème de l'interprétation radicale.

Le problème qu'il s'agit de résoudre est celui de savoir comment, en principe, il est possible d'interpréter le langage d'un locuteur-cible (disons, Kurt) sans avoir de connaissance préalable ni de ce langage, ni des attitudes propositionnelles du locuteur-cible. Les *données de base* sont dans ce cas les énoncés que le locuteur-cible tient pour vrais à un certain moment dans certaines circonstances. Elles ont donc la forme suivante :

Kurt tient pour vrai l'énoncé « p » au moment t dans les circonstances c

Kurt tient pour vrai l'énoncé « p' » au moment t' dans les circonstances c'

…

De ces données de base, on peut inductivement extraire des généralisations qui auront la forme suivante :

Kurt tient pour vrai l'énoncé « p » ssi c

Si Kurt est un locuteur de l'allemand, on aurait ainsi par exemple :

Kurt tient pour vrai l'énoncé « Es regnet » ssi il pleut

Un tel énoncé est un des éléments qui permet de construire une théorie donnant les conditions de vérité des phrases de l'allemand, puisque l'interprétation de ce que disent Kurt et les autres membres de sa communauté linguistique se fait, selon la méthode préférée par Davidson, par la construction d'une théorie de la vérité pour cette langue.

C'est donc dans la construction de cette théorie qu'intervient le principe de charité. Ses formulations sont passablement flottantes, mais voici l'une d'entre elles :

> *« We want a method that satisfies the formal constraints on a theory of truth, and that maximizes agreement, in the sense of making Kurt (and others) right, as far as we can tell, as often as possible. »* (Davidson, 1973)

---

[6] Nous ne prétendons pas ici traiter de la différence entre le statut du principe de charité chez Quine et Davidson. Voir Laugier (1992, p. 43 *sqq*) pour une interprétation soulignant l'écart entre les deux auteurs et les nombreux travaux de P. Engel (notamment Engel, 1994).

Selon cette formulation[7], le principe de charité nous recommande de faire l'hypothèse selon laquelle les croyances du locuteur-cible sont correctes : Kurt croit que p si et seulement si p.

2.2 Principe de correspondance et principe de cohérence

Donner raison à Kurt peut prendre plusieurs aspects, de sorte que le principe de charité de Davidson a en réalité en deux composantes, l'une qui a trait à la vérité, l'autre qui a trait au respect des normes logiques. Davidson distingue ces deux aspects en appelant le premier le « principe de correspondance » et le second le « principe de cohérence » (voir Davidson, 1985, p. 92). Le principe de correspondance applique la charité aux énoncés considérés individuellement : ce que nous tenons pour évidemment vrai doit être en général également vrai pour Kurt. Le principe de cohérence applique la charité aux relations entre les énoncés : les normes que nous appliquons lorsqu'il s'agit d'exclure des énoncés incompatibles ou de tirer d'un ensemble d'énoncés les inférences qui doivent en être tirées doivent également être appliquées par Kurt. L'interprétation apparaît alors comme un processus complexe d'ajustements, qui fait intervenir l'hypothèse de rationalité à la fois dans la détermination des données de base – les croyances élémentaires de Kurt qui donnent accès aux équivalences T ou la traduction des connecteurs logiques – et dans l'évaluation du résultat.

De plus, la détermination de ce que veut dire Kurt n'est qu'une partie du problème lorsque l'on s'intéresse au comportement linguistique *et* non linguistique de Kurt. La compréhension et l'explication de l'ensemble du comportement requiert la détermination des croyances et des désirs de l'agent. En conséquence, Davidson inclut au titre du principe de cohérence les normes du raisonnement pratique – la théorie bayésienne de la décision. Le domaine d'application du principe de charité est alors bien en général celui du *mental*, i.e. de l'attribution d'états mentaux, de l'explication par des états mentaux (de comportements ou d'autres états mentaux) et de la prédiction à partir d'états mentaux (de comportements ou d'autres états mentaux)[8].

On aboutit à ce que l'on pourrait appeler le *principe de charité universel* selon lequel l'hypothèse de rationalité doit valoir pour tout être envisagé du point de vue mental et pour tout type de rationalité – qu'il s'agisse de rationalité théorique ou de rationalité pratique. La partie supérieure de la FIGURE 1 ci-dessous illustre l'esquisse que nous venons de tracer. Le *principe de charité universel* correspond à l'ensemble des domaines représentés tandis que la version plus restreinte qui est invoquée dans la théorie de l'interprétation radicale – la construction d'une théorie de la vérité pour le langage de Kurt – correspond aux domaines coloriés. La partie inférieure de la FIGURE 1 relie les trois modèles de rationalité évoqués dans l'introduction aux trois principaux domaines du principe de cohérence. On notera qu'il n'y pas de telles contreparties pour le principe de correspondance. On remarquera également l'existence d'un domaine axiologique pour ce dernier principe : l'attribution de désirs (ou de valeurs) « corrects » est en effet invoquée par Lewis (1974), Davidson (1985) ou encore

---

[7] Sur les différentes formulations du principe de charité, voir Ludwig (2004, p. 353) et Lepore & Ludwig (2005, pp. 185 *sqq*).

[8] Cela signifie en particulier que le principe de charité ne vaut pas seulement dans le contexte de l'interprétation (radicale ou non), qui ne concerne que la détermination des significations et des croyances. Nous verrons dans la section suivante les justifications du principe de charité, dans le contexte de l'interprétation mais aussi s'agissant d'autres états mentaux que les croyances et d'autres comportements que les comportements linguistiques.

Dennett (1987). S'agissant de Davidson, il n'est cependant pas évident de cerner les motivations qui le font inclure le domaine axiologique dans la mesure où il ne semble requis ni dans l'élaboration de la théorie de l'interprétation ni dans la théorie de la décision qui sont les deux théories que Davidson considère.

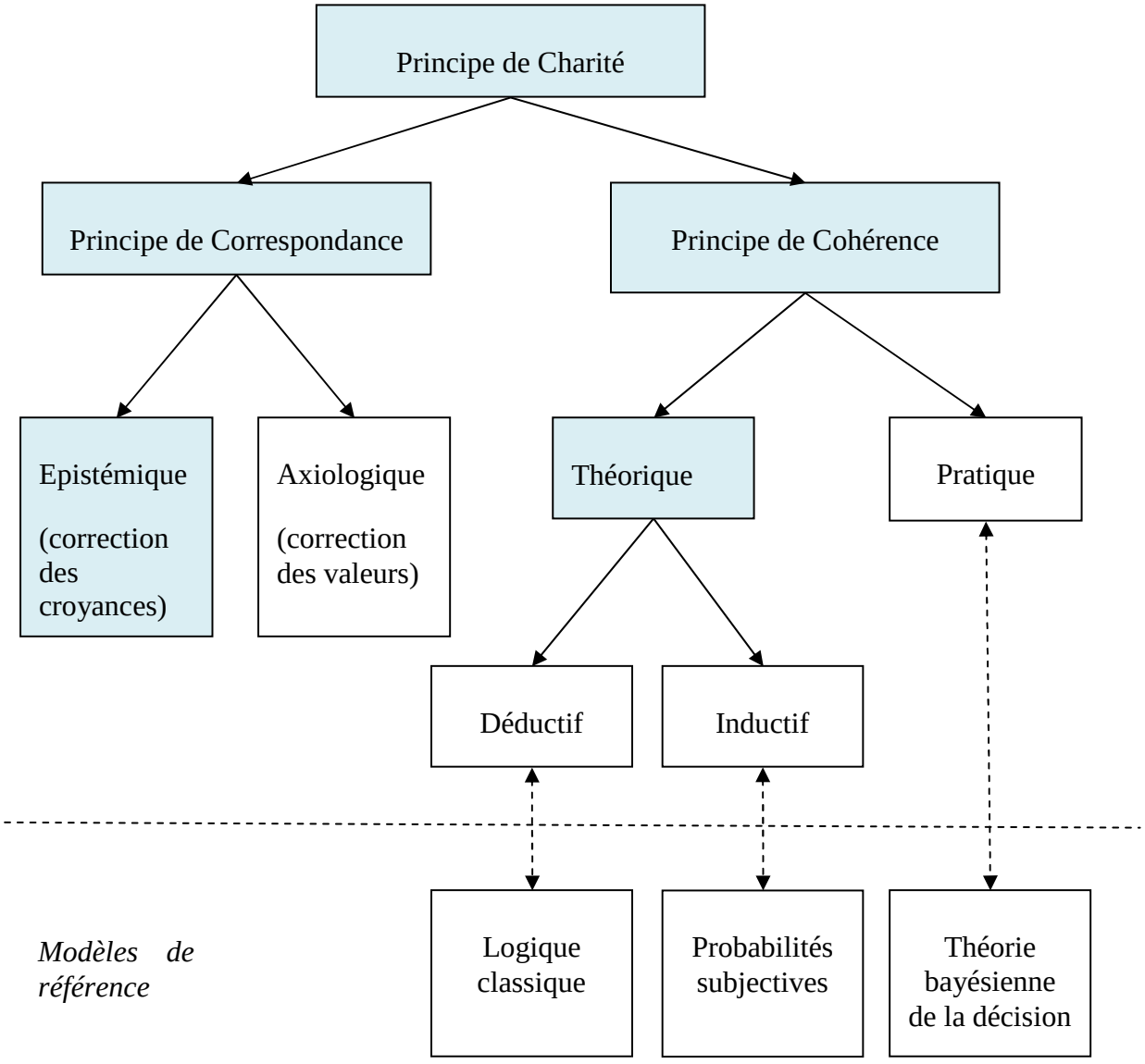

FIGURE 1. Le principe de charité.

3 Pourquoi être charitable ?

3.1 Les conséquences épistémologiques de la charité

L'adoption du principe de charité a des conséquences épistémologiques immédiates. Toute discipline scientifique relevant des sciences humaines vise à expliquer certains comportements ou certains états mentaux, et le fait en invoquant d'autres comportements et d'autres états mentaux[9]. Par exemple, en simplifiant quelque peu, l'économie étudie la manière dont les agents produisent, acquièrent ou cèdent des biens en attribuant à ces agents certaines préférences et certaines croyances et en faisant des hypothèses quant à la manière dont préférences et croyances déterminent leurs comportements économiques. Partant, toute discipline scientifique relevant des sciences humaines doit respecter un principe de charité à la fois dans l'interprétation de ce que disent les agents (s'il y a lieu) et dans la caractérisation des états mentaux qu'elle leur attribue, sous peine de rendre littéralement incompréhensibles les comportements qu'elle étudie.

Cette contrainte normative sur les sciences de l'homme marque une différence majeure avec les sciences de la nature. Dans le cas de celles-ci, les théories ne sont soumises à aucune contrainte « extrinsèque » : la meilleure théorie est celle qui rend compte de la manière la plus adéquate des données (sans bien sûr qu'il soit forcément facile de dire ce que l'on entend par rendre compte de manière adéquate des données). Dans le cas des sciences de l'homme, on attend bien sûr d'une théorie particulière qu'elle rende compte de manière adéquate des données pertinentes, mais en plus cette théorie doit *autant que faire se peut* interpréter les agents comme des agents rationnels. Dans les termes de Davidson,

> *« Each interpretation and attribution of attitude is a move within a holistic theory, a theory necessarily governed by concern for consistency and general coherence with the truth, and it is this that sets these theories forever apart from those that describe mindless objects, or describe objects as mindless. »* (Davidson 1974a, p. 322)

On notera la précaution de Davidson à propos des théories « décrivant des objets considérés comme dépourvus d'esprit ». En effet, il est toujours possible d'imaginer une discipline prenant pour objet certains comportements humains par exemple, mais qui refuserait de les traiter comme des comportements intentionnels. De fait, la psychologie behavioriste, qui explique les comportements futurs uniquement sur la base des comportements passés en refusant d'entrer dans la « boîte noire » de l'esprit, constitue dans une certaine mesure une telle discipline[10]. Parce que les objets à expliquer et les explications à fournir n'incluraient ni

---

[9] Toutes choses égales par ailleurs : on discute plus bas le cas d'une science qui traiterait des objets « humains » sans recourir à une analyse intentionnelle.

[10] Si le principe de charité ne contraint pas les théories behavioristes, on peut s'étonner de ce que le principe fasse son apparition chez Quine dans un contexte qui est, de manière revendiquée, behavioriste. On peut dire au moins deux choses pour dissiper l'effet de surprise. Premièrement, le contexte de la traduction radicale n'est peut-être pas aussi behavioriste que Quine veut bien le dire : en effet, il ne s'agit pas seulement de donner des lois permettant de prédire les comportements (linguistiques) futurs des indigènes à partir de leurs comportements linguistiques passés ; il s'agit également de donner un manuel de traduction, et donc de considérer les énoncés indigènes comme doués de signification. C'est précisément au moment où s'effectue la traduction qu'entre en jeu le principe de charité. A la limite, on pourrait imaginer un ensemble de lois purement behavioriste prédisant au moins partiellement les comportements linguistiques des indigènes qui ne les

énoncés interprétés ni états mentaux, une telle discipline ne serait pas soumise au principe de charité ; reste qu'il serait toujours possible de lui contester le titre de « science de l'homme », dans la mesure où elle ne traiterait pas d'objets humains *en tant qu'humains*.

Les contraintes *a priori* que fait peser le principe de charité sur les sciences de l'homme ne sont pas triviales. Comme on l'a évoqué dans l'introduction, et comme on y reviendra plus en détail dans la section suivante, elles semblent en particulier entrer en conflit avec bien des données recueillies par la psychologie cognitive, données qui suggèrent que nos pratiques inférentielles et notre compréhension naïve des probabilités ne satisfont pas les contraintes de rationalité les plus élémentaires. Il est donc plus que temps d'examiner plus précisément les raisons qui plaident en faveur de l'adoption du principe de charité dans l'interprétation radicale et plus généralement dans la caractérisation d'états mentaux.

3.2 Deux justifications de la charité

Chez Davidson au moins, il existe *deux types de justifications* relativement distinctes du principe de charité : la justification méthodologique et la justification conceptuelle[11]. Nous les envisagerons tour à tour.

(1) **La justification méthodologique**. Selon Davidson, le principe de charité permet de résoudre un problème méthodologique central en théorie de l'interprétation, celui de l'interdépendance entre la signification et la croyance :

> *« This method is intended to solve the problem of the interdependence of belief and meaning by holding belief constant as far as possible while solving for meaning. »* (Davidson 1973, p.137)

Considérons l'exemple suivant : supposons que notre locuteur Kurt tienne pour vrai l'énoncé de son langage « Es regnet. » si, et seulement si, il pleut à proximité de lui. Et demandons-nous ce que l'on peut inférer de telles données quant à la signification de « Es regnet ». Si nous appliquons le principe de charité, plus précisément, le principe de correspondance, nous devons considérer que Kurt tient pour vrai ce qui *est vrai*. Alors, en simplifiant, nous pouvons inférer des données disponibles que « Es regnet » signifie quelque chose de très proche de ce que signifie « Il pleut ».

Supposons, *a contrario*, que l'on n'applique pas le principe de charité et que l'on considère qu'il est parfaitement possible que Kurt tienne pour vrai ce qui est manifestement

---

envisagerait pas comme des énoncés doués de signification et traduisibles dans notre langue. Deuxièmement, précisément parce que Quine envisage la réalisation d'un manuel de traduction d'un point de vue behavioriste, le principe de charité ne relève chez lui que du « *common sense* » : toutes choses égales par ailleurs, il semble raisonnable de considérer qu'une théorie qui ne considère pas les locuteurs indigènes comme systématiquement irrationnels a plus de chances d'être correcte. Pour autant, le principe de charité ne serait pas *pour Quine* un principe *a priori*.

[11] On pourra comparer avec les justifications proposées par Lepore & Ludwig (2005, chap. 13) dans le cadre de la théorie de l'interprétation. D'autres que Davidson proposent d'autres justifications. On trouvera chez Stich (1985) une critique des arguments de Dennett et Cohen. Stein (1996) propose un examen systématique des usages du principe de charité. Il soutient en substance que seules des versions fortes du principe de charité permettent de soutenir que nous sommes rationnels, mais que seules des versions faibles du principe peuvent être justifiées.

faux. Alors nous ne sommes plus en position d'inférer de nos données que « Es regnet » signifie approximativement « Il pleut ». Après tout, peut-être Kurt croit-il à chaque fois qu'il fait beau et dans ce cas « Es regnet » signifie approximativement « Il fait beau ». En d'autres termes, la signification des énoncés d'un langage se retrouve largement sous-déterminée par les énoncés qu'un locuteur de ce langage tient pour vrais. Pour pouvoir inférer la signification à partir des données disponibles, il faut faire des hypothèses supplémentaires sur les croyances du locuteur.

C'est ici que le principe de charité est essentiel. En supposant que Kurt a des croyances adéquates et cohérentes, on détermine partiellement ses croyances, en tout cas suffisamment pour pouvoir commencer à attribuer une signification à certains de ses énoncés. Sans le principe de charité, l'interprète est perdu, car l'attitude de Kurt à l'égard d'énoncés de son langage dépend en effet à la fois de la signification des énoncés et des croyances de Kurt, toutes choses que l'interprète ignore. Avec le principe de charité, l'interprète a un accès partiel aux croyances de Kurt et il peut commencer à trouver son chemin parmi les significations.

Sur ce point, l'analogie avec la théorie de la décision est frappante[12] : le théoricien de la décision fait face à une autre interdépendance, celle des croyances et des désirs de l'individu. Supposons par exemple que Kurt ne sait pas ce que Karl a préparé pour le dîner auquel ce dernier l'invite. Il se peut que Karl ait préparé de la viande, il se peut également qu'il ait préparé du poisson. Kurt doit apporter une bouteille de vin et a le choix entre une bouteille de vin blanc et une bouteille de vin rouge. Supposons que Karl prenne une bouteille de vin blanc. On peut vouloir en inférer que Kurt pense qu'il est plus probable que Karl ait préparé du poisson. Mais on ne fera cette inférence que si l'on suppose que Kurt préfère le vin blanc avec le poisson et le vin rouge avec la viande. Si Kurt a des goûts hétérodoxes et qu'il les impose volontiers aux autres, on ne fera plus cette inférence. Plus subtilement, il se peut qu'il y ait, aux yeux de Kurt, une asymétrie entre le déplaisir occasionné par du vin rouge avec du poisson et le déplaisir occasionné par du vin blanc avec de la viande. De manière générale, la décision dépend conjointement des croyances et des désirs du décideur. La théorie de la décision propose un certain nombre de contraintes de rationalité sur les *préférences* qui permettent de démêler l'interdépendance et de révéler les croyances et les préférences. Ainsi, Savage (1954)[13] montre que si les préférences du décideur satisfont un ensemble restreint de propriétés (parmi lesquelles la transitivité), alors il existe une distribution de probabilité subjective sur les états de la nature possibles et une fonction d'utilité sur les conséquences possibles des actions telles que l'action *a* est préférée à l'action *b* ssi l'espérance d'utilité induite de *a* est supérieure à l'espérance d'utilité induite de *b* (théorème de représentation). La preuve de ce théorème de représentation, que l'on trouve dans tout manuel de théorie de la décision, montre précisément comment on peut (en principe) inférer les probabilités

---

[12] Elle est faite très tôt par Davidson lui-même (1974a, pp.145-46) : « …we should think of meanings and beliefs as interrelated constructs of a single theory just as we already view subjective values and probabilities as interrelated constructs of decision theory. » Rappelons que, au milieu des années 1950, Davidson a été l'un des pionniers de la théorie de la décision expérimentale lors de sa collaboration avec P. Suppes qui a débouché sur Davidson, Suppes & Siegel (1957). Sur le rapport de D. Davidson à la théorie de la décision, on pourra consulter le récent travail de Harnay (2008).

[13] Nous considérons Savage car sa théorie fait aujourd'hui figure de référence en théorie de la décision. Davidson invoque plus volontiers les travaux pionniers de F. Ramsey ou la théorie de R. Jeffrey, plus répandue dans la littérature philosophique.

subjectives puis les utilités à partir des préférences. Le TABLEAU 1 résume le parallèle méthodologique entre la théorie de l'interprétation et la théorie de la décision.

Notons que la solution au problème de l'interdépendance apportée au sein de la théorie de la décision ne dispense pas de résoudre le problème de l'interdépendance tel qu'il se pose pour la théorie de l'interprétation. En effet, la théorie bayésienne de la décision ne se substitue pas à une théorie de l'interprétation, elle la présuppose, puisque les préférences dont elle traite sont des préférences entre propositions. Or pour déterminer le contenu des énoncés, c'est-à-dire les propositions, on a précisément besoin d'une théorie de l'interprétation[14]. Du point de vue méthodologique, le principe de charité apparaît alors comme une des hypothèses clefs qui permet à une théorie du mental d'échapper aux interdépendances entre états mentaux, et entre états mentaux et contenus, et de s'édifier sans une trop grande sous-détermination dans les attributions de contenus et d'états mentaux. D. Davidson est d'ailleurs revenu plusieurs fois sur le projet d'une Théorie Unifiée de l'interprétation et de la décision[15]

	Théorie de l'interprétation	Théorie de la décision
Données de base	Enoncés tenus pour vrais	Préférences entre actions
Objectif	Assignation de signification aux énoncés d'un langage-cible	Attributions de désirs et de croyances à un agent
Problème méthodologique	Interdépendance des croyances et des significations	Interdépendance des croyances et des désirs
Solution	Principe de charité appliqué aux énoncés tenus pour vrais	Principe de charité appliqué aux préférences

TABLEAU 1. Problèmes méthodologiques en theories de l'interprétation et de la decision.

[14] Ce point est explicite chez Davidson : « We generally find out exactly what someone wants, prefers or believes only by interpreting his speech… Clearly, a theory that attempts to elicit the attitudes and beliefs that explain preferences or choices must include a theory of interpretation if it is not to make crippling assumptions » (Davidson, 1990a, p. 318).

[15] Voir Davidson (1985, 1990a). Davidson part d'une théorie de la décision « à la Jeffrey », où les probabilités et les utilités sont attachées aux mêmes entités, mais au lieu que celles-ci soient des *propositions*, ce sont désormais des *énoncés* qu'il faut interpréter. Dans ce cas, il faut enrichir l'axiomatique de Jeffrey pour pouvoir déceler la structure logique des énoncés, structure qui est nécessaire à l'application des axiomes originaux. Cet axiome supplémentaire est présenté dans l'appendice de Davidson (1990a), p. 328.

(2) **La justification conceptuelle**[16]. L'idée est la suivante : il existe un lien de nature *conceptuelle* entre la mentalité (les états mentaux et les comportements intentionnels) et l'hypothèse de rationalité. Comme le dit Davidson,

> *« Si quelqu'un croit que Tahiti est à l'est d'Honolulu, alors il devrait croire qu'Honolulu est à l'ouest de Tahiti. C'est pourquoi, si nous sommes certains que la personne croit qu'Honolulu est à l'ouest de Tahiti, c'est probablement une erreur d'interpréter ce qu'elle dit comment exprimant qu'elle croit également que Tahiti est à l'ouest d'Honolulu. S'il s'agit probablement d'une erreur, ce n'est pas parce que ce serait un fait empirique que les gens ont rarement des conceptions contradictoires, mais parce que les croyances (et les autres attitudes) sont largement identifiées par les relations, notamment logiques, qu'elles entretiennent entre elles ; changer ces relations, c'est changer l'identité de la pensée. »* (Davidson, 1990b, nous traduisons)

Pour autant que les contraintes de rationalité font partie des conditions qui définissent les croyances (ainsi que les autres attitudes) et des critères de leur attribution, il n'est pas possible que ces contraintes ne soient pas satisfaites. Considérons le principe de non-contradiction et imaginons par exemple que, sur la base du comportement de Jean, on considère qu'il ne croit pas que Marie ne viendra pas. Il est alors possible de lui attribuer la croyance que Marie viendra, sans qu'on soit justifié à le faire sans plus d'information. Pour peu que l'on pense que Jean sait *si* Marie viendra, on lui attribuera à bon droit la croyance que Marie viendra. Inversement si le comportement de Jean suggère qu'il croit que Marie ne viendra pas, on s'interdira de lui attribuer la croyance que Marie viendra. Ne pas croire que *non-p* constitue une partie de ce que c'est que croire que *p* : c'est une condition toujours nécessaire et parfois suffisante (quand l'agent est bien informé). De la même façon, on pourra attribuer à un agent une préférence pour *a* relativement à *c* si l'on peut trouver un *b* tel que cet agent préfère *a* à *b* et *b* à *c*. Et comme ne pas préférer *c* à *a* fait également partie de ce que c'est que préférer *a* à *c*, on dira qu'un agent qui préfère *c* à *a*, mais également *a* à *b* et *b* à *c* pour un certain *b*, ne préfère pas réellement *c* à *a*. On peut alors conclure avec Davidson :

> *«la satisfaction des conditions de non contradiction et de cohérence rationnelle [est]constitutive du domaine d'application de concepts comme ceux de croyance, de désir, d'intention et d'action. »* (Davidson 1974b, trad.fr. p. 315)

Les deux justifications pourraient sembler valoir chacune pour l'un des deux versants du principe de charité, la justification méthodologique valant pour le principe de correspondance et la justification conceptuelle pour le principe de cohérence. En effet, c'est le principe de correspondance qui donne accès aux équivalences T pour les énoncés simples, tandis que les relations logiques entre les croyances ou les préférences qui sont constitutives de leur domaine d'application assurent la satisfaction du principe de cohérence. Néanmoins, on peut envisager d'étendre la portée des deux justifications de manière à ce que chacune couvre l'entièreté du principe de charité.

Voyons d'abord le cas de la justification méthodologique et du principe de cohérence. L'obtention des équivalences T n'est pas la seule condition nécessaire à l'élaboration d'une

---

[16] Føllesdal (1982) distingue explicitement la justification méthodologique et la justification conceptuelle : « I agree with Davidson that rationality is constitutive of belief, desire, action, etc. and not just needed in order to find out what beliefs and desires a person has and what actions he performs. »

théorie de la signification pour un langage inconnu. Pour donner une théorie *récursive* de la signification, on doit également être en mesure d'identifier et de traduire les constituants logiques des énoncés indigènes. Et afin de les identifier, on est bien obligé de supposer qu'ils respectent les principes logiques[17]. On justifierait ainsi le principe de cohérence pour son versant théorique. S'agissant du domaine pratique du principe de cohérence, ce sont également les propriétés des croyances et des désirs données par les contraintes de rationalité qui permettent leur révélation ; on ne pourrait pas briser le cercle entre croyance et désir de la façon indiquée par Ramsey et reprise par Davidson si l'on ne pouvait pas s'appuyer sur certaines hypothèses concernant les croyances et les désirs.

Peut-on également développer la justification conceptuelle de manière à englober le principe de correspondance ? Cela semble difficile *: prima facie*, rien dans le concept de croyance ne dit que Kurt doive croire qu'il pleut lorsqu'il pleut, de même que rien ne dit que Kurt doive croire que la neige est blanche. Pourtant, à y regarder de plus près, on peut pourtant soutenir qu'un degré minimum d'adéquation des croyances de Kurt soit conceptuellement nécessaire. D'abord, la nature de l'environnement d'un agent fait bien partie des critères que nous utilisons pour lui attribuer des croyances : toutes choses égales par ailleurs, une bonne raison d'attribuer à quelqu'un la croyance est qu'il pleut est de constater qu'il se trouve dans un environnement où il pleut. Si les croyances sont pour partie définies par des relations logiques, elles sont également pour partie définies en termes d'interactions causales avec l'environnement. Ensuite, les conditions d'application d'autres concepts que les concepts d'attitude sont en jeu. Pour que l'on puisse attribuer à Kurt la croyance qu'il neige, il faut bien qu'il soit possible de lui attribuer des croyances à propos de la neige, et donc qu'il identifie au moins partiellement ce qu'est la neige sur la base de certaines croyances vraies (comme la croyance que la neige est blanche). Sera alors justifié un principe de correspondance appliqué à des vérités générales entrant dans la caractérisation des concepts utilisés dans l'attribution de croyances.

La justification méthodologique et la justification conceptuelle font de l'hypothèse de rationalité quelque chose comme une condition de possibilité de la compréhension. Le principe de charité vaut parce que pour comprendre autrui, on doit supposer qu'il est rationnel. Et l'on doit supposer qu'il est rationnel premièrement parce que la procédure d'interprétation radicale repose en partie sur les contraintes de rationalité et que la possibilité de l'interprétation radicale conditionne la compréhension (justification méthodologique), et deuxièmement parce que les contraintes de rationalité sont constitutives des concepts d'attitudes et que l'attribution d'attitudes conditionne la compréhension (justification conceptuelle). Davidson est alors justifié à s'exprimer de la manière suivante :

> *« Charity is forced on us; whether we like it or not, if we want to understand others, we must count them right in most matters »* (Davidson 1973, p.197)

Fodor et Lepore (1994) semblent ainsi ne faire que prendre acte de l'approche de Davidson lorsqu'ils parlent, au moment de critiquer la justification méthodologique, d'argument « transcendental ». Pourtant Davidson dans sa réponse (Davidson, 1994) n'accepte pas le

---

[17] « We must expect the theory [of truth] to rely on something very like Tarski's sort of recursive characterization of satisfaction, and to describe sentences of the object language in terms of familiar patterns created by quantification and cross-reference, predication, truth-functional connections, and so on. » (Davidson, 1974a)

qualificatif : nous verrons dans le prochain paragraphe que la question de savoir si les justifications envisagées constituent réellement ou non des conditions de possibilité est cruciale.

3.3 Examen critique

Nous nous proposons maintenant de revenir tour à tour sur les justifications invoquées en faveur du principe de charité.

(1) **Critique de la justification méthodologique**. Rappelons l'idée directrice de la justification méthodologique : l'hypothèse de rationalité est justifiée parce qu'elle permet, dans un cas, d'assigner une signification au langage d'un locuteur-cible en brisant le cercle des croyances et des significations, dans l'autre, d'attribuer des croyances et des désirs à un agent-cible sous la forme d'une distribution de probabilité subjective et d'une fonction d'utilité, malgré l'interdépendance des croyances et des désirs.

Considérons d'abord le cas de la théorie de la décision, qui a inspiré l'argument de Davidson en faveur du principe de correspondance et pour laquelle on dispose d'une théorie mathématique établie. Ce que le théorème de représentation montre, c'est que *si* les préférences de l'agent obéissent à certaines propriétés, *alors* on peut rendre compte de ces préférences à partir de l'attribution de croyances dotées d'une certaine structure (celle d'une distribution de probabilité), de l'attribution de désirs dotées eux aussi d'une certaine structure (celle d'une fonction d'utilité) *et* d'un principe qui intègre ces croyances et ces désirs (celui de la maximisation de l'espérance d'utilité). Prêtons attention à la modalité : la conformité aux axiomes sur les préférences rend *possible* une explication mentale d'un certain type. Le problème que l'on rencontre est alors le suivant : le principe de charité affirme qu'il *est nécessaire* de faire l'hypothèse de rationalité, mais ce que semblent montrer, par exemple, un théorème de représentation, c'est que l'hypothèse de rationalité *suffit* pour procéder à des attributions d'un certain type. De manière générale, ce n'est pas parce que l'hypothèse de rationalité constitue (supposons le) *une* solution aux problèmes d'interdépendance entre croyances, désirs et significations que c'est *la seule* solution possible.

Concernant la théorie de la décision, on pourrait toujours être tenté de répliquer ceci : la plupart des axiomes sur les préférences sont également des conséquences *nécessaires* d'une explication en termes de maximisation d'espérance d'utilité. Par exemple, si les croyances d'un agent ont une structure probabiliste, ses désirs celle d'une fonction d'utilité et s'il maximise son espérance d'utilité, alors ses préférences seront transitives, obéiront à l'axiome de la chose sûre, etc. L'objection à la justification méthodologique serait donc fallacieuse parce qu'elle ne considère qu'un seul « sens » du théorème de représentation. Cette réponse n'est cependant pas convaincante en l'état : elle ne vaut que parce que l'on suppose qu'une explication mentale doit reposer sur des croyances probabilistes, la maximisation de l'espérance d'utilité, etc. Or faire cette supposition, c'est accepter la justification conceptuelle, selon laquelle les contraintes de rationalité sont constitutives des concepts d'attitudes. Il apparaît donc que la justification méthodologique n'apporte rien de plus par rapport à la justification conceptuelle, puisqu'elle la présuppose. Et l'on va voir que la justification conceptuelle est problématique.

Une critique analogue vaut dans le cas de l'interprétation. Davidson montre que le principe de charité est suffisant pour expliquer l'interprétation radicale ; il le fait en décrivant

une succession d'étapes par lesquelles l'interprète arrivera à une théorie de la vérité pour le langage-cible en s'appuyant sur le principe de charité. Mais, premièrement, rien ne dit qu'il n'existe pas d'autres moyens pour parvenir à une théorie de la signification en situation d'interprétation radicale. Et deuxièmement, rien ne dit que, pour que l'interprétation soit possible, l'interprétation radicale doive l'être également. Ce second point est l'objet des critiques de Fodor et Lepore (1994), qui nous semblent tout à fait pertinentes. Fodor et Lepore insistent en particulier sur le fait que l'interprétation non-radicale est plus facile que l'interprétation radicale (le linguiste comme l'enfant abordent l'interprétation avec des connaissances d'arrière-plan), de sorte que la nécessité d'accepter une forme forte du principe de charité devient douteuse. Davidson admet lui-même que la portée de ses arguments doit être circonscrite, lorsqu'il explique dans sa réponse à Fodor et Lepore que la question à laquelle il a cherché une réponse est de celle de savoir « ce qu'il suffirait à un interprète de savoir pour pouvoir comprendre le locuteur d'une langue inconnue, et comment il pourrait parvenir à le savoir » (Davidson, 1994, p. 126, nous traduisons).

Une fois reconnu que le principe de charité constitue une condition suffisante plutôt qu'une condition nécessaire, on pourrait considérer que la justification méthodologique conserve une certaine force, en l'absence d'autre explication possible. Malheureusement, il n'en est rien. Il semble bien que dans certains cas il soit possible et préférable pour l'interprétation de se passer du principe de charité. Considérons le scénario suivant que l'on doit à R. Grandy. Kurt arrive à un dîner et dit à un autre convive : « l'homme qui boit un Martini est un philosophe ». Il y a bien un homme dans le champ visuel de Kurt, mais cet homme boit de l'eau dans un verre à Martini et il n'est pas philosophe. Dans une autre pièce, il y un homme qui boit un Martini (c'est le seul convive à boire un Martini), et c'est un philosophe. Si l'on interprète « l'homme qui boit un Martini » comme référant au premier homme (celui qui boit de l'eau dans un verre à Martini), alors il est difficile de considérer comme vrai l'énoncé prononcé par Kurt. Si l'on interprète « l'homme qui boit un Martini » comme référant au second homme (celui qui est dans une autre pièce), alors on peut considérer comme vrai l'énoncé prononcé par Kurt. L'interprète qui veut autant que possible que les assertions de Kurt soient correctes, conformément au principe de charité, penchera pour la seconde interprétation. Mais celle-ci ne semble pas correcte : Kurt n'a aucune connaissance que ce soit de ce second homme qui boit réellement un verre de Martini et tout semble indiquer qu'il réfère à l'homme qui boit de l'eau dans un verre à Martini. « Il vaut mieux attribuer [à Kurt] une fausseté explicable plutôt qu'une vérité mystérieuse » (Grandy 1973, p. 445). On pourrait objecter que le principe de charité laisse bien la place à l'attribution d'erreurs, pourvu que ce soit sur le fond d'un large accord[18]. Mais pourquoi l'accord devrait-il l'emporter sur le désaccord ? Pourquoi ne pourrait-on pas imaginer une situation dans lequel l'interprète, utilisant sa meilleure théorie psychologique, aurait des raisons de supposer que l'indigène risque d'avoir des croyances systématiquement erronées ? De manière générale, il n'y a aucune raison de penser que le principe de charité constitue nécessairement la meilleure

[18] Cette objection est faite par Lepore et Ludwig (2005), qui soutiennent que les positions de Davidson sont plus proches que ne le considèrent la plupart des commentateurs des positions développées par Grandy (1973) pour prendre en compte l'exemple précédent. Certes – et heureusement – Davidson reconnaît la légitimité d'attributions d'erreur. Il n'en reste pas moins qu'elles doivent toujours se faire pour lui sur fond d'accord, alors que l'exemple de Grandy montre très simplement que l'attribution d'erreur est préférable, dès que nos théories sur la formation de croyances (dans ce cas précis, une théorie assez élémentaire) nous disent qu'il est plus plausible que le locuteur se trompe.

manière d'inférer de l'environnement du locuteur à ses croyances afin de briser le cercle des croyances et des significations.

(2) **Critique de la justification conceptuelle**. L'idée directrice de la justification conceptuelle est que les contraintes de rationalité sont constitutives des concepts d'attitudes et que la compréhension d'autrui requiert l'utilisation de ces concepts d'attitudes.

Les deux points sont en réalité sujets à caution. Admettons pour l'instant le premier, admettons également que la théorie de la décision classique énonce des propriétés définitionnelles des concepts classiques de désir, de croyance et d'action intentionnelle. Pourquoi ne serait-il pas possible d'élaborer et d'utiliser une autre théorie énonçant des propriétés définitionnelles d'autres concepts possibles de désir, de croyance et d'action intentionnelle ? Pourquoi ne serait-il pas possible d'élaborer une théorie du mental qui fasse l'économie de contraintes de rationalité d'une allure et d'une force comparable à celles que l'on trouve dans la théorie de la décision classique ? De fait, une des leçons des dernières décennies en théorie de la décision semble bien être que la vie théorique sans l'un ou l'autre des principes de rationalité n'est pas nécessairement chaotique. Considérons par exemple le *principe de la chose sûre*[19] introduit par Savage (1954) dans la théorie du choix en incertitude : ce principe est, du point de vue axiomatique, le nerf du modèle d'espérance d'utilité. Depuis les années 1970, en réaction au fameux « paradoxe d'Ellsberg » dont on considère en général qu'il montre que le principe de la chose sûre *n'est pas* respectée par les individus, une bonne partie de la théorie de la décision en incertitude étudie des modèles qui affaiblissent ou rejettent le principe de la chose sûre. Même si le principe de la chose sûre fait partie de la définition de la notion « classique » de préférence (dans l'incertain), rien n'empêche alors d'utiliser une notion de préférence* dont les propriétés définitionnelles seraient données par l'une des théories non-classiques. On ne voit alors pas au nom de quoi il faudrait être charitable et utiliser le concept de préférence classique plutôt que le concept de préférence* : dans la mesure où il est possible de décrire le comportement intentionnel des agents à l'aide du concept de préférence*, il n'est tout simplement pas nécessaire de supposer qu'ils sont rationnels au sens où ils respecteraient le principe de la chose sûre. Notons que ce genre de problème est un problème général pour les stratégies argumentatives reposant sur l'idée de justification conceptuelle. Ainsi de manière analogue, on pourrait dire que l'axiome des parallèles énonce une propriété définitionnelle de l'espace euclidien « classique », de sorte que l'axiome des parallèles serait « constitutif du domaine d'application du concept d'espace (euclidien) ». Mais de ceci il ne suit pas que l'espace physique satisfasse nécessairement l'axiome des parallèles, puisqu'on peut utiliser un autre concept d'espace, par exemple un concept d'espace riemannien, pour le décrire.

Revenons au premier point, et suivons la critique proposée par Cherniak (1986) de l'hypothèse de rationalité. Admettons qu'on dispose seulement des théories classiques comme théories d'arrière-plan portant sur nos croyances, désirs et utilités. Sans doute doit-on considérer qu'un individu qui violerait systématiquement l'ensemble des contraintes de rationalité – par exemple un individu qui croirait toujours aussi que *non-p* lorsqu'il croit que *p*

---

[19] Soit deux actions *f* et *g* qui ont la même conséquence si un événement *E* survient (par exemple deux paris sur une course de cheval qui ne rapportent rien si la jument Etoile du Nord n'est pas placée). Soient par ailleurs deux actions *f'* et *g'* qui sont respectivement identiques à *f* et *g* quand *E* ne survient pas, et qui ont la même conséquence si *E* survient. Le principe de la chose sûre affirme que, dans ce cas, un individu préfère *f* à *g* ssi il préfère *f'* à *g'*.

– ne pourrait pas être compris. Mais est-il nécessaire pour autant que cet individu respecte systématiquement l'ensemble des contraintes de rationalité ? D'une part, il est tout simplement impossible pour un individu donné de le faire : les agents réels sont des êtres finis qui ne peuvent satisfaire les contraintes idéales des théories de la rationalité. Par exemple, un agent ne peut pas toujours détecter les incohérences dans son système de croyance parce que cette détection lui prendrait un temps dépassant de loin son espérance de vie ! Si les croyances des agents devaient réellement satisfaire les contraintes de rationalité, il n'y aurait tout simplement pas d'agent ayant des croyances. D'autre part, dans le cas des croyances, il semble bien que les relations logiques qui sont constitutives de notre concept de croyance soient relativement minimales. Sans doute la disposition à réviser ses croyances si l'on apprend explicitement que *non-p* est-elle constitutive de la croyance que *p*. Mais si l'on devait dire que croire toutes les conséquences de *p* modulo les croyances déjà acceptées est constitutif de la croyance que *p*, il semblerait que l'on dépasse de loin les exigences liées à l'attribution de la croyance que *p*. Ces deux arguments suggèrent que seule une hypothèse de rationalité minimale, pour reprendre l'expression de Cherniak, compatible avec les limitations cognitives des agents, peut être dérivée de nos concepts d'attitude[20].

Qu'en est-il alors du caractère apparemment transcendantal du principe de charité ? Si les critiques précédentes sont valables, l'hypothèse de charité ne serait pas réellement une condition de possibilité de la compréhension. Une telle conclusion est bienvenue, s'il est vrai que, *de fait*, il peut y avoir compréhension sans « rationalisation ». Considérons la célèbre expérience de D. Kahneman et A. Tversky sur les jugements probabilistes (Tversky & Kahneman, 1983). Voici le scénario présenté :

> *Linda est une jeune femme de 31 ans, célibataire et très intelligente. Elle a étudié la philosophie. Pendant ses études elle était impliquée dans la lutte contre les discriminations et pour la justice sociale, et elle a participé à des manifestations anti-nucléaires.*

On demande au sujet de classer par ordre de probabilité différents énoncés et en particulier les suivants :

> *(1) Linda est active dans le mouvement féministe*
>
> *(2) Linda est banquière*
>
> *(3) Linda est banquière et active dans le mouvement féministe*

Selon la réponse dominante, (3) est plus probable que (2). C'est manifestement une réponse problématique si on la compare aux principes de rationalité les plus enracinés. (3) est la conjonction de (2) et de (1) donc (3) implique à la fois (1) et (2). Il suit des lois du calcul des probabilités que la probabilité de (3) ne peut pas être plus élevée que celle de (2). Intuitivement, la probabilité de (2) est la somme de la probabilité de (3) et de « Linda est banquière et n'est pas active dans le mouvement féministe ».

---

[20] Notons que l'idée d'hypothèses de rationalité minimale n'est pas compatible avec l'idée selon laquelle les hypothèses de rationalité devraient suffire à rendre l'interprétation radicale possible. Sur les bases de principes aussi faibles que, par exemple, la propension à éviter dans certains cas les incohérences, il apparaît difficile de suivre la route indiquée par Davidson pour démêler croyances et significations ou désirs et croyance.

Supposons que la réponse de Kurt au scénario de Linda soit la même que la réponse modale. Il nous semble qu'il y a un sens dans lequel nous comprenons minimalement la réponse de Kurt : c'était notre réponse également lorsque le scénario nous a été soumis pour la première fois. Mais il est difficile d'aménager le principe de charité comme on pourrait le faire dans l'exemple précédemment rapporté de l'homme au Martini : la réponse dominante n'est pas une erreur (facilement) explicable, c'est une violation flagrante d'un principe élémentaire des probabilités. Il semble donc bien qu'il y ait divergence entre compréhension et hypothèse de rationalité[21]. Mais si l'hypothèse de rationalité n'est pas une condition de possibilité de l'attribution d'attitudes, cela veut dire qu'il y a d'autres voies pour l'interprétation et l'attribution de croyances et de désirs que la voie envisagée spéculativement par Davidson. A quelles conditions, qui n'implique pas l'hypothèse de charité, la compréhension est-elle possible ? Nous nous proposons maintenant de discuter cette question à la lumière non pas d'une enquête transcendantale, dont on a vu le statut problématique chez Davidson, mais sur les bases des théories de la mentalisation fournies par les sciences cognitives. Car après tout, puisque la compréhension est un fait, c'est que ses conditions de possibilité sont réalisées, et l'enquête empirique peut nous permettre de les saisir aussi bien que la « philosophie en fauteuil ».

4 Simulation et compréhension

Davidson prend soin d'affirmer que ni la théorie de l'interprétation radicale, ni la théorie de la décision ne *décrivent la façon dont, en fait, nous procédons* à l'interprétation, à l'explication de l'action ou à l'attribution d'états mentaux[22]. Il semble pourtant considérer que ces théories nous éclairent sur la façon dont nous procédons effectivement et sur les propriétés des « produits » de nos pratiques interprétatives, explicatives ou attributives. Elles révèleraient des caractéristiques essentielles des états mentaux, des créatures auxquelles nous les attribuons et des sciences de l'homme par opposition aux sciences naturelles. Si l'hypothèse de rationalité était *nécessaire* à l'interprétation et à la compréhension d'autrui, alors elle devrait se manifester dans nos pratiques *effectives* d'interprétation et de compréhension. Si elle entretient une relation plus complexe à l'interprétation et à la compréhension, dans la mesure où sa justification semble tirée en dernière instance de nos intuitions fondamentales sur ce que c'est que « comprendre » ou « rendre intelligible » un comportement, on en revient de nouveau à nos pratiques effectives d'interprétation et de compréhension. Par conséquent, il nous semble qu'une analyse adéquate du statut de l'hypothèse de rationalité doit s'intéresser aux mécanismes effectifs d'interprétation et de compréhension, et à la place éventuellement occupée par les principes de rationalité dans ces mécanismes. Ce travail sera esquissé dans la présente section, qui commence par présenter les deux grandes théories qui ont émergé au sein des sciences explicatives concernant la nature des mécanismes d'interprétation et de compréhension. Nous rappellerons ensuite les arguments convergents qui existent en faveur d'une de ces deux théories, la théorie de la simulation, avant, de discuter les conséquences de cette théorie pour le principe de charité.

---

[21] Un point analogue est formulé par Goldman (1989).

[22] « The approach to understanding some of the basic psychological facts about an agent I have just sketched does not pretend to describe how we actually learn to understand people. But if it describes a feasible approach, it helps explain, if only in theory, how it is possible to do what we in fact do with surprising ease.» (Davidson 1999, p. 253)

4.1 La mentalisation

Nous avons la capacité de décrire nos semblables en termes *mentaux* : nous leur attribuons des états mentaux comme des attitudes propositionnelles, nous décrivons leurs comportements en leur assignant des intentions, nous expliquons et prédisons leurs comportements en invoquant leurs états mentaux. Cette capacité (ou l'exercice de cette capacité) reçoit tantôt le nom de *mindreading*, tantôt celui de *mentalizing* dans la littérature philosophique et psychologique contemporaine. Faute de mieux, nous parlerons dans ce qui suit de *mentalisation*.

La philosophie et les sciences cognitives ont, durant les deux dernières décennies, activement exploré les *fondements cognitifs* de la mentalisation, et la mentalisation a fait l'objet de nombreuses investigations empiriques, notamment de la part de la psychologie cognitive, de la neuropsychologie et des neurosciences cognitives. La question principale est celle de savoir *comment fonctionne la mentalisation*. Deux théories principales se sont affrontées[23].

(1) Selon la *théorie « théorique »* (*theory theory*), la mentalisation se fonde sur des concepts organisés en une sorte de théorie. On appelle parfois la théorie en question « théorie de l'esprit » : il s'agit, si une telle chose existe, de la théorie du mental dont nous disposons tous et que nous utilisons afin d'attribuer à autrui certains états mentaux ou de prédire de sa part certains comportements. On obtient différentes versions de la théorie « théorique » selon que l'on considère que les concepts en jeu sont innés ou acquis, selon que l'on considère que cette théorie est à mettre au compte d'un module spécifique de l'esprit ou qu'elle est l'œuvre de processus généraux de théorisation, et enfin selon les contenus qu'on lui attribue. Sur le dernier point, selon certains (Lewis 1972, par exemple), notre théorie de l'esprit est une théorie qui ne contient que des platitudes à propos de nos concepts mentaux et comportementaux ; il s'agirait d'une théorie que nous serions tous en mesure d'expliciter. Selon d'autres, la mentalisation s'appuie au contraire sur un grand nombre d'informations largement inaccessibles par introspection. On peut par exemple soutenir que nos attributions d'émotions reposent sur toutes sortes d'indices (comme par exemple les expressions faciales) que nous traitons de manière essentiellement inconsciente. Dans tous les cas, on accepte l'idée que l'attribution d'états mentaux ou la prédiction de comportements se fait à partir des données pertinentes disponibles et d'inférences faites à partir de ces données sur la base de la théorie de l'esprit dont nous disposons. Ainsi le fonctionnement de la mentalisation serait, dans ses grandes lignes, le suivant. Soit *TPN* notre théorie psychologique naïve. Quand nous prédisons un comportement *c* à partir d'un certain nombre d'états mentaux que notre agent-cible est censé posséder (disons un ensemble de croyances *b* et un ensemble de désirs *d*), *c* est le genre de comportement auquel on doit s'attendre quand on considère *TPN* et que l'on prend les états mentaux comme conditions initiales. La FIGURE 2 schématise une activité de mentalisation prédictive : la *TPN* et nos hypothèses sur les croyances et les désirs de l'agent-cible appartiennent à la « boîte à croyances » de celui qui mentalise, lequel opère ensuite un raisonnement théorique pour vérifier que le comportement est bien attendu étant donné ses croyances et hypothèses.

---

[23] Sur le débat entre théorie « théorique » et théorie de la simulation, voir Goldman (1992), Gordon (1992), Heal (1996), Stich & Nichols (1992, 2003).

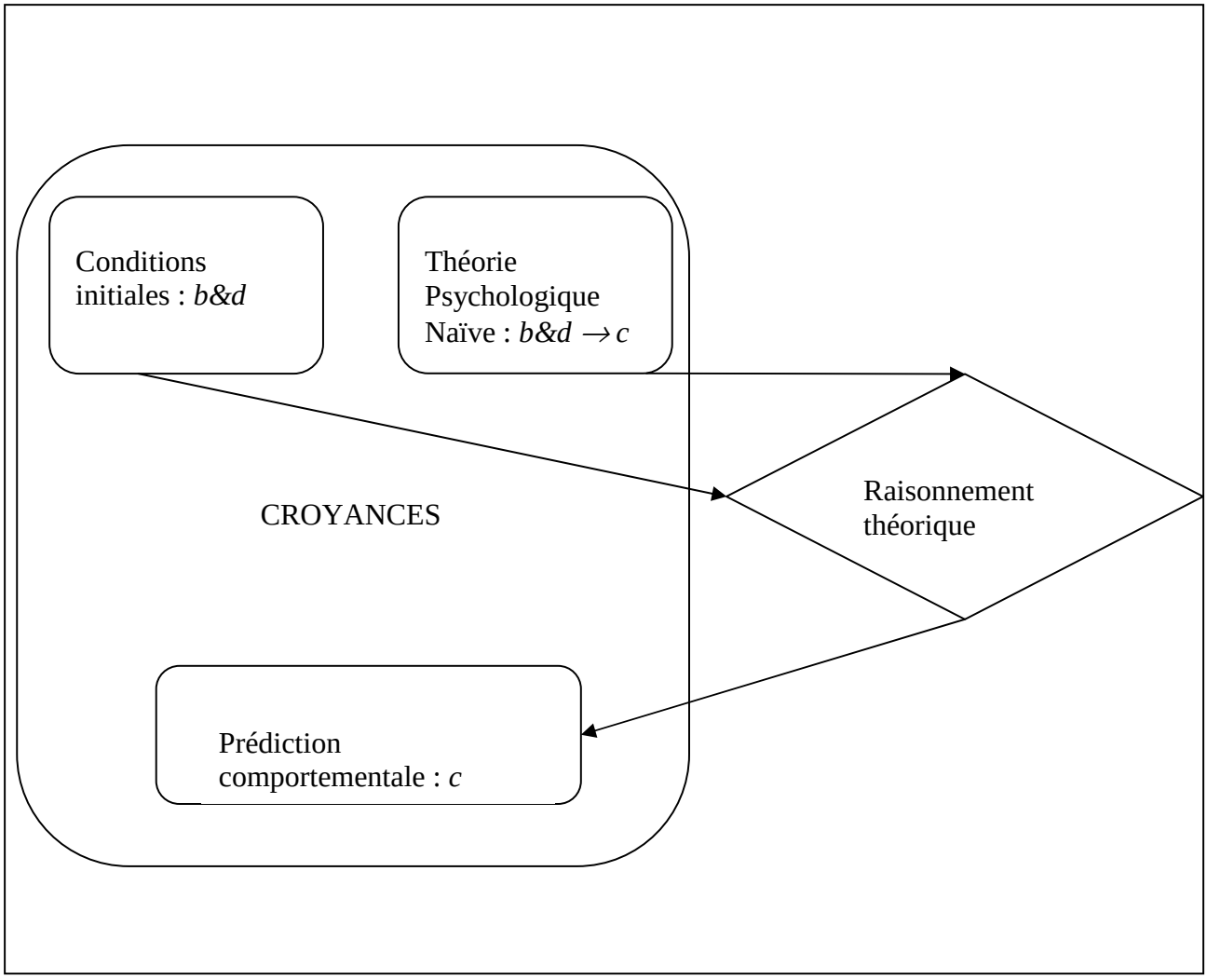

FIGURE 2. Mentalisation prédictive selon la théorie « théorique »[24].

(2) Selon la *théorie de la simulation*, la mentalisation repose essentiellement sur des procédures liées à la simulation mentale (Gordon 1986, Heal 1986, Goldman 1989, voir Davies & Stone 1995). Pour préciser ce qu'on entend par simulation mentale, on peut s'appuyer sur une analogie avec la simulation *physique*. Supposons que l'on veuille prédire le comportement d'un avion dans telles ou telles conditions de vol. Deux options sont concevables. La première option, la prédiction « théorique », consiste à considérer une théorie qui relie les principales variables en jeu, à spécifier les conditions initiales et à calculer la prédiction qui nous intéresse. La seconde option, la prédiction par simulation, est toute différente. Il pourra s'agir, par exemple, de construire un modèle réduit de l'avion et de placer

[24] Les FIGURES 2, 3 et 4 sont inspirées de celles de Stich & Nichols (1992) et Goldman (2006).

ce modèle réduit dans une soufflerie convenablement paramétrée. La prédiction sera alors établie en observant ce qu'il advient du modèle réduit dans ces conditions. On voit bien les avantages potentiels des simulations dans le monde de l'aéronautique comme dans le monde du mental : elles permettent de faire des prédictions en faisant l'économie de long calculs et ne nécessitent de passer par une théorie élaborée. Pour ce faire, les ingénieurs aéronautiques des maquettes et reproduisent dans des souffleries des environnements tels que la situation de la maquette dans la soufflerie soit le plus semblable possible à la situation de l'avion en vol dans des conditions données. Mais quel est l'analogue des maquettes et de la soufflerie dans le cas du mental ? Selon la théorie de la simulation, le « modèle réduit » sur lequel on procède à la simulation n'est rien d'autre que le prédicteur lui-même : il n'est pas nécessaire de construire une maquette, car chacun nous dispose précisément *pour lui-même* de mécanismes permettant d'engendrer des comportements à partir de croyances et de désirs. Supposons ainsi que l'architecture cognitive d'un individu ait à peu près la structure représentée à la FIGURE 3 : l'individu est doté de croyances et de désirs sur la base desquels, après avoir mené un raisonnement pratique, il arrête une décision. Cette décision engendre ensuite un comportement.

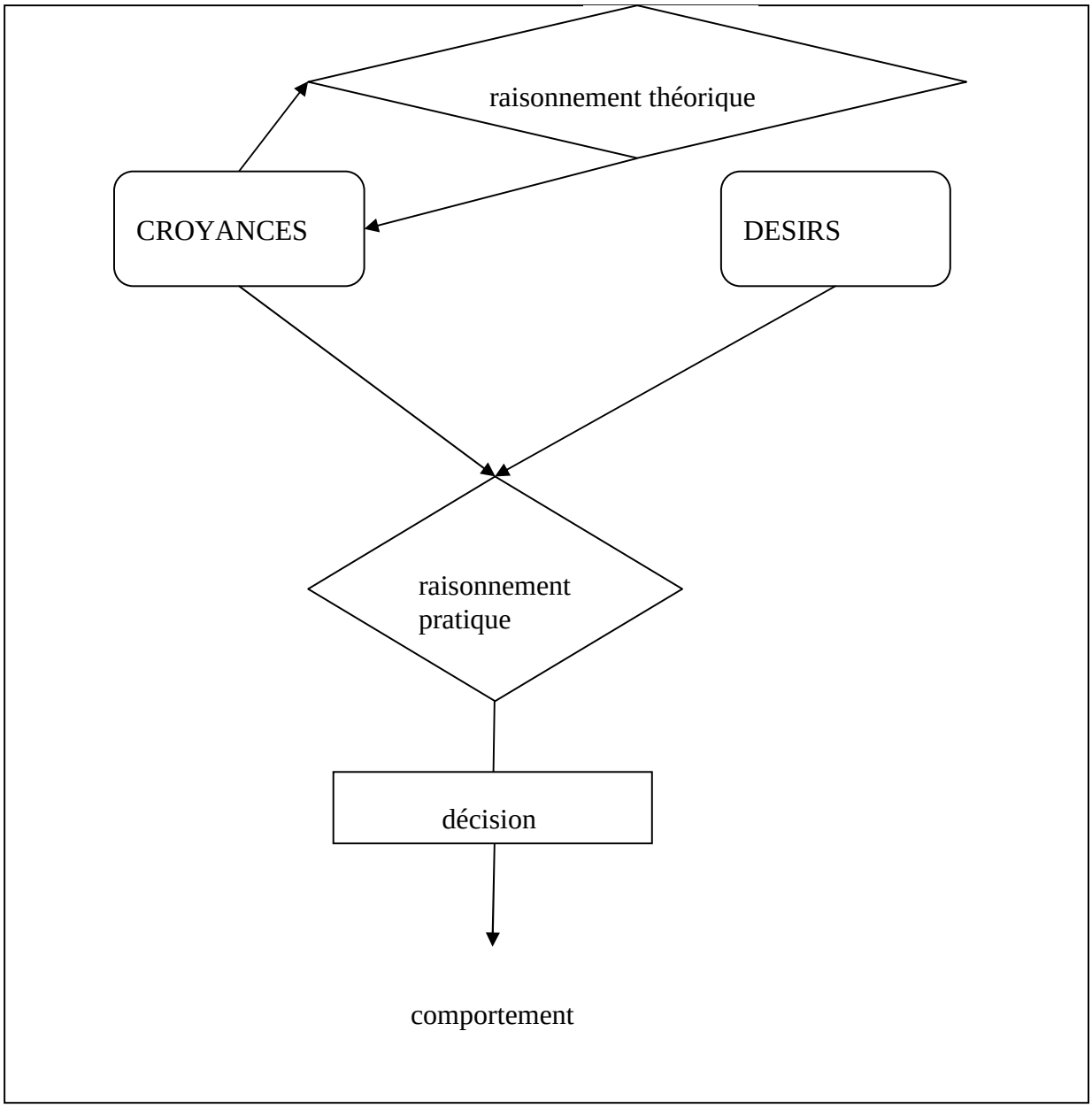

FIGURE 3. Architecture cognitive mobilisée dans l'action.

On peut alors se représenter la mentalisation prédictive selon la théorie de la simulation de la manière suivante : le prédicteur « fait semblant »[25] d'avoir les croyances *b* et les désirs *d*, il opère un raisonnement pratique sur *b* et sur *d* comme il le ferait, autant que possible, s'il avait réellement ces croyances et ces désirs. L'issue de ce raisonnement est une décision, celle d'adopter le comportement *c*. Au lieu de mettre en œuvre le comportement *c*, le prédicteur considère *c* comme sa prédiction de ce que fera l'agent-cible.

[25] Pour une discussion de cette notion, voir notamment Goldman (2006), pp. 46 et suivantes.

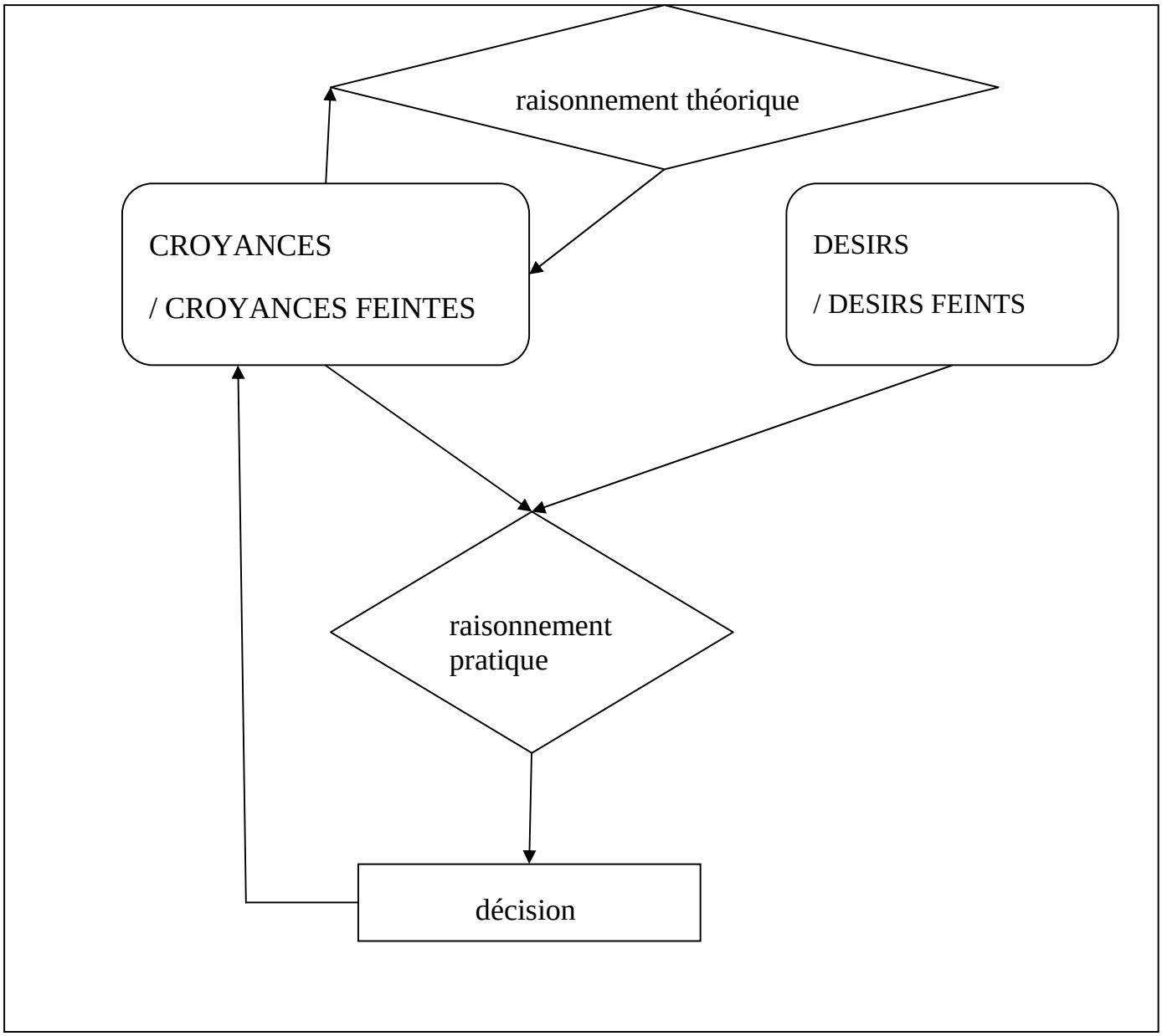

FIGURE 4. Mentalisation prédictive selon la théorie de la simulation.

La simulation mentale peut se concevoir comme un usage « hors-ligne » de ce système complexe de prise de décision. Au lieu de recevoir en entrée les authentiques croyances et désirs de l'agent, le système reçoit certaines croyances et certains désirs simulés. Le système est ensuite laissé à son fonctionnement normal, mais au lieu de déboucher sur un comportement, il s'arrête à la décision d'entreprendre le comportement, et cette décision est alors reversée dans le système effectif de croyances de l'agent, sous la forme d'une prédiction concernant le comportement de l'individu dont on avait testé hors-ligne les croyances et les désirs pertinents.

4.2 La simulation : théories et arguments

Nous n'avons fait que présenter une esquisse des théories de la simulation, une esquisse qui, à vrai dire, n'est probablement pas partagée par *tous* les partisans de la simulation. Il y a en effet des divergences importantes à l'intérieur des théories de la simulation[26]. Quant au *statut* de la théorie de la simulation, tout d'abord. Pour la plupart, en effet, la théorie de la simulation est une théorie empirique (ou un embryon de théorie empirique) dont l'objectif est d'expliquer la mentalisation, mais certains, comme J. Heal, soutiennent qu'il s'agit plutôt d'une hypothèse *a priori*[27]. Le *domaine* de la théorie de la simulation est une autre question délicate. Premièrement, quel type de mentalisation peut être expliqué par la théorie de la simulation ? Dans la section précédente, nous avons évoqué la *prédiction*. La simulation est-elle également appelée à rendre compte de l'*explication*, et en particulier de l'explication de l'action à partir des attitudes propositionnelles de l'agent ? Certains considèrent que l'usage de la simulation pour l'explication est peu plausible. D'autres soutiennent que la simulation joue également un rôle important dans l'explication, mais qu'elle est alors combinée avec une composante « théorique »[28] : cette dernière a la fonction de concevoir des états mentaux qui pourraient expliquer l'action, l'hypothèse étant ensuite « testée » par simulation. Par exemple, si Karl voit Kurt se diriger vers l'interrupteur, il peut d'abord formuler des hypothèses qui prennent la forme de paires croyances-désirs et voir si ces paires aboutiraient bien au comportement observé. Il y a une autre dimension importante dans la question du domaine de la simulation, c'est celle du *type* d'états mentaux qui peuvent être pris en charge par simulation : s'agit-il exclusivement des attitudes propositionnelles, ou également, par exemple, des émotions ? A supposer que différents types d'états mentaux soient pris en charge par simulation, leur prise en charge s'effectue-t-elle de manière analogue ou par des procédures différentes[29] ? Le débat entre théories de la simulation et théories « théoriques » a donné lieu à une très vaste littérature. Nous allons nous contenter de mentionner certains des arguments avancés en faveur des théories de la simulation.

(1) Les partisans de la théorie de la simulation mettent souvent en avant la *simplicité* des processus simulationnistes : pourquoi supposer un vaste ensemble d'hypothèses relatives à la vie mentale d'autrui alors qu'il nous suffit d'extrapoler à partir de la nôtre ? P. Harris (1992) développe l'analogie avec les intuitions grammaticales : si l'on demande à quelqu'un de dire si un locuteur de sa propre langue jugera une certaine suite de signes grammaticale ou non, l'hypothèse la plus probable est que notre sujet-prédicteur se demandera si *lui* trouve la suite de signes grammaticale et ensuite projettera son jugement sur le locuteur-cible. Il serait bien moins plausible de supposer que le sujet-prédicteur dispose d'une théorie sur la façon dont autrui porte des jugements grammaticaux.

---

[26] Voir Goldman (2006), paragraphe 2.6.

[27] Heal (1998) parle de « co-cognition » pour désigner le type de simulation qu'elle défend et d' « off-line simulation » pour les théories empiriques de la simulation.

[28] L'analogie avec la simulation aéronautique prend ici tout sens, puisque la construction d'une maquette d'avion et l'utilisation d'une soufflerie de manière à obtenir des conditions similaires aux conditions de vol nécessitent une bonne dose de théorie, sans que le gain qu'il y ait à simuler plutôt qu'à tout prédire sur la base de théories en soit le moins du monde remis en cause.

[29] Goldman (2006) distingue par exemple la simulation de « bas niveau », qui prend notamment en charge les émotions, de la simulation de « haut niveau » qui concerne les attitudes propositionnelles.

(2) L'*argument de la précision*, proposé par Stich & Nichols (2003), part du constat que nos prédictions naïves sont dans certains domaines comme celui des inférences inductives particulièrement performantes. Ce peut être le cas même quand la situation est relativement inédite et que par conséquent le prédicteur n'est vraisemblablement pas en mesure d'avoir pu former des généralisations empiriques pertinentes.

(3) La théorie de la simulation s'est appuyée dès ses débuts sur les *données développementales*[30]. Tout d'abord, certaines étapes décisives dans le développement de la mentalisation ont lieu au même âge (entre deux et trois ans) que le développement de la capacité de jouer à être une autre personne. Ensuite, les étapes du développement de la mentalisation semblent bien s'accorder avec la théorie de la simulation. La tâche de la croyance fausse est dans certes perspectives l'un des scénarii expérimentaux les plus discutés (Wimmer & Perner, 1983). La tâche consiste essentiellement à attribuer une croyance à un individu-cible. Le sujet voit l'individu-cible observer une scène qui doit produire chez l'individu-cible la croyance *b*. Max voit sa maman ranger une barre de chocolat dans le placard, Max croit que (b) la barre de chocolat est dans le placard. A l'insu de l'individu cible, l'expérimentateur change la scène de sorte que la croyance *b* devient fausse : la maman de Max change la barre de chocolat et la met dans un tiroir à l'insu de Max. Le sujet est informé de ce changement, et on lui demande d'attribuer une croyance à l'individu-cible. Typiquement, l'enfant de moins de trois ans attribue la croyance correcte, disons *b'* et non pas la croyance devenue incorrecte, *b* : l'enfant dit que Max croit que la barre de chocolat est dans le tiroir. C'est entre trois et cinq ans que les enfants se mettent à dire que Max croit que la barre est dans le placard, attribuant (correctement) la croyance fausse *b*. Du point de vue de la théorie de la simulation, on peut rendre compte des performances des jeunes enfants comme d'une incapacité à correctement *neutraliser* leurs propres états mentaux qui sont projetés sur autrui trop massivement – en l'occurrence, sans tenir compte d'une différence d'information cruciale entre soi et l'individu-cible. Les théories « théoriques » ne sont pas sans réponse face à la tâche de la croyance fausse[31]. Elles peuvent affirmer que la Théorie Psychologique Naïve se met en place progressivement et qu'un jeune enfant n'a tout simplement pas encore la *TPN* de l'adulte. C'est pourquoi les tenants de la simulation cherchent désormais des arguments plus fins. Par exemple, Goldman (2006) en appelle à une série d'expériences psychologiques qui montrent une corrélation significative entre échecs aux tâches de croyances fausses et capacités d'inhibition des réponses « naturelles ».

(4) Un autre type d'arguments provient du champ psychopathologique. Des liens ont été proposés entre l'*autisme* (ou du moins certaines formes d'autisme) et certains déficits de mentalisation (Baron-Cohen, Fritch & Leslie, 1985). Or, les enfants autistes manifestent un déficit marqué à entrer dans des jeux d'imagination. Ce déficit conjoint dans la mentalisation et dans les jeux d'imagination est attendu par les théories de la simulation : si l'attribution d'états mentaux repose sur notre capacité à simuler autrui, une forme d'imagination apparaît comme nécessaire au fonctionnement des mécanismes de simulation.

---

[30] Sur le rôle des données développementales dans le débat entre théorie théorique et théorie de la simulation, voir Harris (1992).

[31] Davies (1994) : « It seems fair to say that both the advocates of the theory theory […] and the friends of mental simulation…can provide acceptably principled accounts of the developmental data from the third-person false belief task. »

(5) Il nous faut également mentionner des développements plus récents qui reposent sur certaines avancées des neurosciences cognitives. Des défenseurs de la théorie de la simulation comme Gallese & Goldman (1998) ont cherché à tirer parti de la découverte des *neurones miroirs* chez le singe macaque. Certains neurones du cortex prémoteur sont en effet activés à la fois quand une action est *exécutée* et quand elle est *observée* (Rizzolati & *alii*, 1996). Gallese & Goldman (1998) estiment que les neurones miroirs sous-tendent la simulation, ou sont parmi les précurseurs des processus de simulation. Plus précisément, l'activation de ces neurones correspondrait à l'explication d'un comportement par une intention d'agir ou à l'identification de ce comportement comme action d'un certain type. Goldman (2006) a raffiné sa position et distingue la mentalisation de « bas niveau » de la mentalisation de « haut niveau ». Les systèmes miroirs seraient impliqués dans la première, mais toute activité de mentalisation ne se base pas sur des processus miroirs – ce qui serait peu plausible du point de vue neuroanatomique[32].

(6) Un phénomène particulièrement important dans la discussion sur les théories de la simulation est celui des *biais égocentriques*, c'est-à-dire des situations où le prédicteur ou l'attributeur se trompe sur l'individu-cible en projetant « *trop* » de sa propre condition mentale sur sa cible. Notons que les biais égocentriques incluent les cas où le prédicteur se trompe sur lui-même, dans le passé ou dans le futur (mentalisation intrapersonnelle). Le biais égocentrique est bien sûr à l'œuvre dans la tâche de la croyance fausse. Mais les psychologues ont montré qu'il était également répandu chez l'adulte normal lors de tâches très diverses, qui peuvent mettre en jeu les connaissances, les préférences, les décisions[33] ou les sentiments. Goldman (2006) fait des biais égocentriques l'un de ses arguments principaux *pour* la théorie de la simulation, du moins en ce qui concerne la mentalisation de haut niveau. A la base de la théorie de la simulation, il y a l'idée que, lors de la mentalisation, le prédicteur ou l'attributeur corrige partiellement ce qu'il obtient par simulation sur la base de sa perception des différences entre lui et l'individu-cible. On doit donc s'attendre à des erreurs systématiques induites par une *correction imparfaite* de la simulation, et il s'agit là d'une manière très plausible d'interpréter biais égocentriques. Par exemple, les sujets répondent différemment à la question de savoir si des randonneurs égarés sans eau ni nourriture regretteraient davantage de manquer d'eau ou de nourriture selon qu'ils se trouvent dans un état « normal » où qu'ils viennent de faire 20 minutes d'exercice physique intense (Van Boven & Loewenstein, 2003). Si nos prédictions étaient guidées par une théorie, on n'attendrait pas une telle variation, celle-ci devient plausible si l'on considère que les sujets ne parviennent pas à s'abstraire complètement de leur condition présente lors de prédiction par simulation : leur état actuel de soif influence leur prédiction concernant le désir d'autrui de boire dans des circonstances données. Notons cependant que les biais égocentriques ont été utilisés 'en sens inverse' par Stich & Nichols (2003), qui font de l'existence de certains biais égocentriques leur argument principal *contre* la théorie de la simulation. L'idée est que les échecs de la mentalisation intrapersonnelle en particulier semblent incompatibles avec la théorie de la simulation, puisque l'on est alors dans un cas où un *même* mécanisme, avec les mêmes données en entrées, donne deux résultats différents. L'interprétation de Goldman nous semble néanmoins préférable. S'agissant de la mentalisation intrapersonnelle, même si c'est le même mécanisme qui est employé, il est employé dans un cas avec des croyances et des désirs authentiques,

---

[32] Pour une critique de l'usage des neurones miroirs par les théoriciens de la simulation, voir Jacob & Jeannerod (2005) et Jacob (2008).

[33] Voir *infra* la discussion de l'*effet de dotation*.

dans un autre cas avec des croyances et des désirs feints. La difficulté à ajuster croyances et désirs feints et la prégnance de nos croyances et désirs actuels peut suffire à expliquer nos erreurs de mentalisation intrapersonnelle (autrement dit, dans l'argument de Stich et Nichols, c'est l'hypothèse selon laquelle les entrées sont exactement les mêmes qui est incorrecte). Soulignons enfin que l'analyse proposée par van Boven et Loewenstein eux-mêmes de nos prédictions sur les préférences et les décisions d'autrui est extrêmement proche de conceptions simulationnistes comme celles de Gordon (1986) : pour prédire ce que *fera* autrui, on prédit d'abord ce que l'on *ferait* dans une situation analogue et on corrige la prédiction pour tenir compte d'éventuelles différences entre soi et autrui[34]. Et cette analyse est étayée par une analyse statistique qui met en avant le rôle médiateur de la prédiction par le sujet de ses propres préférences dans sa prédiction de celles d'autrui.

(7) S'ajoutent enfin des données introspectives qui ont été rassemblées par une littérature peu reliée au débat sur la nature de la mentalisation, la littérature qui porte sur le rôle des émotions dans la formation des préférences et dans la décision. Ainsi la majorité des sujets interrogés par Van Boven & Loewenstein (2003) sur la façon dont ils ont procédé pour attribuer des sentiments et des émotions à des personnes décrites dans un scénario affirment en substance qu'ils se sont demandé comment *eux* réagiraient dans une situation comme celle décrite par le scénario. Ces données sont conformes aux intuitions fondamentales des théoriciens de la simulation.

Le débat entre théories « théoriques » et théories de la simulation est loin d'être clos, et l'objet du présent article n'est pas de le trancher. Il n'en est pas moins vrai que l'on assiste aujourd'hui à une forme de rapprochement entre certains des protagonistes, rapprochement qui se manifeste par l'adoption de *théories hybrides* de la mentalisation (Goldman 2006, Stich & Nichols 2003). Dans ce qui suit, nous supposerons que, comme le suggèrent les arguments listés dans le présent paragraphe, une forme modérée de théorie de la simulation est correcte : l'activité de mentalisation repose *au moins en partie* sur des processus de simulation de l'individu-cible.

4.3 Simulation et rationalité

Revenons désormais au scénario de Linda. Après avoir écouté le scénario, Kurt, qui incarne la réponse modale, estime apparemment plus probable que Linda soit banquière *et* militante féministe plutôt que (simplement) banquière. C'est du moins ce que manifestent les réponses qu'il donne au questionnaire du psychologue. Attribuer ces croyances à Kurt viole bien sûr le principe de charité, puisque de telles croyances désobéissent aux principes élémentaires des probabilités. Y a-t-il là quelque chose qui défie la compréhension ? Il nous semble que cette attribution est naturelle. A vrai dire, elle est d'autant plus naturelle que l'attributeur est lui-même susceptible de se montrer irrationnel dans les mêmes conditions. Supposons par exemple que l'attributeur a déjà passé le test de Linda et qu'il a donné les mêmes réponses que Kurt lors du test. Quoi de plus naturel pour lui que de penser que Kurt tombera dans la

[34] Van Boven, Loewenstein & Dunning (2005) : « the results of these two experiments support the thesis that emotional perspective taking entails two judgments : a prediction of one's own preferences and decision in a different emotional situation, and an adjustment of this prediction to accommodate perceived differences between self and others. » Voir aussi Van Boven & Loewenstein (2003).

même piège ? Si l'attributeur avait du faire une *prédiction* sur les réponses qu'allaient donner Kurt au scénario, il aurait probablement prédit que Kurt serait lui aussi irrationnel et ses prédictions auraient probablement été correctes. Tout ceci est attendu du point de vue simulationniste.

Autrement dit, il semble que dans les contextes où l'attributeur peut s'imaginer ne pas être rationnel (c'est *a fortiori* le cas quand, comme on l'a supposé à titre d'exemple, l'attributeur a dans le passé donné la réponse modale qui viole les principes probabilistes), il n'a pas de difficulté particulière à attribuer à autrui des croyances irrationnelles. Dans une perspective simulationniste, si une violation de l'hypothèse de rationalité est simulable, alors elle reste compréhensible. Bien sûr, on doit s'attendre à de vastes zones de convergence entre hypothèse de rationalité et simulation. Mais ce que suggère notre analyse du scénario de Linda, c'est que lorsque hypothèse de rationalité et simulation divergent, la mentalisation se porte du côté de la simulation, pas de l'hypothèse de rationalité. Si ce qui précède est correct, le principe de charité ne vaut pas lorsque la compréhension d'autrui repose sur des mécanismes de simulation qui ne respectent pas eux-mêmes les contraintes de rationalité.

Notre critique du principe de charité passe par une explication simulationniste de notre capacité à comprendre ou à prédire des conduites irrationnelles. Cette explication pourrait être testée empiriquement. Il semble notamment que nous prédisions une corrélation positive entre la propension d'un individu donné à se conduire de manière irrationnelle dans telle ou telle situation (par exemple, dans la situation où il est soumis au test de Linda) et sa propension à attribuer ou à prédire chez autrui dans des situations analogues des croyances irrationnelles. La corrélation statistique mise en évidence par van Boven et Loewenstein (2003) que nous avons évoquée porte sur des *prédictions* nous concernant et des prédictions concernant autrui. Dans notre optique, il serait intéressant d'étudier, pour des scenarii propices à l'irrationalité comme le scénario de Linda, la corrélation entre nos comportements effectifs (produits par notre mécanisme décisionnel fonctionnant « on-line ») et nos prédictions concernant autrui (produits par notre mécanisme décisionnel fonctionnant « off-line » si la théorie simulationniste est correcte).

Revenons au principe de charité. Nous avions vu dans l'examen de ses justifications que Davidson fait appel à nos intuitions communes sur ce que c'est que « comprendre » ou « rendre intelligible » un épisode mental ou comportemental. Ces intuitions communes ne sont rien d'autre que les normes épistémiques qui encadrent la mentalisation. La discussion qui précède suggère que *l'hypothèse de rationalité n'est pas requise par ces intuitions communes*. On pourrait alors être tenté de se demander si la simulation ne fournit pas un principe qui serait la contrepartie simulationniste du principe de charité, c'est-à-dire un principe qui serait requis pour la compréhension d'autrui. En substance, ce principe affirmerait que l'on ne peut comprendre un épisode mental ou comportemental s'il n'est pas *simulable*.

Un tel principe ne nous semble ni guider nos intuitions communes de compréhension, ni être légitime. Nous reviendrons sur cette seconde affirmation dans la section suivante, qui porte sur les implications épistémologiques de notre discussion pour les sciences de l'homme. Concentrons-nous pour le moment sur nos intuitions communes de compréhension. Nombre de phénomènes défient nos performances de simulation. Il y a bien sûr les biais égocentriques que nous avons déjà évoqués – par exemple les prédictions incorrectes engendrées par la différence d'état viscéral entre le prédicteur et sa cible. Songeons également à certaines

pathologies comme le syndrome de Capgras où l'individu croit que ses proches sont en réalité des imposteurs déguisés. Le point crucial est qu'il existe des explications de ces phénomènes. Ces explications ne nous mettent pas en position de simuler le phénomène, mais il nous semble parfaitement naturel de considérer qu'elles nous permettent de *comprendre* le phénomène. Nous en concluons que notre notion intuitive de compréhension est éminemment *plastique* : elle ne se restreint certainement pas à la simulation en première personne[35]. C'est d'ailleurs ce que l'on doit attendre si l'on adopte, comme nous serions volontiers prêts à le faire, une forme hybride de théorie de la simulation pour rendre compte de la mentalisation.

5 Perspectives épistémologiques

Nous avons vu que Davidson tire des implications épistémologiques essentielles du principe de charité : le principe de charité introduirait une différence fondamentale entre les sciences de la nature et les sciences de l'homme. Nous voulons achever notre examen du principe de charité en revenant sur ces implications épistémologiques. Nous n'avons pas l'ambition de proposer une conception aussi aboutie que celle de Davidson, mais nous voudrions mettre en perspective les différents éléments que nous avons fait valoir dans nos analyses précédentes. L'objet de cette section est ainsi de présenter, dans l'attente d'un tableau plus complet, l'esquisse d'une articulation systématique des concepts et théories que nous avons abordés.

5.1 Compréhension et simulation

L'idée de simulation est manifestement très proche de certaines conceptions célèbres en épistémologie des sciences de l'homme (et en particulier des sciences historiques) que l'on appellera par commodité les théories de la Compréhension (ou du *Verstehen*). Ces théories sont anti-naturalistes, au sens que le terme prend en épistémologie des sciences humaines et sociales : elles soutiennent qu'il existe une différence fondamentale entre les sciences de la nature et les sciences de l'homme. Dans les sciences de l'homme, l'objectif épistémique serait en effet essentiellement différent de celui que l'on poursuit dans les sciences de la nature : on cherche bien à comprendre l'objet étudié, mais la nature de l'objet étudié, à savoir l'homme, fait que le *type* de compréhension en jeu est spécifique. Nous désignerons ce type de compréhension comme la « Compréhension ». On la trouve désignée parfois comme «reconstitution (*reenactment*)», « compréhension empathique », « imagination sympathique » ou encore « compréhension en première personne ».

La Compréhension, qui serait propre aux sciences de l'homme, est souvent caractérisée en termes simulationnistes : il s'agirait de « se mettre à la place » des individus auxquels on s'intéresse pour « redécouvrir leurs pensées » (Collingwood, 1946).

> *« Comment l'historien discerne-t-il les pensées qu'il essaie de découvrir ? Il n'y a qu'une seule manière de procéder : en les pensant de nouveau dans son propre esprit…L'histoire consiste entièrement dans la reconstitution de la pensée passée. »* (Collingwood, 1946)

---

[35] Von Wright (1971, p. 6) admet lui aussi la plasticité de nos intuitions communes sur la compréhension: « L'usage ordinaire ne fait pas de distinction stricte entre les mots 'expliquer' et 'comprendre'. Pratiquement toute explication, qu'elle soit causale, téléologique ou d'une autre sorte, peut être dite améliorer notre compréhension. » (nous traduisons)

> *« C'est seulement en se mettant soi-même dans la position de l'agent que l'on peut comprendre pourquoi il a fait ce qu'il a fait »* (Dray, 1957)

Comme l'affirme Stueber (2006), « le regain d'intérêt pour l'empathie rappelle l'enthousiasme des philosophes de l'histoire et des sciences sociales de naguère, qui considéraient l'empathie comme la méthode première pour acquérir des connaissances objectives sur les autres agents. » Notre proposition est de concevoir les théories de la simulation comme des explications psychologiques de cette faculté de Compréhension que les épistémologues anti-naturalistes placent au cœur des sciences de l'homme. Remarquons qu'il n'est pas étonnant que ces derniers accordent une importance toute particulière aux sciences historiques : on y trouve des attributions d'états mentaux et des explications mentalistes qui sont analogues, sinon identiques, à la mentalisation naïve. Ce que notre proposition suggère, c'est que l'on Comprend au sens où l'entendent les anti-naturalistes lorsque l'on arrive à simuler.

Mais il faut immédiatement ajouter qu'une approche naturaliste de la Compréhension suggère également qu'il n'y a guère de raison de *contraindre* les sciences de l'homme à en rester à la sphère du « simulable ». Si la simulation est notre façon *spontanée* de construire des explications et des prédictions mentales, il n'y a guère de raison de supposer que des théories *scientifiques* du comportement humain ne doivent pas s'en émanciper. Nous disposons d'une théorie physique naïve, qui nous fait attribuer certaines propriétés aux corps physiques et nous fait attendre certaines régularités naturelles. Mais les recherches en physique ne sont pas contraintes par les principes de cette physique naïve. Que nous concevions les corps physiques comme des choses pleines et denses n'a pas empêché le modèle atomique de la matière de s'imposer. Pourquoi en irait-il autrement dans le cas des sciences de l'homme ? La psychologie scientifique, et les sciences de l'homme en général, ne sont pas contraintes par les contenus ou les modalités de la psychologie naïve. Nous avons d'ailleurs avancé, à la fin de la section précédente, que notre mentalisation elle-même n'était pas rigidement contrainte par une exigence de « simulabilité ».

Notre attitude vis-à-vis de la Compréhension se rapproche de celle de Hempel (1942) qui lui assignait le statut d'*heuristique*. Hempel examine en effet le rôle de ce qu'il appelle *« empathetic understanding »* dans le cadre d'une discussion sur la nature des explications historiques. Il défend la thèse générale selon laquelle les explications en histoire sont homogènes aux explications dans les sciences de la nature. Toutes deux reposent essentiellement sur l'utilisation de *lois* générales à partir desquelles les faits à expliquer sont déduits, moyennant les conditions initiales appropriées. Hempel soutient que la Compréhension ne saurait se substituer à l'explication proprement dite[36], précisément parce qu'elle ne repose sur l'utilisation de lois générales. Ce que nous retenons n'est pas tant l'insistance sur les lois que l'argument qu'Hempel oppose, dans ce contexte, à l'idée selon laquelle la Compréhension serait indispensable à l'explication historique :

> *« Un historien peut, par exemple, ne pas être capable de se mettre dans la peau d'une personnalité historique paranoïaque, and pourtant il peut tout à fait être*

---

[36] « This method of empathy is, no doubt, frequently applied by laymen and by experts in history. But it does not in itself constitute an explanation; it rather is essentially a heuristic device; its function is to suggest certain psychological hypotheses which might serve as explanatory principles in the case under consideration » (1942, p. 44)

> *capable d'expliquer certaines de ses actions; en particulier en se rapportant aux principes de la psychopathologie. »* (Hempel, 1942, p.44)

Autrement dit, la Compréhension ne constitue pas une limite *a priori* aux possibilités théoriques des sciences de l'homme. Bien au contraire, une partie au moins des sciences de l'homme (dans l'exemple de Hempel, la psychopathologie) visent à développer des théories qui permettent d'expliquer ce qu'on ne peut pas Comprendre.

5.2 Les trois domaines

Le principe de charité procède à une sorte d'*écrasement* des trois domaines suivants :

1. Les théories de rationalité
2. Les théories scientifiques du mental (de la cognition et du comportement)
3. La mentalisation

Les principes de rationalité sont en effet pour Davidson les hypothèses fondamentales des théories générales du mental, comme la Théorie Unifiée que nous avons évoquée précédemment, et ce sont également des formulations explicites des conditions de possibilité de notre compréhension d'autrui. A vrai dire, ce sont les hypothèses fondamentales des théories du mental précisément *parce que* ce sont les conditions de possibilité de notre compréhension d'autrui. Nous pensons au contraire que ces trois domaines ne coïncident pas et *n'ont pas à* coïncider. Les théories de la rationalité sont des théories normatives, les théories scientifiques de la cognition et du comportement sont des théories descriptives et la mentalisation relève de l'épistémologie naïve. Nous allons désormais illustrer et étayer notre affirmation en considérant la non-coïncidence entre les trois paires possibles de domaines. Nous serons assez brefs concernant les deux paires (1,3) et (1,2) que nous avons déjà implicitement traitées précédemment. Nous détaillerons davantage la paire (2,3).

(1,3) THEORIES DE LA RATIONALITE ET MENTALISATION

Que la compréhension et la prédiction naïves ne coïncident pas avec les principes de rationalité, c'est ce qui ressort des analyses concluant la quatrième section, et en particulier des discussions du scénario de Linda. S'il nous fallait, par exemple, prédire les réponses d'un sujet au scénario de Linda, nous nous fonderions probablement sur la réponse qui *nous* paraît aller de soi à la première lecture du scénario, à savoir celle qui implique une violation des axiomes des probabilités.

(1,2) THEORIES DE LA RATIONALITE ET THEORIES SCIENTIFIQUES DU MENTAL

Que les principes de rationalité et les théories descriptives du mental ne coïncident pas, c'est ce dont témoigne à notre sens l'immense littérature psychologique sur les déviations par rapport aux principes de rationalité, tant dans le domaine de l'inférence « théorique » (dont le scénario de Linda fait partie) que dans celui de la rationalité pratique. Les travaux de Kahneman et Tversky, qui couvrent les deux domaines, soutiennent avec particulièrement de force cette affirmation. Une des tâches de la psychologie scientifique est précisément d'identifier et d'expliquer les biais de raisonnement qui marquent l'écart entre la manière dont nous raisonnons et la manière dont nous devrions raisonner.

Passons enfin aux relations entre mentalisation et théories scientifiques du mental. La compréhension et la prédiction naïves ne coïncident pas non plus avec les théories descriptives du mental. Considérons par exemple l'*effet de dotation* (*endowment effect*, Thaler, 1980) mis au jour en théorie de la décision comportementale et mobilisé dans le débat sur la mentalisation par Stich & Nichols (2003). L'effet de dotation est le phénomène suivant : la valeur (monétaire du moins) que les individus attribuent à un objet varie selon qu'ils possèdent cet objet ou non. Typiquement, on attribue une valeur significativement supérieure à un objet quand on le possède. Il s'agit d'un phénomène extrêmement robuste. Loewenstein & Adler (1995) ont montré que les individus ne prédisent pas correctement l'effet de dotation – même quand il s'agit de *leur propre* comportement ! Ainsi, la valeur moyenne prédite par les sujets testés s'élèvent à 3.73 $ quand ils ne possèdent pas l'objet à évaluer. La valeur moyenne des mêmes sujets quand, un peu plus tard, leur est donné l'objet à évaluer est de 5.40$ (Loewenstein & Adler, 1995). Ces résultats ont été étendus au cas des prédictions sur la valeur accordée par *autrui* à un objet qu'il possède. Autre exemple : Van Boven, Loewenstein & Dunning (2005) parlent d' « illusion du courage » pour désigner le fait que, en général, nous surestimons la disposition d'autrui à s'engager dans des situations socialement embarrassantes. L'illusion du courage est conçue comme un cas particulier du « fossé empathique chaud / froid » qui consiste en ce que les individus, lorsqu'ils *ne sont pas* dans certains états psychologiques (par exemple des états émotionnellement intenses) sous-estiment l'impact sur leurs préférences et décisions du fait d'être dans ces états psychologiques. Tout comme dans le cas des écarts avec les normes de rationalité, une des tâches de la psychologie scientifique est d'identifier ces écarts avec nos prédictions spontanées.

Ce qui précède ne doit pas cacher le fait qu'*il y a de larges aires de recouvrement entre les trois domaines.* Dans bien des situations, les états mentaux que les individus ont effectivement, ou les comportements qu'ils adoptent effectivement, obéissent aux principes de rationalité et sont aisément prédictibles ou explicables par mentalisation naïve. Il n'en reste pas moins, si notre critique du principe de charité est correcte, qu'il n'y a pas de raison *a priori* de supposer que ces trois domaines se recouvrent systématiquement, et qu'on doit au contraire considérer que le degré auquel ils se recoupent ou non est une question scientifique de plein droit. L'intérêt de cette question apparaît de manière particulièrement vive dans une optique évolutionniste : si nos capacités de raisonnement et de décision sont le fruit de la sélection naturelle, on peut supposer que ces capacités sont sinon les plus adaptées, du moins raisonnablement adaptées à notre environnement. Comment expliquer alors que ces capacités ne fonctionnent pas selon ce que les théories de la rationalité nous indiquent être les meilleurs modes de raisonnement ? Faut-il en déduire que les principes de rationalité ne sont pas des manières optimales de raisonner, étant donné notre environnement et nos contraintes biologiques ?

5.3 Questions ouvertes

Le tableau que nous venons d'esquisser laisse de nombreuses autres interrogations en suspens.

La première d'entre elles porte sur ce que nous avons appelé les théories scientifiques du mental et sur leur relation avec la mentalisation naïve. Les sciences cognitives et les sciences

de la décision ont connu ces dernières décennies des avancées remarquables, aussi bien empiriques que théoriques, mais il est clair qu'on ne dispose pas aujourd'hui d'une théorie descriptive unifiée du comportement et des états mentaux qui serait analogue à la Théorie Unifiée que Davidson propose. La question qui se pose alors est celle de savoir ce qui restera de notre psychologie naïve dans les futures théories scientifiques du mental.

Cette première question en appelle immédiatement une seconde : à supposer que nous disposions de théories scientifiques de la cognition et du comportement abouties, faut-il que ces théories se substituent, dans les différentes sciences humaines, aux différents modèles ou hypothèses (éventuellement, tacites) qui sont faites sur le comportement et la cognition ? Et si oui, comment ? Il est clair que cette seconde question est partiellement solidaire de la première : dans les disciplines où c'est essentiellement la mentalisation naïve qui est à l'œuvre, tout dépend, en principe, de l'écart entre mentalisation naïve et théories scientifiques. Il se peut, par exemple, que la mentalisation naïve fournisse des approximations parfaitement acceptables, ou que les données nécessaires au fonctionnement explicatif ou prédictif des théories scientifiques ne soient tout simplement pas disponibles dans la discipline en question.

La première question porte sur les relations entre la psychologie naïve et la psychologie scientifique, la seconde sur les relations entre la psychologie scientifique et les autres sciences humaines. La dernière question que nous voudrions soulever est assez différente puisqu'elle porte sur les théories normatives. Nous avons considéré qu'il existe des divergences importantes entre ce que nous faisons ou pensons et ce que nous devrions faire ou devrions penser. Mais si cet écart se laisse constater, c'est d'abord parce que nous sommes capables d'expliciter certaines normes de rationalité. Se pose alors la question de savoir d'où proviennent ces normes de rationalité, et comment il est possible que ces normes divergent de nos intuitions naïves. Certains comme L.J. Cohen (1981) ont affirmé, en substance, que puisque la réponse la plus plausible à la première question est que les normes de rationalité codifient et systématisent nos intuitions naïves, il ne pouvait y avoir de divergence essentielle entre ces normes et la « compétence » psychologique des individus. Si cette thèse ne nous convainc guère, il faut admettre que la question de la formation et de l'acceptation des théories normatives est largement ouverte [37].

6 Références

Baron-Cohen S., Leslie A.M. & Frith U (1985), « Does the autistic child have a 'theory of mind'? », *Cognition*, 21(1), pp. 37–46.

Cherniak, Ch. (1985). *Minimal Rationality*, Cambridge, Mass: MIT Press.

Cohen, L.J. (1981), « Can Human Irrationality Be Experimentally Demonstrated?», *The Behavioral and Brain Sciences*, 4, 317-70.

[37] L'idée très populaire selon laquelle les normes de rationalité seraient le résultat d'un *équilibre réfléchi* fournit un premier élément de réponse à la question, mais un élément très incomplet.

Collingwood, R.G. (1946) *The idea of history*, Oxford : Oxford University Press.

Davidson, D. (1973) « Radical Interpretation » *in Inquiries into Truth and Interpretation*, Oxford : Clarendon Press, 1984.

Davidson, D. (1974a), « Belief and the Basis of Meaning », *in Inquiries into Truth and Interpretation*, Oxford : Clarendon Press, 1984.

Davidson, D. (1974b), « Psychology as Philosophy », *in Essays on Actions and Events*, Oxford : Clarendon Press, 1980.

Davidson, D. (1975), « Thought and Talk », *in Inquiries into Truth and Interpretation*, Oxford : Clarendon Press, 1984.

Davidson, D. (1976), « Hempel on Explaining Action », *in Essays on Actions and Events*, Oxford : Clarendon Press, 1980.

Davidson, D. (1980), *Essays on Actions and Events*, Oxford : Clarendon Press ; trad.fr. P. Engel, *Actions et événements*, Paris : PUF, 1993.

Davidson, D. (1985), « A New Basis for Decision Theory», *Theory and Decision*, 18(1), pp. 87-98

Davidson, D. (1987), « Problems in the Explanation of Action », *in Problems of Rationality*, Oxford : Clarendon Press, 2004, pp. 101-116.

Davidson, D. (1990a), « The Structure and Content of Truth », *The Journal of Philosophy*, 87(6), pp. 279-328.

Davidson, D. (1990b), « Representation and Interpretation », *in Problems of Rationality*, Oxford: Clarendon Press, 2004, pp. 87-100.

Davidson, D. (1991), « Three Varieties of Knowledge », *in Subjective, Intersubjective, Objective*, Oxford: Clarendon Press, 2001, pp. 193-204.

Davidson, D. (1994), « Radical Interpretation Interpreted », *Philosophical Perspectives*, 8, pp. 121-128.

Davidson, D. (1995), « Could There Be a Science of Rationality », *in Problems of Rationality*, Oxford: Clarendon Press, 2004, pp. 117-134.

Davidson, D. (1999), « Replies » *in* Hahn, L.E. (ed.), *The Philosophy of Donald Davidson*, The Library of Living Philosophers, vol. XXVII, Chicago & La Salle : Open Court

Davidson, D., Suppes, P. & Siegel, S. (1957), *Decision Making: An Experimental Approach*, Stanford: Stanford University Press

Davies, M. (1994), « The mental simulation debate » in C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness: Current Issues in the Philosophy of Mind*. Proceedings of the British Academy, vol. 83, Oxford : Oxford University Press, 1994), 99-127.

Davies, M. & Stone, T. (1995), *Mental Simulation*, Oxford : Blackwell.

Dennett, D. (1987) *The Intentional Stance*, Cambridge, MA: Bradford Books/MIT Press.

Dray, W. (1957), *Laws and Explanation in History*, Oxford: Oxford University Press.

Engel, P. (1994), *Davidson et la philosophie du langage* , Paris : P.U.F.

Fodor, J. & Lepore, E. (1994), « Is radical interpretation possible? », *Philosophical perspectives*, vol. 8, pp. 101-119.

Føllesdal, D. (1982) «The Status of Rationality Assumptions in Interpretation and in the Explanation of Action », *Dialectica,* 36 (4), pp. 301-316.

Gallese, V. & Goldman, A.I. (1998), « Mirror Neurons and the Simulation Theory of Mind-Reading », *Trends in Cognitive Sciences*, 12(2), pp. 493-501.

Goldman, A.I. (1989), « Interpretation Psychologized », *Mind and Language*, 4(3), pp. 161-185.

Goldman, A.I. (1992), « In Defense of the Simulation Theory », *Mind and Language*, 7(1-2), pp. 104-119.

Goldman, A.I. (2006), *Simulating Minds*, Oxford: Oxford UP.

Gopnik, A. & Wellman, H.M., (1992), « Why the Child's Theory of Mind Really Is a Theory », *Mind and Language*, 7(1-2), pp. 145-171.

Gordon, Robert M. (1986), « Folk psychology as simulation », *Mind and Language*, 1, pp. 158-171.

Gordon, R.M. (1992), « The Simulation Theory: Objections and Misconceptions », *Mind and Language*, 7(1-2), pp. 11-34.

Grandy, R. (1973) « Reference, Meaning and Belief », *The Journal of Philosophy*, 70(14), pp. 439-52.

Harnay, P-V. (2008), *La décision de l'expérimentation à l'interprétation : l'apport de Donald Davidson*, Thèse de doctorat, Université Paris I Panthéon Sorbonne

Harris, P.L. (1992), « From Simulation to Folk Psychology: The Case for Development », *Mind and Language*, 7(1-2), pp. 120-144.

Heal, J. (1996), « Simulation and Cognitive Penetrability », *Mind and Language*, 11(1), pp. 44-67.

Heal, J. (1998), « Co-Cognition and Off-Line Simulation: Two Ways of Understanding the Simulation Approach », *Mind and Language*, 13(4), pp. 477-98.

Heal, J. (2000), « Other Minds, Rationality and Analogy », *Proceedings of the Aristotelian Society*, suppl.vol., 74, pp. 1-19

Hempel, C. G. (1942), «The Function of General Laws in History», *The Journal of Philosophy*, 39, pp. 35-48.

Jacob, P. & Jeannerod, M. (2005) « The motor theory of social cognition: a critique », *Trends in Cognitive Sciences*, 9 (1), pp. 21-25.

Jacob, P. (2008), « What do Mirror Neurons Contribute to Human Social Cognition ? » *Mind and Language*, 23 (2), pp. 190—223.

Laugier, S. (1992) *L'anthropologie logique de Quine*, Paris: Vrin.

Lepore, E. & Ludwig, K. (2005), *Donald Davidson : Meaning, Truth, Language And Reality*, Oxford : Clarendon Press.

Lewis, D.K. (1972), « Psychophysical and Theoretical Identifications » *Australasian Journal of Philosophy*, 50, pp 249-58.

Lewis, D.K. (1974), « Radical Interpretation », *Synthese*, pp 331-44.

Loewenstein, G. & Adler, D. (1995), « A Bias in the Prediction of Tastes », *Economic Journal*, 105, pp. 929-937.

Ludwig, K. (2004) « Rationality, Language and the Principle of Charity » *in* Mele,A. & Rawling, P. (eds) *The Oxford Handbook of Rationality*, Oxford: Oxford UP.

Quine, W.V.O., (1960) *Word and Object*, MIT Press, Cambridge, Mass., tr. fr. J. Dopp et P.Gochet, *Le mot et la chose*, Flammarion 1977.

Quine, W.V.O. (1970) *Philosophy of Logic*, Harvard University Press, tr. fr. J. Largeault, *Le mot et la logique*, Aubier, rééd. 2008.

Rawlings, P. (2003), « Radical Interpretation », *in* Ludwig, K. (ed), *Donald Davidson*, Cambridge : Cambridge UP.

Rizzolatti, G., Gallese, V., Fadiga, L. & Fogassi, L. « Action recognition in the premotor cortex » (1996), *Brain*, 119, pp. 593-609.

Savage, L. (1954), *The Foundations of Statistics*, 2nde ed. 1972, New York : Dover

Stein, E. (1996) *Without Good Reason*, Oxford : Clarendon Press.

Stich, S. (1985) Could man be an irrational animal?, *Synthese*, 64, pp. 115-135.

Stich, S. & Nichols, S. (1992), « Folk Psychology : Simulation or Tacit Theory ? », *Mind and Language*, 7(1), pp. 35-71

Stich, S. & Nichols, S. (2003) « Folk Psychology », *in* Stich, S. & Warfield, T.A. (eds), *The Blackwell Guide to Philosophy of Mind*, London : Blackwell, pp. 235-55.

Stone, T. & Davies, M. (1996),« The Mental Simulation Debate : A Progress Report », *in* Carruthers, P. & Smith, P.K. (eds), *Theories of Theories of Mind*, Cambridge: Cambridge University Press.

Stueber, K. (2006) *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*, Cambridge, Mass: MIT Press.

Thaler, R. (1980). « Toward a positive theory of consumer choice ». *Journal of Economic Behavior and Organization*, 1, pp. 39-60.

Tversky, A. & Kahneman, D. (1983), « Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment », *Psychological Review*, 90(4), pp. 293-315

Van Boven, L., & Loewenstein, G. (2003), « Projection of transient drive states », *Personality and Social Psychology Bulletin*, 29, pp. 1159-1168.

Van Boven, L., Loewenstein, G., & Dunning, D. (2005), « The illusion of courage in social predictions: Underestimating the impact of fear of embarrassment on other people ». *Organizational Behavior and Human Decision Processes*, 96, pp. 130-141.

Wilson, N. L. (1959) « Substances without substrata », *Review of Metaphysics*,12, pp. 521-539.

Wimmer, H. & Perner, J. (1983) «Beliefs about beliefs: representations and constraining function of wrong beliefs in young children's understanding of deception », *Cognition*, 13, pp. 103-128.

Von Wright, G.H. (1971) *Explanation and Understanding*, New York : Cornell University.