La philosophie de l'esprit

[pour le *Précis de philosophie analytique*]

(Jérôme Dokic)

I. Introduction

La philosophie de l'esprit est une réflexion sur la possibilité d'une ontologie des phénomènes mentaux. Elle tente en effet de répondre à deux questions fondamentales : *Les phénomènes mentaux existent-ils? S'ils existent, de quelle nature sont-ils?*

La philosophie de l'esprit se distingue à la fois de la psychologie cognitive et de la philosophie de la psychologie (McGinn, 1996). La psychologie cognitive est une étude empirique de phénomènes mentaux intéressant des organismes réels. La philosophie de la psychologie est une étude des méthodes, concepts et résultats de la psychologie empirique. La philosophie de l'esprit vise à rendre compte de l'essence des phénomènes mentaux, non pas directement sur des bases empiriques, mais par le biais d'une analyse des *concepts mentaux*. Ce qu'il faut entendre par « analyse » dépend du cadre philosophique adopté. Beaucoup ont renoncé à la notion étroite d'analyse réductive *a priori*, en termes de conditions nécessaires et suffisantes. Certains philosophes reconnaissent la légitimité d'analyses circulaires et néanmoins informatives (des analyses qui réutilisent le concept mental à définir dans le *definiens*), ou d'analyses qui reposent essentiellement sur des faits observés et des théories empiriques.

L'analyse des concepts mentaux soulève une difficulté fondamentale, sur laquelle s'attarde tout particulièrement la philosophie de l'esprit. Les concepts mentaux ont des conditions d'application à première vue hétérogènes. D'un côté, nous nous les attribuons « de l'intérieur », sur la base de la connaissance immédiate que nous semblons avoir de nos propres états ou événements mentaux. De l'autre côté, nous attribuons des concepts mentaux à autrui, sur la base de l'observation de son comportement « extérieur » (linguistique ou non). La difficulté majeure consiste à expliquer comment les *mêmes* concepts mentaux, et donc les *mêmes* phénomènes réels, peuvent être visés au travers de ces deux perspectives, en première et en troisième personne, apparemment incommensurables.

On a souvent été tenté d'accorder la priorité ontologique à l'une des deux perspectives au détriment de l'autre. Le *cartésianisme* est la conception (plus ou moins fidèle à la véritable philosophie de Descartes) selon laquelle la perspective en première personne est prioritaire, l'autre perspective ne permettant qu'un accès indirect aux phénomènes mentaux. Selon le *béhaviorisme*, au contraire, la perspective en troisième personne est fondamentale et la connaissance intime que nous croyons avoir de nos propres états mentaux est illusoire. La philosophie de l'esprit récente se caractérise, de manière générale, par la recherche d'une troisième voie entre le cartésianisme et le béhaviorisme.

II. Le mental et le physique

Le problème métaphysique des rapports entre le corps et l'esprit est une illustration de la même difficulté. L'esprit se manifeste normalement par des activités corporelles, et cependant l'introspection et la réflexion peuvent nous persuader que l'esprit en est essentiellement indépendant. Nous avons une « image manifeste » des phénomènes mentaux (selon laquelle ils se présentent en première personne comme aucun autre phénomène) que nous devons réconcilier avec notre « image scientifique » du monde, telle que la définit les sciences actuelles (en particulier la physique). Le problème n'engage pas seulement la philosophie de l'esprit mais la philosophie des sciences toute entière, dominée après le positivisme logique par le réalisme scientifique. En fait, de nombreux aspects de la discussion actuelle des rapports entre le corps et l'esprit concernent de manière plus générale la question du statut des sciences spéciales (psychologie, biologie, neurosciences) par rapport à la physique théorique.

Le dualisme des substances

Selon le *dualisme des substances*, aucune substance ou entité ne peut avoir simultanément les deux jeux de propriétés suivants : d'un côté la subjectivité, la conscience (notamment la conscience de soi), la rationalité et la liberté, et de l'autre côté les qualités premières objectives qui caractérisent les corps physiques. Le corps est une chose, mais l'esprit en est une autre, immatérielle et inétendue. Le reproche le plus couramment adressé à cette position cartésienne est que l'interaction manifeste entre l'esprit et le corps est laissée

dans l'ombre, ou qu'elle apparaît comme inintelligible si on la conçoit sur le modèle d'un échange causal entre deux substances indépendantes. La philosophie de l'esprit récente est globalement *anti-cartésienne* en ce sens. Elle s'attache à définir une conception du mental compatible avec un *monisme de la substance*, selon lequel une même chose (par exemple, une personne) peut avoir les deux jeux de propriétés cités (Strawson 1959). Toutefois, aucun consensus ne s'est établi sur la forme spécifique que doit prendre un tel monisme.

Le naturalisme physicaliste

Selon le *physicalisme* – une forme importante de monisme – tout ce qui existe est de nature physique. La science physique jouit d'un statut fondamental par rapport aux autres sciences naturelles (y compris la psychologie empirique) : tout ce qui est empiriquement accessible dépend ontologiquement d'entités et de propriétés physiques. Le physicalisme est souvent assorti de la thèse selon laquelle toute explication de phénomènes réels est une explication *causale* entièrement spécifiable en termes physiques. Le physicalisme est une forme de *naturalisme*. Le naturalisme est la conception selon laquelle tous les traits du monde peuvent être entièrement décrits, en principe, par les sciences naturelles. (Bien que les naturalistes soient souvent des physicalistes, un naturalisme non-physicaliste est possible ; cf. section VI.)

Les physicalistes fournissent rarement un critère rigoureux de ce qui compte comme étant de nature physique. Du point de vue de la philosophie de l'esprit, le problème central est ailleurs : il concerne la relation entre les phénomènes mentaux et les phénomènes du genre de ceux décrits par la physique contemporaine. Plus spécifiquement, il s'agit de se demander quelle forme de dépendance existe entre ces deux types de phénomènes, et si les phénomènes mentaux peuvent être entièrement assimilés ou réduits à des phénomènes physiques, en particulier cérébraux. Les physicalistes admettent couramment un principe de *complétude physique* (que certains tirent de la pratique des physiciens eux-mêmes) : tout effet physique a une cause physique complète. Le monde physique est causalement clos. Le fait que les phénomènes mentaux ont (apparemment) des effets physiques a-t-il pour conséquence qu'ils sont eux-mêmes physiques ?

Il existe deux formes principales de physicalisme. La première est la *théorie de l'identité* : chaque phénomène mental est identique à quelque phénomène physique. La deuxième forme de physicalisme est la *théorie de la survenance* : les phénomènes mentaux

dépendent unilatéralement des phénomènes physiques, sans leur être à proprement parler identiques. (Sur le reste de cette section, cf. Engel 1992 : ch. 1, Pacherie 1993, Jacob 1997.)

Les théories de l'identité

La théorie de l'identité se présente en fait sous plusieurs versions. Selon la *théorie de l'identité-type*, particulièrement radicale, chaque *type* de phénomène mental est identique à un *type* d'état cérébral, ou plus généralement physique (Lewis 1966). De même que la chaleur est en réalité un certain type d'agitation moléculaire et l'eau un liquide composé surtout de molécules H_2O, un type spécifique de douleur pourra être identifié (*a posteriori* mais nécessairement) à un type déterminé de stimulation neurale (tel qu'il pourrait être décrit par les neurosciences).

La théorie de l'identité-type suscite l'objection intuitive selon laquelle un même phénomène mental (par exemple, un type de douleur) peut être réalisé de différentes manières physiologiques d'un cerveau humain à l'autre, et en principe chez d'autres espèces intelligentes (par exemple dont le tissu organique serait à base de silicone). C'est l'objection de la *réalisabilité multiple*, ou du « *chauvinisme* » (Putnam 1990, Fodor 1968).

Selon une version plus modeste de la théorie de l'identité, chaque *occurrence* d'événement mental est identique à une *occurrence* d'événement physique (Davidson 1984). La théorie de l'identité-type implique cette *théorie de l'identité-occurrence* (si les types psychologiques coïncident avec des types physiques, leurs occurrences coïncideront aussi), mais la réciproque n'est pas vraie : toutes les occurrences d'un même phénomène mental ne relèvent pas nécessairement du même type physique. La théorie de l'identité-occurrence reste essentiellement lacunaire en l'absence d'une réponse détaillée à la question de savoir en vertu de quoi différentes occurrences physiques réalisent le même phénomène mental. Elle remplace le dualisme des substances par un *dualisme des propriétés*, physiques et psychologiques. Elle est relativement neutre en ce qui concerne les rapports entre ces deux propriétés, à tel point qu'elle est compatible en principe avec la conception anti-réaliste du mental selon laquelle il n'y a pas d'essences ou de « genres naturels » mentaux.

Le fonctionnalisme

Certains auteurs ont affirmé, parfois en s'inspirant des remarques de Wittgenstein (1953) sur la signification comme usage dans un jeu de langage, que les concepts mentaux pouvaient être définis en termes causaux, c'est-à-dire par les causes et les effets typiques des phénomènes mentaux qu'ils décrivent (Lewis 1966). La douleur, par exemple, est une expérience typiquement causée par des stimulis nocifs associés à un dommage corporel, et elle cause typiquement un jugement du type « J'ai mal » et des comportements caractéristiques comme retirer sa main du feu, gémir, se soigner. Certains philosophes (cf. Braddon-Mitchell et Jackson 1966) considèrent que les définitions causales des concepts mentaux peuvent être tirées *a priori* de la *psychologie populaire*, c'est-à-dire l'ensemble des principes du sens commun que nous mobilisons pour donner un sens au comportement d'autrui. De telles définitions impliquent une forme de *holisme* relatif à l'attribution des états mentaux : l'attribution d'un état mental à un organisme suppose l'attribution au moins implicite de beaucoup d'autres états mentaux causalement et fonctionnellement liés au premier.

On appelle *fonctionnalisme* la thèse selon laquelle les phénomènes mentaux sont *constitués* par leur position relationnelle dans un réseau de relations causales entre des entrées sensorielles, d'autres états mentaux et des sorties motrices. Le fonctionnalisme est souvent associé à une théorie de l'identité-occurrence. Chaque occurrence d'état mental est identique à une occurrence d'état physique (cérébral), mais un *type* d'état mental est identifié à une propriété d'ordre supérieur : le *rôle causal* (ou *fonctionnel*) que jouent les occurrences de cet état dans l'économie mentale du sujet. Un rôle causal déterminé peut être réalisé dans des cas particuliers par des occurrences qui relèvent de types physiques divers ; l'objection du chauvinisme est ainsi levée. Le fonctionnalisme est même compatible avec le dualisme des substances, à condition que l'on puisse donner un sens à l'idée d'un état de la *res cogitans* qui joue un rôle causal déterminé. Aux yeux du physicaliste, la force du fonctionnalisme réside précisément dans le fait que les rôles fonctionnels n'ont pas besoin d'être remplis par les états d'une substance immatérielle ; ils *peuvent* être occupés par des états physiques. Le fonctionnalisme physicaliste identifie les types mentaux non pas directement à des types physiques de premier ordre, mais à des types fonctionnels, intermédiaires entre le niveau intentionnel proprement dit (auquel se place la psychologie populaire) et celui de la réalisation physique des occurrences mentales.

Une version assez différente du fonctionnalisme est le *fonctionnalisme homonculariste* de Dennett (1983). Selon cette conception, les capacités apparemment miraculeuses d'un système intentionnel intelligent peuvent être expliquées par une hiérarchie de fonctions qui pourraient être remplies par des « homoncules » de plus en plus stupides, et en définitive par de simples mécanismes.

La théorie de la survenance

Le concept de *survenance* [*supervenience*] vient de la philosophie morale (Moore 1903) et de la philosophie des sciences du début du siècle (notamment l'émergentisme ; cf. Kim 1994). Il a été récemment utilisé dans l'espoir qu'il permettait une formulation plus rigoureuse des rapports entre le physique et le mental (Davidson 1984). La survenance est une relation de dépendance ontologique asymétrique (contrairement à l'identité, qui est une relation symétrique). Un phénomème mental survient sur un ensemble de phénomènes physiques lorsque toute différence mentale implique nécessairement quelque différence sur le plan physique, même si la réciproque ne vaut pas (le mental ne peut pas varier indépendamment du physique). La survenance est *locale* lorsque les phénomènes physiques sous-jacents concernent la personne ou son cerveau ; elle est *globale* lorsqu'ils coïncident avec le monde physique dans son ensemble.

Pour le naturalisme physicaliste, la thèse de la survenance ne se résume pas au fait que deux organismes ou deux mondes indiscernables au point de vue physique le sont également au point de vue psychologique ; elle doit accréditer l'idée selon laquelle les faits physiques *fixent* ou *constituent* les faits psychologiques. Le physique a une priorité ontologique sur le mental. On peut avoir l'intuition que les propriétés esthétiques d'une œuvre d'art sont fixées par les propriétés physiques de l'œuvre et de son contexte. Le physicaliste qui invoque la thèse de la survenance cherche à exploiter une intuition analogue concernant les rapports entre le physique et le mental ; une fois l'ensemble des faits physiques (locaux ou globaux) spécifiés, tous les faits mentaux sont automatiquement fixés.

En tant que telle, la thèse de la survenance ne fournit aucune *explication* du rapport entre le physique et le mental. Comme la théorie de l'identité-occurrence, elle implique un dualisme des propriétés (mentales et physiques), mais le rapport spécifique entre ces propriétés doit encore être précisé. Ce n'est pas un rapport causal car les faits mentaux et physiques ne sont pas suffisamment distincts les uns des autres. De plus, la thèse de la

survenance est compatible avec la théorie de l'identité-occurrence, mais elle ne l'exige pas. Elle établit une relation de dépendance entre les propriétés mentales et les propriétés physiques, mais elle n'implique pas que ces propriétés s'attachent aux mêmes occurrences.

Le naturalisme physicaliste est une position *réductionniste* qui vise à assimiler les phénomènes mentaux à des phénomènes physiques. Comme on vient de le voir, la version la plus radicale de cette position est la théorie de l'identité-type, selon laquelle les types psychologiques et physiques sont coextensifs. Le réductionniste qui rejette cette version par trop chauvine se voit contraint de relâcher le lien entre les propriétés psychologiques et les propriétés physiques. Il troque ainsi le chauvinisme contre une certaine autonomie du mental par rapport au physique. Cette autonomie ne doit pas être trop grande, car elle pourrait rendre le mental *épiphénoménal*, étant donné la complétude revendiquée de la physique. Le réductionniste sophistiqué doit donc trouver un équilibre délicat entre le réductionnisme classique (la théorie de l'identité-type) et le sevrage complet du mental par rapport au physique, qui ne répond plus aux intuitions physicalistes.

Pour les détracteurs du physicalisme, un tel équilibre est impossible à atteindre : trop de différences opposent le mental et le physique. Au moins deux dimensions apparemment constitutives du mental semblent résister à la réduction physicaliste : la *conscience* et l'*intentionnalité*. La conscience est liée à l'effet subjectif que cela fait d'être dans un état mental (Nagel 1979). L'intentionnalité est la propriété qu'ont les états mentaux de présenter ou de représenter des objets et des états de choses – de les viser ou d'être dirigés vers eux. C'est la propriété par laquelle les états mentaux engagent des *contenus mentaux* (Brentano 1924-1928, Searle 1983).

III. La conscience

La conscience et l'intentionnalité

Une question centrale de la philosophie de l'esprit concerne les relations entre la *conscience* et l'*intentionnalité*. Certains états mentaux (comme celui dans lequel nous nous trouvons lorsque nous avons mal) semblent être essentiellement conscients. Quel sens y a-t-il à parler d'une douleur qui n'est pas effectivement ressentie, même si elle ne reçoit pas forcément pas toute notre attention ? D'autres états mentaux ne sont pas conscients, comme

certaines croyances non explicitement considérées par leur sujet. Peut-on considérer ces croyances comme des dispositions à entrer dans des états conscients, par exemple des capacités à former des jugements ?

La philosophie de l'esprit actuelle est partagée entre deux attitudes, qui correspondent à des stratégies méthodologiques divergentes. Certains philosophes défendent la thèse selon laquelle l'intentionnalité est un phénomène plus fondamental que la conscience, de sorte qu'une théorie de l'intentionnalité a une priorité logique sur une théorie de la conscience (Dennett 1991, Dretske 1995). Cette thèse est souvent associée à une approche « ascendante » des problèmes de la philosophie de l'esprit, qui préconise la considération d'organismes primitifs intuitivement dépourvus de conscience mais éventuellement capables de représentations rudimentaires (Proust 1997, Jacob 1997). D'autres prétendent au contraire que la notion de conscience est première, et qu'elle doit être invoquée pour rendre compte de l'intentionnalité (Searle 1983, 1992). Ils privilégient typiquement une approche directe des phénomènes cognitifs conscients de l'être humain, sur laquelle ils fondent une théorie de l'intentionnalité. Les capacités de représentation apparemment attribuables à des organismes dépourvus de conscience peuvent éventuellement se comprendre par analogie (mais par analogie seulement) à nos états cognitifs conscients.

Le programme intentionnaliste

Certains états mentaux semblent avoir une « qualité phénoménale » qui définit, selon l'expression consacrée, « l'effet que cela fait » d'être dans ces états (Nagel 1979). On parle ainsi des *qualia* associés aux phénomènes mentaux, et immédiatement accessibles à la conscience. Les *qualia* incluent souvent ce que la tradition appelle des *sensations*, et en particulier des sensations corporelles, comme la douleur. L'*intentionnalisme* est la thèse générale selon laquelle les qualités phénoménales peuvent être entièrement définies en termes intentionnels. Les *qualia* ne sont pas des propriétés *de* l'expérience, mais des propriétés (premières ou secondes) présentées *dans* l'expérience. Toute différence phénoménale accessible au sujet percevant est en réalité une différence relative à la scène objective qui lui est présentée.

L'intentionnalisme est plausible lorsqu'il est appliqué à certaines attitudes propositionnelles comme les croyances, auxquelles il n'est pas évident que soient associés des *qualia* caractéristiques. La tentation d'introduire des *qualia* distincts des contenus

intentionnels est plus grande pour les états mentaux non dispositionnels et essentiellement conscients. Considérons l'expérience perceptive. La tradition postule des sensations visuelles, tactiles, etc., qui constituent les « éléments » de l'expérience. Bien que les sensations soient conscientes, elles ne sont pas (normalement) perçues. Elles ne participent donc pas du contenu intentionnel de l'expérience. Du point de vue intentionnaliste, en revanche, la conscience perceptive est vide : les sensations de la tradition sont assimilées à des propriétés présentées *dans* l'expérience.

Les émotions et les sensations corporelles résistent tout particulièrement à la réduction intentionnaliste. De ce point de vue, il faut montrer qu'en dépit des apparences, ces phénomènes ont un contenu ; ils représentent le monde (objectif ou corporel) d'une certaine manière. Par exemple, Dretske (1995) affirme que l'expérience de douleur a un objet intentionnel spécifique : un dommage corporel physique. Les détracteurs du programme intentionnaliste rétorquent que si l'expérience de douleur a un objet apparent, celui-ci ne peut pas exister sans être effectivement ressenti, ce qui n'est pas le cas de la blessure physique qui résiste à l'anesthésie. C'est donc un « objet » spécial, dont le statut ontologique est très différent de celui du membre meurtri. (Mais l'intentionnalisme n'est pas forcément physicaliste.)

Selon une version moins radicale de l'intentionnalisme, les *qualia* peuvent être définis par des contenus intentionnels *et* par des propriétés fonctionnelles. La douleur est ce qu'elle est, au point de vue phénoménal, non seulement parce qu'elle représente un certain type d'état de choses (un dommage corporel), mais aussi parce qu'elle joue un rôle fonctionnel caractéristique dans l'économie mentale du sujet (Lycan 1996).

Les qualia *sont-ils naturels ?*

Les *qualia* posent le *problème du fossé explicatif* : il semble que même une description adéquate et complète du fonctionnement du cerveau soit incapable d'expliquer pourquoi tel état mental (par exemple, la perception visuelle d'un ciel bleu) est associé à tel « effet » qualitatif plutôt qu'à tel autre. Il est apparemment impossible de comprendre comment les *qualia* peuvent émerger d'un système purement physique. Un certain nombre d'expériences de pensée ont été proposées pour illustrer ce problème. (Sur ces expériences, cf. Block 1994 et Clémentz 1997, et les autres références données dans ces essais.)

Considérons l'*argument de la connaissance* de Jackson (1990). Marie est confinée depuis sa naissance dans une pièce en noir et blanc, sans jamais entrer en contact visuel avec un objet coloré. Par hypothèse, Marie possède une connaissance complète des mécanismes physiques et neurophysiologiques qui sous-tendent la perception chromatique. Comment décrire sa situation cognitive lorsqu'elle quitte la pièce pour la première fois et se trouve devant un ballon rouge ? Selon l'argument de Jackson, elle prend connaissance de faits inédits, irréductibles à des faits physiques, en particulier l'effet que cela fait de voir un ballon rouge. Une autre réponse consiste à dire que Marie apprend à reconnaître et à imaginer des couleurs, mais que cette connaissance est de l'ordre d'un savoir-faire et non d'un savoir purement théorique. On peut aussi affirmer que Marie apprend de nouveaux concepts (des concepts « récognitionnels »), mais que ceux-ci désignent des propriétés auxquelles Marie avait déjà accès au travers des théories qu'elle avait maîtrisées dans son environnement en noir et blanc (cf. Lycan 1990 : VII).

Les défenseurs des *qualia* ont fait valoir, plus spécifiquement, leur irréductibilité aux propriétés fonctionnelles d'un système cognitif. Selon l'*argument des* qualia *absents*, on peut concevoir un système cognitif fonctionnellement équivalent à notre esprit, c'est-à-dire un organisme (naturel ou artificiel) qui réagit comme nous dans n'importe quelle situation réelle ou contrefactuelle, mais entièrement ou partiellement dépourvu des *qualia* qui caractérisent notre expérience. Certains organismes pourraient être des « zombies » en ce sens, mais nous ne pourrions pas le savoir en observant seulement leur comportement linguistique et non-linguistique (dans toutes les situations possibles) (Kirk 1994). Le problème que soulève cet argument est qu'il est difficile de dissocier la possibilité des zombies du corollaire intuitivement dérangeant selon lequel *nous* sommes des zombies.

L'argument des *qualia* absents peut conduire à une objection plus générale contre le fonctionnalisme. Le fonctionnalisme échappe au chauvinisme au prix d'un libéralisme outrancier. Le réseau fonctionnel auquel on identifie un ensemble d'états mentaux est caractérisé de manière tellement abstraite qu'il pourrait être réalisé par des systèmes physiques intuitivement dépourvus de propriétés mentales (et pas seulement de *qualia*), comme les événements microscopiques d'un étang ou l'évolution d'une nation comme la Chine dont les habitants pourraient imiter par hasard le comportement fonctionnel de groupes de neurones.

D'autres arguments font intervenir des expériences de pensée moins radicales. C'est le cas de l'*argument du spectre inversé*, qui a des origines classiques (par exemple chez Locke). Supposons que le *quale* associé au mot « rouge » dans mon esprit soit associé au mot « bleu »

dans le vôtre, et vice-versa. Cette différence peut ne jamais se révéler dans notre comportement (linguistique ou non). Par exemple, nous donnons tous deux notre assentiment à la phrase française « Le sang et le drapeau chinois sont rouges ». On peut objecter à cet argument que l'espace logique des qualités comporte un certain nombre d'asymétries qui rendent détectable en principe une inversion qualitative (Harrison 1973), mais il n'est pas évident que ce type d'asymétrie soit un trait nécessaire de tout espace qualitatif. Or l'argument ne requiert que la *possibilité* de telles inversions. D'un autre côté, si cette possibilité ne concerne pas notre expérience réelle, nos intuitions la concernant seront moins assurées.

Les arguments qui viennent d'être évoqués peuvent encourager une division entre le contenu intentionnel d'un état mental et son contenu qualitatif (les *qualia*). Ils pourraient aussi légitimer une distinction entre deux formes de conscience : une *conscience d'accès*, explicable en termes fonctionnels, et une *conscience phénoménale*, qui échappe à l'emprise fonctionnaliste. Cette distinction (proposée par Ned Block) peut être acceptée par un physicaliste, par exemple si la conscience phénoménale est expliquée en termes de propriétés physiques non-fonctionnelles des états cérébraux.

Dennett (1991) a présenté un argument célèbre contre l'existence des *qualia*. Il est prématuré de décrire les cas d'inversions spectrales comme impliquant un changement dans la corrélation entre les rôles fonctionnels de nos états mentaux et des *qualia*. Selon lui, en effet, les *faits* justifient aussi bien l'hypothèse selon laquelle notre expérience qualitative a changé que l'hypothèse rivale selon laquelle ce sont nos réponses qui se sont modifiées. Il n'y aurait donc pas de faits déterminés concernant l'identité des *qualia* – pas d'identité, donc pas d'entité. Pour Dennett, la thèse selon laquelle il existe des *qualia* ineffables, intrinsèques, immédiatement accessibles à la conscience et éventuellement privés n'est qu'un avatar du modèle cartésien du mental. Il rejette en particulier la ligne de partage implicite dans ce modèle entre les processus préconscients et les états mentaux conscients. Il n'y a aucun « moment » temporel ou logique où les informations recueillies par le cerveau sont rassemblées et unifiées pour être présentées à la conscience.

La conscience et la réflexion

Selon une conception de la conscience actuellement très débattue, un état mental est conscient lorsqu'il fait lui-même l'objet d'une représentation d'ordre supérieur,

immédiatement et non-inférentiellement fondée sur lui. Cette méta-représentation peut être considérée comme un jugement ou comme une expérience (le sens interne étant alors conçu sur le modèle de l'observation externe). Elle peut être occurrente ou seulement dispositionnelle, consciente ou non (Rosenthal 1993). Dans tous les cas, toute conscience implique une forme de réflexion, puisque *qua* consciente elle est un objet intentionnel pour le sujet. Cette conception de la conscience peut être combinée au fonctionnalisme. Les *qualia* d'une expérience de premier ordre (par exemple, la perception d'une chaise) consistent dans la prise de conscience, sur un plan réflexif, des propriétés fonctionnelles de cette expérience (Lycan 1996).

On peut objecter que la conception métareprésentationnelle de la conscience ne rend pas justice à notre expérience ordinaire. Lorsque je vois un zèbre, je ne suis pas normalement conscient de tous les aspects de mon expérience, comme le nombre de rayures présentées dans mon champ visuel. Suivant Dretske (1993), certaines rayures sont telles que j'en ai conscience sans avoir la moindre propension à me représenter le fait que j'en suis conscient. Par ailleurs, il faut tenir compte de la possibilité que la métareprésentation *change* les propriétés qualitatives de l'état mental représenté. Si la nature phénoménale d'un état mental est considérée comme étant nécessairement consciente, et si elle est manifestement modifiée par l'introspection et le jugement réflexif, elle ne dépend pas toujours d'une représentation d'ordre supérieur.

Des travaux récents (cf. Campbell, 1994) tentent d'expliciter, en collaboration avec la psychologie cognitive, les relations entre deux formes de conscience : la conscience immergée, pratique et égocentrique d'un côté et la conscience réflexive, détachée et allocentrique de l'autre. Deux questions centrales se posent. Premièrement, la conscience immergée implique-t-elle vraiment une capacité réflexive, comme le préconise la conception métareprésentationnelle ? Ne peut-on pas considérer par exemple que la capacité complexe de navigation spatiale de certains animaux non-linguistiques témoigne de la maîtrise pratique d'une conception rudimentaire de l'espace ? Les organismes incapables de réflexion sont-ils pour autant entièrement dépourvus de conscience ? Deuxièmement, peut-on concevoir la conscience réflexive comme une forme élaborée de conscience immergée ? La théorie est-elle entièrement issue de la pratique et de l'expérience, comme le pensent les pragmatistes et les empiristes, ou y a-t-il un saut qualitatif entre une conception immergée et une conception détachée du monde ? Ces questions concernent les conditions du *décentrement cognitif* (spatial, temporel ou social) et les capacités cognitives qui sous-tendent celui-ci (comme le

raisonnement spatial, le souvenir autobiographique ou la « théorie de l'esprit » des psychologues du développement).

IV. L'intentionnalité

Le problème de l'intentionnalité consiste à expliquer comme un état mental attribuable à un individu peut viser un objet ou un état de choses en dehors de cet individu. On peut distinguer deux aspects de ce problème. Premièrement, l'intentionnalité engage des relations « verticales » entre le sujet et les objets intentionnels de ses états mentaux. Deuxièmement, elle fait également intervenir des relations « horizontales » entre des états mentaux et entre ceux-ci et le comportement. Comme nous le verrons, la relation entre ces deux aspects est controversée.

Internalisme ou externalisme ?

Dans quelle mesure l'esprit en général, et les phénomènes mentaux en particulier, doivent-ils leur existence et leur nature à celles d'états de choses qui existent en dehors de l'esprit ? Peut-on faire une « géographie de l'esprit » sans faire un peu de vraie géographie (McGinn 1996) ? Autrement dit, peut-on décrire les aspects pertinents des phénomènes mentaux sans être amené à former des hypothèses plus ou moins substantielles sur la structure du monde naturel indépendant de l'esprit ?

L'internalisme affirme l'existence d'une ligne de partage franche entre ce qui est mental ou intérieur, et ce qui est non-mental ou extérieur à l'esprit. L'internalisme n'est pas toujours dualiste comme chez Descartes. Dans la conception cartésienne, l'esprit est une substance autonome, ontologiquement indépendante de l'environnement extérieur sur lequel portent ses perceptions et ses pensées. Selon le *monisme internaliste*, l'esprit est de même nature que cet environnement (par exemple physique), mais il lui reste tout à fait extérieur, même si l'extériorité est ici littéralement spatiale (McDowell 1986). Les deux formes d'internalisme – dualiste ou moniste – partagent la conviction selon laquelle les phénomènes mentaux sont essentiellement indépendants des objets et états de choses extérieurs sur lesquels ils portent. Par opposition, l'externalisme est la position selon laquelle l'esprit et les phénomènes mentaux sont constitués, au moins en partie, par l'environnement. De même que

l'internaliste peut être moniste, l'externalisme est compatible avec le dualisme, du moins celui qui concerne les propriétés.

Certains externalistes tentent de montrer, souvent par le biais d'expériences de pensée, que nos concepts doivent au moins en partie leur contenu à la nature de l'environnement. Certains concepts sont « déférentiels » eu égard à la société, à la nature, ou aux deux à la fois. Putnam (1975) imagine une planète en tous points semblable à la Terre (« Terre-Jumelle »), sauf en ce qui concerne le liquide qui ressemble le plus à l'eau terrienne, et qui n'est pas composé de molécules H_2O (sa composition chimique est très différente). Selon Putnam, la pensée d'un terrien relative à l'eau n'est pas exactement la même que la pensée analogue dans l'esprit d'un habitant de Terre-Jumelle, bien qu'elles s'expriment toutes deux au moyen des mêmes phonèmes, par exemple « Voici de l'eau ». Les deux pensées ne portent pas sur la même espèce naturelle ; elles n'ont pas les mêmes conditions de vérité en vertu de leur ancrage différent au contexte.

L'*externalisme social* concerne spécifiquement le rapport entre la pensée et le langage public. Selon cette position, les ressources expressives du sujet ne sont pas indépendantes de celles d'autres sujets parlant la même langue : c'est l'argument de la « division du travail linguistique » offert par Putnam (1975) et modifié par Burge (1979). L'environnement social contribue à la détermination de la nature intrinsèque de la pensée.

L'*externalisme biologique* est la conception selon laquelle l'identité de nos pensées dépend de certaines fonctions biologiques, qui supposent à leur tour une histoire évolutionniste appropriée (cf. les références dans Engel 1992, Pacherie 1993, Proust 1997 et Jacob 1997). De ce point de vue, un organisme spontanément issu d'un marais par une réunion miraculeuse de molécules disparates n'aurait pas de contenus mentaux, puisqu'il n'est pas le produit de l'évolution (Davidson 1987).

Une autre forme d'externalisme concerne les *pensées singulières*, dont l'expression implique la référence à un objet particulier. Par exemple la pensée déictique exprimée dans un contexte particulier par « Ceci est une table » ne peut être formée qu'en présence perceptive de cette table. La pensée est *de re*, et dépend de l'existence réelle de son objet intentionnel (Evans 1982, McDowell 1984, Recanati 1993, Corazza 1995).

On peut distinguer l'*externalisme descriptif*, qui s'intéresse à l'analyse de nos concepts existants, et l'*externalisme transcendental*, qui concerne les conditions générales de possibilité de la pensée et de l'expérience. La théorie de l'interprétation radicale de Davidson (cf. section VI plus bas) est une illustration de l'externalisme transcendental ; elle revendique

l'effacement de la ligne de partage entre les représentations intérieures (nos croyances) et le monde extérieur tel qu'il est en fait (la vérité de nos croyances).

Selon les formes transcendentales d'externalisme, le sujet percevant et pensant n'est pas autonome relativement au monde extérieur ; il doit pouvoir compter, la plupart du temps de manière tacite, sur le fait que son environnement extérieur est structuré de diverses façons, sans quoi certaines pensées lui seraient inaccessibles. Le fait que nous sommes capables de telles pensées atteste indirectement de l'existence d'un monde extérieur structuré, que les structures en question soient des formes de vies, des jeux de langage, des contingences factuelles, des essences naturelles, ou d'autres choses encore. L'externalisme semble donc autoriser la connaissance *a priori* d'un ensemble de faits apparemment contingents (Corazza et Dokic 1996).

D'autres conséquences de l'externalisme doivent être mentionnées, qui concernent respectivement l'*explication psychologique* et l'*autorité de la première personne sur ses propres états mentaux*.

Il semble que le même comportement intentionnel puisse être expliqué à partir de contenus externes différents. Lorsque j'ai soif, je saisis le verre d'eau en face de moi. Lorsque mon sosie sur Terre-Jumelle a soif, il fait *de même* : il saisit le verre en face de lui. Selon l'externalisme, nous n'avons pas les mêmes croyances : je crois que j'ai de l'eau en face de moi, mon sosie croit autre chose (que je ne peux pas exprimer directement dans mon langage). Or la différence de contenu intentionnel entre nos croyances ne semble jouer aucun rôle dans l'explication de notre comportement. Cette situation est problématique, puisque la notion de contenu est introduite, en philosophie de l'esprit comme en psychologie populaire, comme un moyen d'expliquer le comportement intentionnel.

Pour certains philosophes, l'explication psychologique n'a rien à voir avec les conditions de vérité des croyances : ce qui compte, c'est le *contenu étroit*, que mon sosie et moi saisissons également, et non le *contenu large*, qui dépend de la structure invisible de notre environnement. Il faut donc accepter le « solipsisme méthodologique » rejeté par Putnam : au niveau qui concerne l'explication du comportement, les contenus que nous saisissons ne s'alignent pas sur la structure réelle de l'environnement, mais sur l'apparence phénoménale, c'est-à-dire sur le monde tel que nous le voyons « de l'intérieur » (Fodor 1981). Certes, les contenus larges sont rapportés par les comptes rendus d'attitudes propositionnelles, comme « Pierre croit qu'il y a de l'eau devant lui », mais les contenus à l'œuvre dans la rationalisation du comportement sont étroits : ils se calquent sur la perspective propre de l'agent.

Un problème potentiel concerne la nature du contenu étroit. Il est par définition *ineffable*, puisque toute tentative de le spécifier dans un langage public réintroduit le contenu large et donc les éléments externalistes dont on tente de se débarrasser. On ne peut pas exprimer une apparence phénoménale en termes neutres, indépendamment du contexte particulier dans lequel on se trouve.

Le philosophe qui rejette la distinction entre contenu étroit et contenu large peut « externaliser » le comportement lui-même, et donc l'explication psychologique correspondante. Même si mon sosie et moi sommes dans les mêmes états cérébraux et fonctionnels, qui expliquent causalement les mêmes mouvements musculaires, notre comportement est différent, car il dépend du contexte particulier dans lequel nous évoluons. Alternativement, on peut être amené à modifier le modèle causal de l'explication souvent présupposé par l'internalisme. Le rôle du contenu des croyances et des désirs dans l'explication du comportement ne se résume pas à celui d'un « pouvoir causal » (cf. plus bas, « La causalité mentale »).

L'autre difficulté majeure pour l'externalisme concerne la connaissance intime que nous semblons avoir de nos propres états mentaux. Beaucoup de philosophes par ailleurs anti-cartésiens retiennent la thèse de Descartes selon laquelle le sujet est une autorité sur l'identité des contenus mentaux qu'il saisit. Cette thèse est-elle compatible avec l'externalisme ? Aucune analyse chimique du liquide que j'ai devant moi n'est nécessaire pour que je sache que c'est de l'eau. J'ai un accès direct, par la réflexion ou l'introspection, au contenu mental de mes croyances et de mes désirs (du moins ceux qui sont conscients). Or je serais incapable de distinguer, uniquement par introspection, mes contenus de croyance de ceux de mon sosie sur Terre-Jumelle (si j'étais capable de les saisir). On peut rétorquer que l'autorité du sujet sur lui-même ne s'applique pas aux contenus mentaux, mais seulement aux sensations (comme la douleur). Une autre réponse est que l'autorité du sujet s'étend aux contenus mentaux, mais que cela n'implique pas l'existence d'un contenu étroit, insensible à la structure propre de l'environnement. Simplement, la réflexion reprend les contenus réfléchis. Je crois qu'il y a là de l'eau, et je me rends compte que je crois qu'il y a là de l'eau ; la structure de la réflexion m'assure que le même concept d'eau est engagé dans ma croyance de premier ordre et dans ma prise de conscience réflexive (Davidson 1987, Burge 1988).

La naturalisation des contenus mentaux

Pour le naturalisme physicaliste, l'esprit n'est rien d'autre qu'un système physique complexe. Par conséquent, les propriétés logiques, sémantiques et intentionnelles d'un état mental doivent être entièrement dérivables de propriétés physiques. Cette conviction est à la base d'un programme de recherche visant à *naturaliser l'intentionnalité*, c'est-à-dire à rendre compte des représentations et de leurs propriétés normatives (en particulier, le fait que les représentations sont correctes ou incorrectes) en termes purement physiques (Fodor 1990, Engel 1992 : ch. 5, Pacherie 1993, Dretske 1995, Proust 1997, Jacob 1997).

Le point de départ d'une naturalisation de l'intentionnalité est une analogie entre les états mentaux et les *signes ou indices naturels* qui ont une valeur de représentation en vertu de dépendances causales et plus généralement nomiques. C'est ainsi que la fumée est un signe naturel du feu, de même que les cernes de l'arbre constituent un signe naturel de son âge (Grice 1957). L'indication naturelle suppose l'existence de lois (causales ou non) entre des occurrences ou des états de choses, qui justifient des énoncés contrefactuels (« S'il n'y avait pas de feu, il n'y aurait pas de fumée », « Si l'arbre n'avait pas *n* ans, il n'aurait pas *n* cernes »). L'indication naturelle est objective dans la mesure où elle existe indépendamment d'un interprète. L'analogie entre l'indication naturelle et l'intentionnalité soulève un certain nombre de problèmes dont la résolution constitue l'*agenda* du programme de naturalisation de l'intentionnalité.

Un premier problème concerne la *densité informationnelle* des signes naturels. Le contenu de certains états mentaux comme les croyances est *conceptuel*, alors que celui des signes naturels semble être *non-conceptuel*. Premièrement, une empreinte représente la forme spécifique du pied ou de la chaussure, qu'aucun concept général ne peut rendre aussi précisément (Evans 1982). Deuxièmement, l'information naturelle relative à la forme est enchâssée dans d'autres informations naturelles, par exemple relative à la taille. La seule information véhiculée par la croyance qu'il pleut, en revanche, est relative à la présence de la pluie ; elle n'est pas forcément enchâssée dans d'autres informations, par exemple relative à l'intensité de la pluie (Dretske 1981).

Le *problème de l'opacité* concerne le degré d'intensionnalité élevé des contenus mentaux. Ma représentation de F peut différer de ma représentation de G, même si F et G sont les mêmes propriétés au point de vue nomologique. La même propriété (l'eau) peut être présentée sous deux modes différents (comme « eau » et comme « H_2O ») sans que le sujet s'en rende compte (cf. Frege 1971). Par contre, si un état véhicule une information naturelle

sur la présence d'eau, elle véhicule *eo ipso* une information sur la présence d'un liquide composé (surtout) de molécules H_2O, puisque l'eau est nécessairement un liquide de ce genre. Pour résoudre le problème de l'opacité, le partisan de l'intentionnalité naturalisée peut faire appel à certains éléments de la conception millienne des propositions en philosophie du langage (Jacob 1997). Par exemple, il peut affirmer que les modes de présentations ne sont pas sémantiques, mais syntaxiques : c'est une différence non-sémantique (peut-être compositionnelle) entre les véhicules de l'information qui rend compte de l'opacité des représentations mentales (Fodor 1990). Il reste que cette différence correspond à des potentiels inférentiels distincts, ce qui nous amène au problème du holisme.

Le *problème du holisme* est que les croyances forment nécessairement un système, alors qu'un état informationnel est en principe indépendant d'autres états informationnels. La sémantique informationnelle s'oppose à la *sémantique des rôles fonctionnels*, pour laquelle les relations inférentielles entre les états mentaux participent du contenu de ceux-ci (Loar 1981, Pacherie 1993 : ch. 6). La sémantique des rôles fonctionnels reprend les principes du fonctionnalisme dans le cadre d'une théorie des contenus mentaux. Ceux-ci sont étroits lorsque la définition fonctionnaliste s'arrête aux stimulis et aux comportement « proximaux » ; ils sont larges lorsque des objets et des comportements externes sont pris en compte – on parle alors de fonctionnalisme « à bras long » (Pacherie 1993). Il a été suggéré que la sémantique informationnelle peut emprunter des éléments à la sémantique des rôles conceptuels, notamment pour rendre compte de l'intensionnalité des contenus mentaux (Jacob 1997).

Le *problème de la méreprésentation* (ou de la méprise représentationnelle) est que les signes naturels (par définition) ne mentent pas, alors qu'une croyance peut être fausse. Le « contenu » d'un signe naturel dépend d'une covariation fiable entre la présence du signe et celle de l'état de choses représenté par le signe ; sans l'état de choses représenté, pas de signe naturel. Selon une version du problème de la méreprésentation, le contenu des signes naturels est disjonctif. Supposons que mon expérience visuelle d'un cheval soit considérée comme un signe naturel de la présence d'un cheval (un bon détecteur de cheval). La corrélation entre mon expérience et la présence d'un cheval est imparfaite ; dans certains cas, j'ai la même expérience causée par la présence d'une vache vue de loin ou de nuit. Dans ces cas, mon expérience est intuitivement illusoire, mais pourquoi ne pas dire qu'elle représente correctement la présence d'un cheval *ou* d'une vache (Fodor 1990) ? Un autre problème apparenté est celui de la *distalité*. Si la même expérience est causée par un cheval et par une vache, c'est (en partie) parce que les mêmes simulations rétiniennes sont en jeu. Pourquoi ne

pas dire alors que mon expérience est un signe naturel de ces stimulations plutôt que de la présence d'un objet « distal » (Proust 1997) ?

Plusieurs solutions au problème de la méreprésentation ont été proposées, qui impliquent une distance plus ou moins grande par rapport au noyau dur de la sémantique informationnelle. Une première division est celle entre les théories qui tentent de résoudre le problème en termes purement informationnels et celles qui introduisent la notion supplémentaire de *fonction*. Dans la première catégorie, on tente de montrer que les représentations incorrectes dépendent des représentations correctes, mais non réciproquement. Si ma représentation visuelle d'un cheval est parfois causée par la présence nocturne d'une vache, c'est qu'elle est par ailleurs nomiquement liée à une présence chevaline, mais l'inverse n'est pas vrai. C'est la théorie de la *dépendance nomique asymétrique* défendue par Fodor (1990).

L'autre stratégie consiste à rendre compte de l'erreur en invoquant, outre les relations informationnelles, la *fonction* remplie par le signe naturel. On peut ainsi expliquer la différence entre ce que le signe indique dans un contexte particulier et ce qu'il est *supposé* indiquer de par sa fonction. On s'intéresse alors aux conditions dans lesquelles un signe est *sélectionné* par un système cognitif. Le processus sélectif peut être d'origine ontogénétique ou phylogénétique. Dans le premier cas, le signe naturel est recruté comme indicateur fiable d'un certain état de choses au cours d'un processus d'apprentissage individuel. Dans le deuxième cas, le processus sélectif est la sélection naturelle et la fonction pertinente est biologique. C'est la *sémantique téléologique* ou *téléosémantique*. Les deux cas peuvent être combinés, pour rendre compte de l'intentionnalité respective des croyances et des expériences (Dretske 1995). Une troisième option consiste à rejeter tout élément informationnel dans la définition des contenus mentaux pour ne retenir que la dimension téléologique. C'est ainsi qu'on a affirmé qu'une téléosémantique à base informationnelle n'est pas compatible avec un point de vue fondé sur l'adaptation (Millikan 1984). La téléosémantique pure est la conception selon laquelle le contenu est entièrement fixé par les conditions de réussite de la croyance telles qu'elles sont déterminées par des fonctions biologiques. Cette conception reprend la thèse pragmatiste de Ramsey (1978) selon laquelle les croyances sont des cartes mentales grâce auxquelles nous nous orientons dans le monde, mais il reste à montrer qu'elle est réellement tributaire d'hypothèses téléologiques.

En fait, le recours à la sélection naturelle pour naturaliser l'intentionnalité n'a pas manqué de soulever plusieurs objections. La critique la plus courante consiste à faire valoir que l'attribution d'une fonction à un état mental, organe ou artefact est toujours relative à

l'observateur, et ne correspond à aucune propriété intrinsèque de la situation décrite (Dennett 1983, Searle 1992). Selon une autre critique, la sélection naturelle est aveugle et ne peut rendre compte du contenu *déterminé* de nos états mentaux (Fodor 1990).

Le problème de la causalité mentale

Il semble qu'un élément important de la psychologie populaire concerne le rôle causal de nos états mentaux : j'ai frappé la balle *parce que* je *voulais* la renvoyer à mon adversaire. Non seulement les états mentaux sont des causes, mais leur *contenu* (leurs propriétés sémantiques) doit jouer un rôle substantiel dans les transitions rationnelles entre états mentaux, et dans la production du comportement. Lorsque la voix du chanteur soprano brise le verre en cristal, c'est le son physique, et non la signification des paroles chantées, qui joue le rôle causal pertinent. Le contenu des paroles est *épiphénoménal* dans le processus qui conduit à la brisure du verre. Mais comment montrer que le contenu de nos états mentaux n'est pas *toujours* épiphénoménal en ce sens ? C'est le *problème de la causalité mentale*, particulièrement épineux pour le naturalisme physicaliste.

Pour certains philosophes, la causalité mentale n'est possible que si les propriétés sémantiques des états mentaux sont causalement efficaces dans la production du comportement. Cette conviction soulève deux problèmes distincts : la *menace de la préemption* (ou le *problème de l'exclusion explicative*) et le *problème de l'externalisme* (Engel 1992 : ch. 5, Jacob 1997 : ch. 7). Selon le premier problème, la causalité se joue entièrement au niveau physique sous-jacent ; les propriétés sémantiques survenantes sont inertes. Selon le second problème, la causalité mentale est locale et intrinsèque, alors que le contenu de états mentaux est relationnel et (au moins en partie) extrinsèque à l'individu, ce qui semble également impliquer l'inertie du mental.

Différentes stratégies ont été proposées pour sortir de l'impasse de l'épiphénoménalisme. On peut renoncer au modèle de la causalité efficiente ou « déclenchante » au profit d'autres notions de causalité (par exemple « structurante » chez Dretske 1995). On peut aussi tenter de montrer que les transitions causales rationnelles se font *en vertu* du contenu des états mentaux, même si ceux-ci ne sont pas au sens strict les *termes* de relations causales. La référence au contenu des états mentaux permet de donner une *explication causale* du comportement, même si la causalité proprement dite est en définitive purement locale.

V. La rationalité

Le modèle computationnel de l'esprit

Les états mentaux ont un *rôle sémantique* (ou contenu intentionnel) : ils représentent des objets et états de choses dans le monde. Mais ils ont aussi un *rôle inférentiel* : ils entrent en relation rationnelle avec d'autres états mentaux, et avec l'action. Si l'on s'accorde en général sur l'idée qu'il doit y avoir une certaine harmonie entre les deux rôles, la question de la priorité d'un rôle sur l'autre est controversée. Pour la sémantique des rôles fonctionnels, c'est le rôle inférentiel qui détermine le rôle sémantique (ou du moins le contenu étroit) des états mentaux. Les sémantiques informationnelle et téléologiques accordent typiquement la priorité au rôle sémantique, d'où découle éventuellement le rôle inférentiel.

Une autre question se pose, qui concerne la relation entre le rôle inférentiel des états mentaux et le rôle causal qu'ils semblent également jouer. Les états mentaux (doués de contenu) sont unis par des relations rationnelles, sur lesquelles repose le raisonnement théorique et pratique. Ces relations rationnelles ont une dimension *normative*. Par exemple, lorsqu'une croyance implique logiquement une autre, le sujet qui a la première croyance *devrait* avoir la seconde (ou il serait *approprié* qu'il l'ait), même si cela ne se vérifie pas toujours sur le plan empirique. D'où vient alors la portée normative des relations rationnelles ?

Selon une réponse à cette question, le rôle inférentiel des états mentaux est au fond de nature causale et l'explication rationnelle est une espèce d'explication causale. Selon le *modèle computationnel de l'esprit,* les relations rationnelles entre les états mentaux sont réductibles à des relations causales qui dépendent de la *forme* syntaxique ou quasi-syntaxique des états mentaux. Les états mentaux, ou du moins les attitudes propositionnelles, reposent sur un *langage de la pensée* ou « Mentalais » (Fodor 1975). L'hypothèse du langage de la pensée (qui, pour Fodor, est une de nature empirique) est invoquée pour rendre compte de la productivité et de la systématicité de la pensée, notamment la capacité de former un nombre indéfini de pensées en recombinant les mêmes concepts.

La thèse selon laquelle la pensée doit avoir une « syntaxe » a été critiquée par les partisans du *connexionnisme* (cf. Bechtel 1990). Un système connexionniste repose sur un

certain nombre d'unités simples de traitement de l'information entre lesquelles s'établissent des relations dynamiques de forces diverses. Si ces unités sont considérées comme les véhicules de contenus mentaux, leur structure n'a pas besoin de reproduire la structure logico-sémantique de ces contenus. La question pertinente est de savoir si l'architecture connexionniste d'un système cognitif dont la complexité approche celle de notre esprit ne réalisera pas en définitive un langage de la pensée.

On a critiqué le *formalisme* inhérent au modèle computationnel de l'esprit. La forme des phrases mentales, sur laquelle Fodor fait reposer l'inférence, est supposée être spécifiable en termes purement physiques. Or Davidson (1982) a affirmé que les relations rationnelles n'ont aucun écho dans la théorie physique ; elles ne sont pas *codifiables* (cf. section VI). Une transition causale entre deux états mentaux particulier peut instancier une relation rationnelle, mais on ne peut pas réduire, de manière générale, les relations rationnelles à des relations causales. Même les relations rationnelles de type déductif ne sont pas réductibles à des transitions purement causales, indépendamment des contenus engagés dans la déduction (McDowell 1985). On a également fait valoir que l'inférence mentale est *située*, et ne repose pas uniquement sur la forme de phrases mentales (Barwise et Perry 1983). Pour Searle (1983, 1992), elles dépendent aussi d'un *arrière-plan* de capacités non-représentatives dont le fonctionnement ne s'explique pas entièrement en termes de règles formelles (Searle 1983, 1992).

Par ailleurs, la réduction du rôle inférentiel des états mentaux à leur rôle causal pose de manière particulièrement aiguë le problème de la relation entre la syntaxe et la sémantique de la pensée. Comme le fait observer Searle, les propriétés syntaxiques sont relatives à l'observateur ; la sémantique n'est jamais intrinsèque à la syntaxe (c'est la portée de l'*argument de la chambre chinoise* in 1992). Il reste donc à montrer que la syntaxe de la pensée est liée de manière non-triviale à sa sémantique constitutive. Il n'est pas évident qu'une hypothèse d'« harmonie préétablie » entre les rôles causal et sémantique de nos états mentaux (Perry 1999) rende justice à notre intuition selon laquelle une inférence mentale s'articule sur des *contenus mentaux*, et pas seulement sur des formes syntaxiques. (C'est un autre aspect du problème de la causalité mentale.)

VI. Les formes de l'anti-réductionnisme

Comme nous l'avons vu, plusieurs éléments supposés constitutifs du mental semblent résister au naturalisme physicaliste : les *qualia*, les contenus de représentation, la causalité mentale et les relations rationnelles entre états mentaux. Un certain nombre de doutes se sont ainsi élevés contre le projet même de naturalisation, et plusieurs conceptions anti-naturalistes, ou du moins anti-réductionnistes, ont été proposées.

Les conceptions anti-réductionnistes partent souvent du même constat : les lois de la psychologie populaire ont un statut tout à fait différent des lois physiques. Davidson affirme ainsi qu'il n'y a pas de lois psychologiques *strictes* ; toute généralisation psychologique comporte une clause *ceteris paribus*, c'est-à-dire ne vaut que dans certaines conditions normales, qui ne peuvent d'ailleurs pas être spécifiées sans réutiliser le vocabulaire psychologique. (Il en va de même des lois psycho-physiques.) La question est de savoir quelles sont les conséquences du statut particulier des lois psychologiques sur notre conception de la *réalité* des phénomènes mentaux (Engel 1996). L'anti-réductionnisme peut être réaliste ou anti-réaliste.

Une forme radicale d'anti-réalisme est le *non-factualisme* attribué à Wittgenstein par Kripke (1982), selon lequel il n'y a pas de *faits* (réels ou objectifs) concernant les règles et la signification, et donc les contenus mentaux.

L'*éliminativisme* est une autre forme d'anti-réalisme (cf. Churchland 1990). C'est une position fortement physicaliste : les phénomènes mentaux ne sont pas physiques, *donc* ils n'existent pas ; ils doivent être *éliminés* de notre conception objective du monde. Les concepts mentaux de la psychologie populaire ont un statut analogue à celui des concepts de sorcellerie et de phlogiston ; ils sont condamnés à disparaître au profit d'une théorie scientifique plus rigoureuse (dans le cas qui nous intéresse, les neurosciences). L'*éliminativisme* est, comme le dit Fodor, tout simplement difficile à *croire* ; il repose par ailleurs sur la thèse contestable selon laquelle la psychologie populaire est une théorie *empirique* (Engel, 1992 : ch. 2).

D'autres positions anti-réalistes moins radicales ont été proposées. Selon Dennett, les explications de la psychologie populaire ne sont pas descriptives ; elles impliquent des règles normatives à la lumière desquelles nous interprétons le comportement d'autrui. Dennett rejette la distinction absolue entre l'intentionnalité réelle ou intrinsèque et l'intentionnalité dérivée, comme celle que l'on peut attribuer à un programme qui joue aux échecs. L'intentionnalité n'existe que par la *posture intentionnelle* que nous adoptons en tant

qu'interprètes pour rendre compte, à un niveau commode de description, du comportement d'organismes ou d'artefacts. Les croyances et les désirs ne sont pas *réellement* des états de quoi que ce soit, bien que leur attribution fondée présuppose, outre des hypothèses normatives d'optimalité, une certaine complexité réelle du comportement décrit.

Kripke (1980) se range du côté des réalistes non-physicalistes. Il présente un certain nombre de considérations modales contre les théories de l'identité dans les deux versions distinguées plus haut (section I). Intuitivement, la sensation subjective de chaleur n'est liée que de manière contingente à un certain degré d'agitation moléculaire. Mais si les états mentaux ne sont pas *nécessairement* identiques aux états physiques, ils ne le leur sont pas identiques du tout.

Pour Davidson (1982), l'anti-réalisme n'est pas une conséquence nécessaire de l'irréductibilité de l'explication psychologique à l'explication physique. Davidson défend une version de la thèse de l'identité-occurrence, le *monisme anomal*. Le monisme anomal repose sur trois prémisses. Premièrement, des relations causales s'établissent entre des états mentaux (par exemple, des croyances et des désirs) et des états physiques (par exemple, des mouvements musculaires). Deuxièmement, toute relation causale entre deux événements suppose l'existence d'une loi stricte les concernant. Troisièmement, il n'y a pas de lois psychophysiques strictes (les concepts mentaux sont anomaux). Il s'ensuit que chaque état ou événement mental est également physique.

L'identité des états mentaux d'un sujet est entièrement fixée par ce qu'une *théorie de l'interprétation radicale* peut en dire. (La notion d'interprétation radicale est inspirée de celle de traduction radicale de Quine 1960). L'interprétation du comportement linguistique et non-linguistique du sujet est fondée sur des principes constitutifs et *a priori* de rationalité, comme le *principe de charité*. Ce principe exige que nous tirions du comportement du sujet interprété le maximum de cohérence et de vérité. La fausseté et l'irrationalité n'ont de sens que sur un fond de vérité et de rationalité. *Avoir* une croyance, c'est avoir l'aptitude à reconnaître cette croyance chez autrui. Toute créature qui a des croyances a donc le *concept* de croyance. La *triangulation*, c'est-à-dire ce processus d'interprétation qui engage deux sujets en communication mutuelle à propos d'un tiers objet, est une condition minimale nécessaire de l'*existence* d'états mentaux (Davidson 1991).

Les théories de l'interprétation de Dennett et de Davidson sont parfois opposées à la *théorie de la simulation*. Selon cette théorie, la connaissance que nous avons des états mentaux n'a pas exclusivement la forme d'un ensemble articulé d'hypothèses théoriques. Elle repose plutôt sur la simulation, une forme de savoir-faire basée sur l'expérience (Goldman

1993). La simulation est un processus qui nous permet d'exécuter des processus mentaux « hors circuit », c'est-à-dire indépendamment de la présence d'entrées perceptives et de sorties comportementales réelles. Lorsque nous attribuons à autrui, et peut-être à nous-mêmes, des états mentaux, nous tâchons de nous mettre « dans sa peau » pour prédire ou expliquer son comportement (Gordon 1995). Dans la théorie de l'interprétation, l'existence d'une croyance présuppose l'attribution possible de celle-ci ; dans la théorie de la simulation, c'est l'attribution de la croyance qui présuppose son existence, réelle ou simulée.

Selon le *naturalisme biologique* de Searle (1992), les phénomènes mentaux font partie de notre histoire biologique au même titre que la digestion ou la reproduction. Ils sont causés par des processus neurobiologiques dans le cerveau. Ce sont des traits cérébraux d'ordre supérieur – de même que la solidité de la table est un trait d'ordre supérieur causé par le comportement d'éléments d'ordre inférieur : les molécules dont la table est entièrement composée. Searle partage la conviction des physicalistes selon laquelle tous les phénomènes sont physiques, mais refuse d'admettre que les phénomènes mentaux sont réductibles à d'autres phénomènes physiques. Pour McDowell (1994), une telle position ne rend pas justice à la dimension essentiellement normative des concepts mentaux. Le réalisme est de rigueur : les phénomènes mentaux sont naturels. Mais la nature ne se résume pas à la nature physique ; elle inclut notre *seconde nature*. Les être humains, contrairement aux autres animaux, acquièrent une seconde nature en maîtrisant un système de concepts par lequel ils entrent en relation normative avec le monde.

REFERENCES

Barwise J. et Perry J., 1983, *Situations and Attitudes*, Cambridge (Mass.) : MIT Press.

Bechtel W., 1990, « Connectionism and the Philosophy of Mind : An Overview », in Lycan (1990).

Block N., 1994, « *Qualia* », in Guttenplan (1994).

Braddon-Mitchell D. et Jackson F., 1996, *Philosophy of mind and cognition*, Oxford : Blackwell.

Brentano F., 1924-1928, *Psychologie vom Empirischem Standpunkt*, 3 vol., Leipzig : Felix Meiner. Trad. fr. M. de Gandillac, *Psychologie du point de vue empirique*, Paris : Aubier-Montaigne, 1944.

Burge T., 1979, « Individualism and the Mental », in *Midwest Studies in Philosophy* IV.

Burge T., 1988, « Individualism and Self-Knoweldge », *Journal of Philosophy* 85.

Burwood S., Gilbert P. et Lennon K., 1999, *Philosophy of Mind*, London : UCL Press.

Campbell J., 1994, *Past, Space, and Self*, Cambridge (Mass.) : MIT Press.

Churchland P. M., 1990, « Neurophilosophy and Connectionism », in Lycan (1990).

Clémentz F., 1997, « *Qualia* et contenus perceptifs », in J. Proust (dir.), *Perception et intermodalité*, Paris : PUF.

Corazza E., 1995, *Propositions, Contexte et Attitudes*, Montréal/Paris: Vrin/Bellarmin.

Corazza E. et Dokic J., 1996, « Un aspect du cartésianisme en philosophie de l'esprit », *Studia Philosophica*, 55, 1996.

Davidson D., 1984, *Inquiries into Truth and Interpretation*, Oxford : Clarendon Press, 1984. Trad. fr. P. Engel, *Enquêtes sur la vérité et l'interprétation*, Nîmes : Chambon, 1993.

Davidson D., 1987, « Knowing one's own Mind », *Proceedings of the Aristotelian Societey*, 60.

Davidson D., 1991, « Animaux rationnels », in *Paradoxes de l'irrationalité*, trad. fr. P. Engel, Combas : l'Eclat.

Dennett D., 1983, *The Intentional Stance*, Cambridge (Mass.) : MIT Press. Trad. fr. P. Engel, *La stratégie de l'interprète*, Paris : Gallimard, 1990.

Dennett D., 1991, *Consciousness Explained*, New York : Littlebrown. Trad. fr. P. Engel, *La conscience expliquée*, Paris : Odile Jacob, 1994.

Dretske F., 1981, *Knowledge and the Flow of Information*, Cambridge (Mass.) : MIT Press.

Dretske F., 1995, *Naturalizing the Mind*, Cambridge (Mass.) : MIT Press.

Engel P., 1992, *Etats d'esprit. Questions de philosophie de l'esprit*, Aix-en-Provence : Alinea.

Engel P., 1996, *Philosophie et psychologie*, Paris : Gallimard (Folio).

Evans G., 1982, *The Varieties of Reference*, Oxford : Blackwell.

Fodor J., 1968, *Psychological Explanation*, New York : Random House.

Fodor J., 1975, *The Language of Thought*, New York : Thomas Crowell.

Fodor J., 1981, *Re-presentations*, Cambridge (Mass.) : MIT Press.

Fodor J., 1990, *A Theory of Content*, Cambridge (Mass.) : MIT Press.

Frege G., 1971, *Ecrits logiques et philosophiques*, trad. fr. C. Imbert, Paris: Seuil.

Goldman A., 1993, « The Psychology of Folk Psychology », *Behavioral and Brain Sciences* 16.

Gordon R. M., 1995, « Folk Psychology as Simulation », in M. Davies et T. Stone (eds), *Mental Simulation*, Oxford : Blackwell.

Grice P., 1957, « Meaning », in *Studies in the Way of Words*, Cambridge (Mass.) : Harvard UP.

Guttenplan S. (ed.), 1994, *A Companion to the Philosophy of Mind*, Oxford : Blackwell.

Harrison B., 1973, *Form and Content*, Oxford : Blackwell.

Jacob P., 1997, *Pourquoi les choses ont-elles un sens ?*, Paris : Odile Jacob.

Jackson F., 1990, « Epiphenomenal Qualia », in Lycan (1990).

Kim J., 1994, « Supervenience », in Guttenplan (1994).

Kirk R., 1994 *Raw Feeling*, Oxford : Clarendon Press.

Kripke S., 1980, *Naming and Necessity*, Cambridge (Mass.) : MIT Press. Trad fr. P. Jacob et F. Recanati, *La logique des noms propres*, Paris : Minuit, 1982.

Kripke S., 1982, *Wittgenstein on Rules and Private Language*, Oxford : Blackwell. Trad. fr. Paris : Seuil, 1995.

Lewis D., 1966, « An Argument for the Identity-Theory », in *Philosophical Papers I*, Oxford : OUP.

Loar B., 1981, *Mind and Meaning*, Cambridge : CUP.

Lycan W. (ed.), 1990, *Mind and Cognition*, Oxford : Blackwell.

Lycan W., 1996, *Consciousness and Experience*, Cambridge (Mass.) : MIT Press.

McDowell J., 1984, « *De Re* Senses », in C. Wright (ed.), *Frege: Tradition and Influence*, Oxford : Blackwell.

McDowell J., 1985, « Functionalism and anomalous monism », in E. LePore et B. McLaughlin (eds), *Actions and events : perspectives on the philosophy of Donald Davidson*, Oxford : Blackwell.

McDowell J., 1986, «Singular Thought and the Extent of Inner Space», in P. Pettit et J. McDowell (eds), *Subject, Thought, and Context*, Oxford : Clarendon Press.

McDowell J., 1994, *Mind and World*, Cambridge (Mass.) : Harvard UP.

McGinn C., 1996, *The Character of Mind. An Introduction to the Philosophy of Mind*, 2nd edition, Oxford : OUP.

Millikan R. G., *Language, Thought, and Other Biological Categories*, Cambridge (Mass.) : MIT Press.

Moore G. E. M., 1903, *Principia Ethica*. Trad. fr. M. Gouverneur, Paris : PUF, 1998.

Nagel T., 1979, *Mortal Questions*, Cambridge : CUP. Trad. fr. P. Engel et C. Engel-Tiercelin, *Questions mortelles*, Paris : PUF, 1983.

Pacherie E., 1993, *Naturaliser l'intentionnalité*, Paris : PUF.

Perry J., 1999, *Problèmes d'indexicalité*, Stanford : CSLI.

Putnam H., 1990, « The Nature of Mental States », in Lycan (1990).

Putnam H., 1975, « The Meaning of 'Meaning' », in *Philosophical Papers II*, Cambridge : CUP.

Ramsey F., 1978, *Foundations*, London : Routledge.

Recanati F., 1993, *Direct Reference. From Language to Thought*, Oxford: Blackwell.

Rosenthal D., 1993, « Thinking that One Thinks », in M. Davies et G. W. Humphreys (eds), *Consciousness*, Oxford : Blackwell.

Proust J., 1997, *Comment l'esprit vient aux bêtes. Essai sur la représentation*, Paris : Gallimard.

Quine W. V. O., 1960, *Word and Object*, Cambridge (Mass.) : MIT Press. Trad. fr. P. Gochet et J. Dopp, Paris : Flammarion, 1980.

Searle J., 1983, *Intentionality*, Cambridge : CUP. Trad. fr. C. Pichevin, Paris: Minuit, 1987.

Searle J., 1992, *The Rediscovery of the Mind*, Cambridge (Mass.) : MIT Press. Trad. fr. C. Engel-Tiercelin, *La redécouverte de l'esprit*, Paris : Gallimard, 1994.

Wittgenstein L., 1953, *Philosophical Investigations*, éd. par E. Anscombe et R. Rhees, Oxford. Trad. fr. P. Klossowski, *Investigations philosophiques*, Paris : Gallimard, 1961.