J. SEARLE
La Redécouverte de l'esprit
Gallimard 1995

CHAPITRE II

L'histoire récente du matérialisme : une répétition de la même erreur

Le mystère du matérialisme

À quoi se ramène au juste la doctrine connue sous le nom de « matérialisme » ? À la thèse, pourrait-on penser, selon laquelle la microstructure du monde est entièrement faite de particules matérielles. Mais le problème est que cette thèse est compatible avec n'importe quelle philosophie de l'esprit ou presque, à l'exception peut-être de la conception cartésienne qui veut qu'en plus des particules physiques, il existe des âmes « immatérielles » ou des substances mentales, des entités spirituelles qui survivent, immortelles, à la destruction de nos corps. De nos jours, en revanche, pour autant que je le sache, personne ne croit en l'existence de substances spirituelles immortelles, si ce n'est pour des raisons religieuses. Il n'y a, à ma connaissance, aucune motivation purement philosophique ou scientifique qui fasse que l'on accepte l'existence de substances mentales immortelles. Si on laisse donc de côté l'opposition à la croyance religieusement motivée en l'immortalité de l'âme, la question qui demeure est celle-ci : à quoi au juste le matérialisme se ramène-t-il en philosophie de l'esprit ? À quelles conceptions est-il censé s'opposer ?

Si on lit les premiers travaux de nos contemporains qui se décrivent eux-mêmes comme des matérialistes –

J. J. C. Smart (1965), U. T. Place (1956) et D. Armstrong (1968), par exemple – il paraît clair que, lorsqu'ils affirment l'identité du mental et du physique, ils soutiennent quelque chose de plus que le simple rejet du dualisme cartésien des substances. Il me semble que leur désir est de nier l'existence de tout phénomène irréductiblement mental quel qu'il soit dans le monde. Ils veulent nier l'existence de propriétés phénoménologiques irréductibles telles que la conscience ou les *qualia*. Or, pourquoi tiennent-ils tant à nier l'existence de phénomènes mentaux intrinsèques irréductibles ? Pourquoi ne concèdent-ils pas tout simplement que ces propriétés sont des propriétés biologiques ordinaires, de niveau supérieur, de systèmes neurophysiologiques tels que des cerveaux humains ?

La réponse à cette question est extrêmement complexe. Elle a trait en partie au fait qu'ils acceptent les catégories cartésiennes traditionnelles comme le vocabulaire afférent et ce qu'il implique. À cet égard, admettre l'existence et l'irréductibilité des phénomènes mentaux reviendrait à admettre une forme ou une autre de cartésianisme. Dans leurs termes, il pourrait s'agir d'un « dualisme des propriétés » plutôt que d'un « dualisme des substances », mais de leur point de vue, le dualisme des propriétés serait tout aussi contradictoire avec le matérialisme que le dualisme des substances. Chacun aura pour lors compris que je suis contre les présupposés sous-jacents à leur conception. J'aimerais souligner, sans relâche, que l'on peut accepter les faits évidents de la physique – par exemple, que le monde est entièrement fait de particules physiques dans des champs de force – sans en même temps nier les faits évidents relatifs à nos propres expériences – par exemple, que nous sommes tous conscients et que nos états conscients ont des propriétés phénoménologiques *irréductibles* tout à fait spécifiques. L'erreur est de supposer que ces deux thèses sont contradictoires, et cette erreur provient de ce que l'on accepte les présupposés inhérents au vocabulaire traditionnel. Que les choses soient

claires : je ne défends pas une forme de dualisme. Je rejette le dualisme des propriétés comme celui des substances ; mais précisément les mêmes raisons qui me font rejeter le dualisme me font rejeter tout autant le matérialisme et le monisme. La grossière erreur est de supposer que l'on doive choisir entre ces positions.

C'est faute de voir que mentalisme naïf et physicalisme naïf sont compatibles qu'ont pu surgir ces discussions très curieuses, au début de l'histoire relative au problème qui nous intéresse, au cours desquelles des auteurs essaient de trouver un vocabulaire « topiquement neutre » ou d'éviter quelque chose qu'ils appellent les « baladeurs nomologiques » (*nomological danglers*) (Smart, 1965). On notera au passage que personne ne songe à dire, par exemple, que la digestion doit être décrite dans un vocabulaire « topiquement neutre ». Personne n'éprouve le besoin de dire : « Quelque chose se passe en moi qui ressemble à ce qui se passe quand je digère une pizza. » Bien qu'on éprouve le besoin de dire : « Il y a quelque chose qui se passe en moi qui ressemble à ce qui se passe lorsque je vois une orange. » Ce qu'il faut, c'est essayer de trouver une description des phénomènes qui ne fasse aucun usage du vocabulaire mentaliste. Mais où cela mène-t-il ? Les faits demeurent les mêmes. Le fait est que les phénomènes mentaux ont des propriétés mentalistes, tout comme ce qui se passe dans mon estomac a des propriétés digestives. On ne se débarrasse pas de ces propriétés en se contentant de trouver un autre vocabulaire. Le souhait des philosophes matérialistes est de nier l'existence des propriétés mentales sans nier la réalité de *certains* phénomènes qui sous-tendent l'usage de notre vocabulaire mentaliste. Il leur faut donc trouver un autre vocabulaire pour décrire les phénomènes [1]. Mais à mon avis, tout cela est une vaste perte de temps. On devrait se contenter de reconnaître les phénomènes mentaux (et partant, physiques) d'entrée de jeu, de

la même manière qu'on reconnaît les phénomènes digestifs dans l'estomac.

Si je retrace de manière cursive l'histoire du matérialisme telle qu'elle s'est déroulée depuis un demi-siècle, je mettrai en évidence le schème plutôt surprenant mais très révélateur d'arguments et de contre-arguments qui ont eu cours dans la philosophie de l'esprit depuis le positivisme des années 30. Ce schème n'est pas toujours visible à l'œil nu. Il n'est pas même visible à l'œil nu qu'on ait affaire aux mêmes questions. Je crois pourtant qu'il y a bien eu, contrairement aux apparences, un seul sujet majeur de discussion dans la philosophie de l'esprit des cinquante dernières années ou presque : le problème des rapports du corps et de l'esprit. Souvent les philosophes prétendent parler de quelque chose d'autre – de l'analyse de la croyance, par exemple, ou bien de la nature de la conscience –, mais on s'aperçoit presque invariablement que les caractéristiques spécifiques de la croyance ou de la conscience ne les intéressent pas vraiment. Ils ne s'intéressent pas à la manière dont croire diffère de supposer et d'émettre des hypothèses ; ce qu'ils veulent plutôt éprouver, c'est leurs convictions à propos du problème des rapports du corps et de l'esprit sur l'exemple *de la croyance.* Il en va de même avec la conscience. On discute peu, curieusement, de la conscience comme telle ; les matérialistes voient plutôt en la conscience un « problème » particulier qui se pose à une théorie matérialiste de l'esprit. En d'autres termes, ils veulent trouver une manière de « s'occuper » de la conscience, étant donné leur matérialisme [2].

Le schéma que l'on retrouve presque invariablement dans ces discussions est le suivant : un philosophe avance une théorie matérialiste de l'esprit. Il le fait, avec l'intime conviction qu'une version ou une autre de la théorie matérialiste de l'esprit doit être la bonne – après tout, ne savons-nous pas, grâce aux découvertes de la science, qu'il n'y a vraiment rien d'autre dans l'univers que des particules physiques et

des champs de force agissant sur des particules physiques ? Or, à l'évidence, on doit pouvoir donner une explication des êtres humains qui soit compatible et cohérente avec notre explication de la nature en général. Et ne s'ensuit-il pas, à l'évidence, de tout cela que notre explication des êtres humains doit être un matérialisme intégral ? Voilà donc notre philosophe en quête d'une explication matérialiste de l'esprit. C'est alors qu'il rencontre des difficultés. Il semble toujours qu'il laisse quelque chose de côté. Si l'on s'en tient au schéma général de la discussion, on peut voir que les critiques de la théorie matérialiste de l'esprit prennent ordinairement une forme plus ou moins technique ; mais en réalité, sous les objections techniques, se cache une objection plus profonde, laquelle peut se formuler plus simplement comme suit : la théorie en question a laissé de côté l'esprit ; elle a laissé de côté quelque caractéristique essentielle de l'esprit, comme la conscience ou les « *qualia* » ou le contenu sémantique. On rencontre sans cesse ce schéma. Une thèse matérialiste est avancée. Mais la thèse rencontre des difficultés ; les difficultés prennent différentes formes, mais elles sont toujours la manifestation d'une difficulté sous-jacente plus profonde, à savoir que la thèse en question nie des faits évidents que nous connaissons tous sur notre esprit. Et cela conduit à des efforts encore plus effrénés pour s'accrocher à la thèse matérialiste et essayer de contrer les arguments avancés par ceux qui s'obstinent à vouloir préserver les faits. Après quelques années de manœuvres désespérées pour expliquer les difficultés, on avance tel ou tel nouveau développement qui est censé résoudre les difficultés, quitte à s'apercevoir alors que de nouvelles difficultés surgissent, à ceci près qu'elles ne sont pas si nouvelles que cela – ce sont les mêmes que par le passé.

Si nous considérions la philosophie de l'esprit des cinquante dernières années sous la forme d'un seul et unique individu, nous dirions de lui qu'il est atteint de névrose obsessionnelle,

et que sa névrose prend la forme d'une compulsion à la répétition du même schéma comportemental. D'après l'expérience que j'en ai, on ne peut pas guérir une névrose en l'attaquant de front. Il ne suffit pas de mettre le doigt sur les erreurs logiques qui sont commises. La réfutation directe conduit simplement à une répétition du schéma comportemental névrotique. Nous devons donc aller voir derrière les symptômes et chercher les présupposés inconscients qui ont induit dès le départ le comportement. Je suis à présent convaincu, après plusieurs années de discussions sur ces questions, qu'à de très rares exceptions près, toutes les parties prenantes dans les débats qui ont cours aujourd'hui en philosophie de l'esprit sont prisonnières d'un certain ensemble de catégories verbales. Elles sont en proie à une certaine terminologie, une terminologie qui remonte à Descartes si ce n'est plus avant, et pour réussir à nous rendre maîtres du comportement obsessionnel, nous aurons à examiner les origines inconscientes des controverses. Il nous faudra essayer de mettre à nu ce que tout le monde tient pour acquis et qui contribue à alimenter incessamment la controverse.

Qu'on ne se méprenne pas : l'utilisation que je fais de l'analogie thérapeutique n'implique pas que je prenne en règle générale pour argent comptant les modes d'explication psychanalytiques en matière intellectuelle. Transformons donc la métaphore thérapeutique comme suit : mon entreprise présente ressemble un peu à celle d'un anthropologue qui s'emploierait à décrire le comportement exotique d'une tribu lointaine. La tribu fait montre d'un ensemble de schémas comportementaux et d'une métaphysique qu'il nous faut essayer de mettre à nu et de comprendre. Il est facile de se moquer des vieilleries de la tribu des philosophes de l'esprit, et je dois avouer que je n'ai pas toujours réussi à résister à cette tentation. Mais au début, au moins, je dois souligner que la tribu c'est nous – nous sommes les dépositaires des présupposés métaphysiques qui rendent possible le comportement de la tribu. Aussi voudrais-je, avant de proposer vraiment une analyse et une critique du comportement de la tribu, présenter l'idée qui fait réellement partie de notre culture scientifique contemporaine. Et pourtant, je soutiendrai plus loin que cette idée est incohérente ; ce n'est qu'un symptôme du même schéma névrotique.

L'idée est celle-ci : nous pensons que la question : comment est-il possible à des morceaux de matière dépourvus d'intelligence de produire de l'intelligence ? doit avoir un sens. Comment est-il possible aux morceaux de matière dépourvus d'intelligence qui composent notre cerveau de produire le comportement intelligent que nous produisons tous ? Voilà, nous semble-t-il, une question parfaitement intelligible. À la vérité, cela paraît être un projet de recherche très valable, et c'est en fait un projet de recherche qui est mené sur une large échelle [3] et, ajouterais-je, très bien subventionné.

Comme nous trouvons la question intelligible, nous trouvons plausible la réponse suivante : les morceaux de matière dénués d'intelligence peuvent produire de l'intelligence en raison de leur *organisation*. Les morceaux de matière dénués d'intelligence sont *organisés* d'une certaine manière dynamique, et c'est l'organisation dynamique qui est constitutive de l'intelligence. En vérité, nous pouvons réellement reproduire artificiellement la forme de l'organisation dynamique qui rend possible l'intelligence. La structure sous-jacente de cette organisation s'appelle un « ordinateur », le projet de programmation de l'ordinateur s'appelle l'« intelligence artificielle » ; et lorsqu'il fonctionne, l'ordinateur produit de l'intelligence parce qu'il réalise le bon programme d'ordinateur avec les bonnes entrées et les bonnes sorties.

Cette histoire ne vous paraît-elle pas à tout le moins plausible ? On peut lui donner une parfaite allure de plausibilité à mes yeux, et si elle ne vous paraît pas même lointainement plausible, c'est probablement que vous n'êtes pas un membre parfaitement bien intégré à notre culture

intellectuelle contemporaine. Assurément, je montrerai plus loin que la question comme la réponse sont incohérentes. Lorsque nous posons la question et donnons la réponse en ces termes, nous n'avons pas, en réalité, la moindre idée de ce dont nous parlons. Mais je présente ici cet exemple parce que je veux qu'il vous paraisse naturel, et même prometteur, en tant que projet de recherche.

Donc, l'histoire du matérialisme philosophique au xx^e^ siècle révèle un curieux schéma, traversé par une tension récurrente entre le besoin du matérialiste de donner une explication des phénomènes mentaux qui ne fasse aucune référence à quoi que ce soit d'intrinsèquement ou d'irréductiblement mental, d'un côté, et, de l'autre, l'exigence intellectuelle générale que rencontre tout chercheur qui est de ne rien dire qui soit manifestement faux. On va le voir dans ce bref aperçu, aussi neutre et objectif que possible, du schéma de thèses et de réponses dont les matérialistes sont l'incarnation. J'entends justifier les thèses que j'ai soutenues au premier chapitre en donnant des illustrations réelles des tendances que j'ai identifiées.

Le behaviorisme

Au commencement était le behaviorisme. Le behaviorisme s'est présenté sous deux espèces : le « behaviorisme méthodologique » et le « behaviorisme logique ». Le behaviorisme méthodologique est une stratégie de recherche en psychologie d'après laquelle une science de la psychologie doit consister en la découverte des corrélations qui existent entre les entrées de stimuli et les sorties comportementales (Watson, 1925). Une science empirique rigoureuse, selon cette théorie, ne fait absolument pas référence à de mystérieux éléments introspectifs ou mentalistes, quels qu'ils soient.

Le behaviorisme logique va même encore plus loin et

souligne qu'il n'y a pas d'éléments de ce genre auxquels faire référence, si ce n'est dans la mesure où ils existent sous forme de comportement. Selon le behaviorisme logique, c'est une question de définition, une question d'analyse logique, que les termes mentaux puissent se définir en termes de comportement, que les phrases sur l'esprit puissent se traduire intégralement en des phrases sur le comportement (Hempel, 1949 ; Ryle, 1949). Selon le behavioriste logique, bon nombre de phrases de la traduction seront de forme hypothétique, parce que les phénomènes mentaux en question consistent non en trames comportementales se produisant effectivement, mais plutôt en dispositions comportementales. Ainsi, d'après l'explication behavioriste classique, dire que Jean croit qu'il va pleuvoir c'est simplement dire que Jean sera disposé à fermer les fenêtres, à ranger les outils de jardinage et à emporter un parapluie s'il sort. Sur le mode matériel du discours, le behaviorisme soutient que l'esprit n'est que comportements et dispositions comportementales. Sur le mode formel, cela consiste à soutenir que les phrases portant sur les phénomènes mentaux peuvent se traduire en des phrases portant sur un comportement réel et possible.

Les objections faites au behaviorisme peuvent se diviser en deux catégories : des objections de sens commun et des objections plus techniques. Une évidente objection de sens commun est que le behavioriste semble laisser à l'écart les phénomènes mentaux en question. Il ne reste rien dans l'explication behavioriste de l'expérience subjective consistant à penser et sentir ; celle-ci se ramène à des trames comportementales objectivement observables.

Plusieurs objections plus ou moins techniques ont été faites au behaviorisme logique. En premier lieu, les behavioristes n'ont jamais réussi à donner un sens parfaitement clair à l'idée de « disposition ». Personne n'est jamais parvenu à donner une explication satisfaisante du genre d'antécédents qu'il devrait y avoir dans les énoncés hypothétiques pour

produire une analyse dispositionnelle adéquate des termes mentaux en termes dispositionnels (Hampshire, 1950 ; Geach, 1957). En second lieu, il semble qu'il y ait un problème concernant une certaine forme de circularité dans l'analyse : pour donner une analyse de la croyance en termes comportementaux, on doit apparemment faire référence au désir ; pour donner une analyse du désir, on doit apparemment faire référence à la croyance (Chisholm, 1957). Ainsi, pour reprendre notre précédent exemple, nous essayons d'analyser l'hypothèse que Jean croit qu'il va pleuvoir en termes de l'hypothèse que si les fenêtres sont ouvertes, Jean va les fermer, et autres hypothèses du même genre. Nous voulons analyser l'énoncé catégorique que Jean croit qu'il va pleuvoir en termes de certains énoncés hypothétiques sur ce que Jean fera dans certaines conditions. Toutefois, la croyance de Jean qu'il va pleuvoir ne se manifestera dans le comportement consistant à fermer les fenêtres que si nous postulons d'autres hypothèses telles que Jean ne veut pas que l'eau de pluie puisse entrer par les fenêtres et Jean croit que le fait que les fenêtres sont ouvertes permet à l'eau de pluie d'entrer. S'il n'y a rien qu'il aime tant que de l'eau de pluie se déversant par les fenêtres, il ne sera pas disposé à les fermer. Sans une telle hypothèse sur les désirs de Jean (et sur ses autres croyances), nous ne pouvons, semble-t-il, commencer à analyser la moindre phrase portant sur les croyances qu'il avait au départ. Des remarques similaires peuvent s'appliquer à l'analyse des désirs ; ces analyses semblent exiger que l'on fasse référence à des croyances.

Une troisième objection technique au behaviorisme était qu'il laissait de côté les relations causales entre états mentaux et comportement (Lewis, 1966). En identifiant, par exemple, la douleur à la disposition au comportement de douleur, le behaviorisme laisse de côté le fait que la douleur *cause* le comportement. Pareillement, si nous essayons d'*analyser* les croyances et les désirs en termes comportementaux, nous ne

sommes plus en mesure de dire que croyances et désirs *causent* le comportement.

Bien que la plupart peut-être des discussions techniques dans la littérature philosophique aient trait aux objections « techniques », ce sont en fait les objections de sens commun qui sont les plus embarrassantes. L'absurdité du behaviorisme tient à ce qu'il nie l'existence d'états mentaux internes qui s'ajouteraient au comportement externe (Ogden et Richards, 1926). Ce qui, nous le savons, va totalement à l'encontre de notre expérience de l'effet que cela fait d'être un être humain. Pour cette raison, les behavioristes ont déchaîné les sarcasmes et ont été accusés de « feindre l'anesthésie [4] » ; ils ont aussi été la cible d'un certain nombre de mauvaises plaisanteries (du genre de celle-ci : un behavioriste à un autre behavioriste, juste après avoir fait l'amour : « C'était formidable pour toi, comment était-ce pour moi ? »). On a parfois présenté cette objection de sens commun au behaviorisme sous la forme d'arguments faisant appel à nos intuitions. L'une d'entre elles est l'objection du superacteur/superspartiate (Putnam, 1963). On peut sans peine imaginer un acteur particulièrement doué qui donnerait une imitation parfaite du comportement de quelqu'un qui souffre, bien que l'acteur en question ne souffre pas du tout, et l'on peut aussi imaginer un superspartiate capable d'endurer de la douleur sans en manifester le moindre signe extérieur.

Les théories de l'identité de type à type

Le behaviorisme logique était censé être une vérité analytique. Il affirmait une connexion définitionnelle entre les concepts mentaux et comportementaux. Dans l'histoire récente des philosophies matérialistes de l'esprit, il s'est vu remplacé par la « théorie de l'identité » : c'est un fait contingent, synthétique, empirique que les états mentaux sont

identiques à des états du cerveau et du système nerveux central (Place, 1956 ; Smart, 1965). Selon les théoriciens de l'identité, il n'y avait rien d'absurde, logiquement parlant, à supposer qu'il pût y avoir des phénomènes mentaux distincts, indépendants de la réalité matérielle ; il s'est seulement trouvé qu'en fait, nos états mentaux, les douleurs par exemple, sont identiques à des états de notre système nerveux. Et en ce cas, on a soutenu que les douleurs sont identiques à des stimulations des fibres-C [5]. Descartes *aurait* pu avoir raison en considérant qu'il existe des phénomènes mentaux distincts ; il s'est seulement trouvé qu'*en fait* il avait tort. Les phénomènes mentaux n'étaient rien que des états du cerveau et du système nerveux central. L'identité entre l'esprit et le cerveau était censée être une identité empirique, tout comme l'identité entre l'éclair et les décharges électriques (Smart, 1965), ou l'identité entre l'eau et les molécules d'H_2O (Feigl, 1958 ; Shaffer, 1961), étaient censées être des entités empiriques et contingentes. Mais il s'est trouvé qu'en fait, comme l'a établi la découverte scientifique, les éclairs ne sont que des flux d'électrons, et que l'eau sous toutes ses formes n'est rien d'autre que des collections de molécules d'H_2O.

Comme dans le cas du behaviorisme, on peut diviser les difficultés de la théorie de l'identité entre les objections « techniques » et les objections de sens commun. En ce cas, l'objection de sens commun prend la forme d'un dilemme. Supposons que la théorie de l'identité soit, comme le prétendent ses défenseurs, une vérité empirique. S'il en est ainsi, alors il doit y avoir des traits logiquement indépendants du phénomène en question qui lui permettent d'être identifié du côté gauche de l'énoncé d'identité d'une manière différente de la manière dont il est identifié du côté droit de l'énoncé d'identité (Stevenson, 1960). Si, par exemple, les douleurs sont identiques à des événements neurophysiologiques, alors il doit y avoir deux ensembles de caractéristiques, des caractéristiques de douleur et des caractéristiques neurophysiologiques, et ces deux ensembles de caractéristiques nous permettent de fixer les deux côtés de l'énoncé d'identité synthétique. Supposons, par exemple, que nous ayons un énoncé de la forme suivante :

> L'événement douloureux x est identique à l'événement neurophysiologique y.

Nous comprenons un tel énoncé parce que nous comprenons qu'un seul et même événement a été identifié en vertu de deux sortes différentes de propriétés, des propriétés de douleur et des propriétés neurophysiologiques. Mais s'il en est ainsi, il nous faut alors faire face à un dilemme : ou bien les caractéristiques de la douleur sont des caractéristiques subjectives, mentales et introspectives, ou bien elles ne le sont pas. Si elles le sont, alors nous ne nous sommes vraiment pas débarrassés de l'esprit. Nous en sommes toujours à une certaine forme de dualisme, à cela près qu'il s'agit d'un dualisme des propriétés plutôt que d'un dualisme des substances. Nous en sommes toujours à un ensemble de propriétés mentales, même si nous nous sommes débarrassés des substances mentales. Si, en revanche, nous essayons de considérer que la « douleur » n'est pas le nom d'une caractéristique mentale subjective de certains événements neurophysiologiques, alors sa signification reste totalement mystérieuse et inexpliquée. Car nous n'avons aucun moyen de spécifier ces caractéristiques mentales subjectives de nos expériences.

On se rend bien compte, je l'espère, qu'il ne s'agit que de la répétition de l'objection de sens commun faite au behaviorisme. Dans le cas présent, nous l'avons mise sous la forme d'un dilemme : le matérialisme dans le style de l'identité ou bien laisse de côté l'esprit ou bien il ne le laisse pas ; s'il le laisse, il est faux ; dans le cas contraire, ce n'est pas du matérialisme.

Les théoriciens australiens de l'identité pensaient avoir une réponse à cette objection. Ils essayèrent de décrire les

prétendues caractéristiques mentales dans un vocabulaire « topiquement neutre ». L'objectif étant de parvenir à une description des caractéristiques mentales qui ne mentionnât pas le fait qu'elles sont mentales (Smart, 1965). On peut sûrement le faire : on peut faire état de douleurs sans faire état du fait qu'il s'agit de douleurs, tout comme on peut faire état d'avions sans faire état du fait qu'il s'agit d'avions. En d'autres termes, on peut faire état d'un avion en disant : « Un certain bien appartenant à Air France », et on peut faire référence à une image après coup de jaune-orange en disant : « Un certain événement qui se passe en moi ressemble à l'événement qui se produit en moi lorsque je vois une orange. » Mais le fait qu'on puisse faire état d'un phénomène sans spécifier ses caractéristiques essentielles ne veut pas dire qu'il n'existe pas et qu'il n'a pas ces caractéristiques essentielles. C'est toujours une douleur ou une image après coup, ou un avion, même si nos descriptions ne réussissent pas à faire état de ces faits.

Une autre objection plus « technique » à la théorie de l'identité était celle-ci : il paraît peu probable que pour tout type d'état mental il y ait un seul et unique type d'état neurophysiologique auquel il soit identique. Même si ma croyance que Denver est la capitale du Colorado est identique à un certain état de mon cerveau, il paraît exagéré d'espérer que quiconque croit que Denver est la capitale du Colorado doive avoir une trame neurophysiologique identique dans son cerveau (Block et Fodor, 1972 ; Putnam, 1967). Et d'une espèce à l'autre, même s'il est vrai que chez tous les humains les douleurs sont identiques aux événements neurophysiologiques humains, nous ne voulons pas exclure la possibilité que dans une autre espèce il puisse y avoir des douleurs qui soient identiques à un autre type de trame neurophysiologique. Bref, il paraît exagéré d'espérer que tout *type* d'état mental soit identique à un *type* d'état neurophysiologique. Et, en vérité, cela semble relever d'une sorte de « chauvinisme neuronal » (Block, 1978) que de supposer que seules des entités dotées de neurones comme les nôtres puissent avoir des états mentaux.

Une troisième objection « technique » à la théorie de l'identité provient de la loi de Leibniz. Si deux événements ne sont identiques que s'ils ont toutes leurs propriétés en commun, alors les états mentaux ne peuvent semble-t-il être identiques à des états physiques, parce que les états mentaux ont certaines propriétés que n'ont pas les états physiques (Smart, 1965 ; Shaffer, 1961). Par exemple, ma douleur est dans mon orteil, mais mon état neurophysiologique correspondant fait tout le chemin de l'orteil au thalamus et au-delà. Où donc se trouve vraiment la douleur ? Les théoriciens de l'identité n'ont guère vu de difficulté à cette objection. Ils ont fait remarquer que l'unité de l'analyse est réellement l'*expérience* consistant à éprouver de la douleur, et que cette expérience (ainsi que l'expérience de l'image corporelle globale) se produit vraisemblablement dans le système nerveux central (Smart, 1965). Sur ce point il me semble que les matérialistes ont absolument raison.

Une objection technique plus radicale à la théorie de l'identité a été soulevée par Saul Kripke (1971), qui s'appuyait sur l'argumentation modale suivante : s'il était réellement vrai que la douleur soit identique à la stimulation des fibres-C, il devrait alors s'agir d'une vérité nécessaire, comme l'est l'énoncé d'identité : « La chaleur est identique au mouvement des molécules. » Et ce, parce que, dans les deux cas, les expressions situées de chaque côté de l'énoncé d'identité sont des « désignateurs rigides ». Par quoi il veut dire que chaque expression identifie l'objet auquel elle fait référence d'après ses propriétés essentielles. Ce sentiment de douleur que j'ai en ce moment est *essentiellement* un sentiment de douleur parce que tout ce qui est identique à ce sentiment devrait être une douleur, et cet état cérébral est *essentiellement* un état cérébral parce que tout ce qui lui est identique devrait

être un état cérébral. Apparemment donc, le théoricien de l'identité qui affirme que les douleurs sont certains types d'états cérébraux, et que cette douleur particulière est identique à cet état cérébral particulier, serait tenu de soutenir les deux choses suivantes : c'est une vérité nécessaire qu'en général les douleurs soient des états cérébraux, et c'est une vérité nécessaire que cette douleur particulière soit un état cérébral. Or, ni l'un ni l'autre ne semble correct. Il ne semble correct de dire ni que les douleurs en général sont nécessairement des états cérébraux ni que ma douleur présente est nécessairement un état cérébral, car il est apparemment facile d'imaginer un être d'un genre quelconque qui pourrait avoir des états cérébraux comme ceux-ci sans avoir de douleurs, et des douleurs comme celles-ci sans être sans des états cérébraux de ce genre. On peut même concevoir une situation dans laquelle j'aurais cette douleur précise sans avoir cet état cérébral précis, et dans laquelle j'aurais cet état cérébral précis sans avoir de douleur.

Cela fait des années que l'on discute de la force de cette argumentation modale et les discussions vont toujours bon train (Lycan, 1971, 1987 ; Sher, 1977). Si l'on se place du point de vue qui nous intéresse ici, je voudrais attirer l'attention sur le fait qu'il s'agit essentiellement de l'objection de sens commun sous une forme sophistiquée. L'objection de sens commun à toute théorie de l'identité est que vous ne pouvez identifier du mental à du non-mental, sans laisser de côté le mental. L'argumentation modale de Kripke est que l'identification des états mentaux à des états cérébraux devrait être nécessaire ; or, elle ne peut l'être, puisque le mental ne saurait être nécessairement physique. Comme le dit Kripke, citant Butler, « tout est ce qu'il est et pas autre chose [6] ».

Quoi qu'il en soit, il paraissait vraiment exagéré de soutenir que tout type d'état mental est identique à un type quelconque d'état neurophysiologique. Mais il semblait possible de préserver la motivation philosophique sous-jacente au matérialisme en défendant une thèse beaucoup plus faible : à savoir que pour toute occurrence particulière (*token*) d'état mental, il y aura une occurrence événementielle neurophysiologique particulière à laquelle cette occurrence particulière est identique. On a parlé au sujet de ces conceptions de « théories de l'identité de token à token » (*token-token identity theories*), lesquelles ne tardèrent pas à remplacer les théories de l'identité de type à type. Certains auteurs eurent en effet l'impression qu'une théorie de l'identité de token à token pouvait échapper à la force des arguments modaux de Kripke [7].

Les théories de l'identité de token à token

Les théoriciens de l'identité de token à token héritèrent de l'objection de sens commun aux théories de l'identité de type à type, c'est-à-dire de l'objection selon laquelle ils donnaient toujours l'impression d'en rester à une forme ou une autre de dualisme des propriétés ; mais ils rencontrèrent d'autres difficultés bien à eux.

Voici l'une d'entre elles. Si deux personnes qui sont dans le même état mental sont dans des états neurophysiologiques différents, alors qu'est-ce qui de ceux-ci fait le même état mental ? Si vous et moi croyons que Denver est la capitale du Colorado, alors qu'avons-nous en commun qui fait que nos différentes trames neurophysiologiques constituent la même croyance ? Notons que les théoriciens de l'identité de token à token ne peuvent donner la réponse de sens commun à cette question ; ils ne peuvent pas dire que deux événements neurophysiologiques sont le même type d'événement mental du fait qu'ils ont le même type de caractéristiques mentales, puisque c'était précisément l'élimination ou la réduction de ces caractéristiques mentales

que cherchait à effectuer le matérialisme. Ils doivent trouver une réponse non mentaliste à la question : « Qu'est-ce qui, dans deux états neurophysiologiques différents, fait qu'il s'agit de tokens du même type d'état mental ? » Étant donné toute la tradition dans laquelle ils travaillaient, la seule réponse plausible à donner ne pouvait se faire que dans un style behavioriste. Aussi répondaient-ils ceci : un état neurophysiologique est un état mental particulier en vertu de sa fonction, ce qui conduit naturellement à la conception suivante.

Le fonctionnalisme de la boîte noire

Ce qui fait de deux états neurophysiologiques des tokens du même type d'état mental, c'est qu'ils accomplissent la même fonction dans la vie globale de l'organisme. La notion de fonction est relativement vague, mais les théoriciens de l'identité de token à token l'ont élaborée de la manière suivante. Deux tokens différents d'état cérébral sont des tokens du même type d'état mental si et seulement si les deux états cérébraux ont les mêmes relations causales avec le stimulus d'entrée que reçoit l'organisme, avec ses divers autres états « mentaux » et avec son comportement de sortie (Lewis, 1972 ; Grice, 1975). Ainsi, par exemple, ma croyance qu'il est sur le point de pleuvoir sera un état en moi qui est causé par ma perception de l'accumulation de nuages et de l'intensification des coups de tonnerre ; ce qui à son tour, allant de pair avec mon désir de ne pas voir la pluie entrer par les fenêtres, me conduira à les fermer. On notera qu'en identifiant les états mentaux en termes des relations causales qui sont les leurs – pas seulement avec des stimuli d'entrée et un comportement de sortie, mais aussi avec d'autres états mentaux –, les théoriciens de l'identité de token à token ont immédiatement évité deux objections que l'on adressait au

behaviorisme. La première était que le behaviorisme avait négligé les relations causales des états mentaux, et la seconde qu'il y avait une circularité dans le behaviorisme, puisque les croyances devaient être analysées en termes de désirs, et les désirs en termes de croyances. Le théoricien de l'identité de token à token, d'obédience fonctionnaliste, n'a aucun mal à accepter cette circularité : il lui suffit de soutenir que l'on peut monnayer l'ensemble du système conceptuel dans les termes d'un système de relations causales.

Le fonctionnalisme disposait d'un magnifique procédé technique qui lui permettait de rendre parfaitement clair ce système de relations sans invoquer les moindres « entités mentales mystérieuses ». Ce procédé est ce que l'on appelle un énoncé de Ramsey[8], et il fonctionne comme suit : supposons que Jean ait la croyance que p, et que ceci soit causé par sa perception que p ; et que, conjointement à son désir que q, la croyance que p cause son action a. Comme nous sommes en train de définir les croyances en termes de leurs relations causales, nous pouvons éliminer l'usage explicite du mot « croyance » dans la phrase précédente, et dire simplement qu'il y a *quelque chose* qui se trouve dans telles ou telles relations causales. En termes formels, la manière dont nous éliminons la mention explicite de la croyance consiste simplement à mettre une variable « x » à la place de n'importe quelle expression faisant référence à la croyance de Jean que p ; et nous apposons à la phrase entière un quantificateur existentiel (Lewis, 1972). Toute l'histoire concernant la croyance de Jean que p peut alors se raconter comme suit :

> $(\exists x)$ (Jean a x & x est causé par la perception que p & x conjointement à un désir que q cause l'action a).

En outre les énoncés de Ramsey sont censés se débarrasser de l'occurrence des derniers termes psychologiques tels que « désir » et « perception ». Une fois que les énoncés de Ramsey

sont épelés de cette manière, on se rend compte que le fonctionnalisme a l'avantage crucial de montrer qu'il n'y a rien de particulièrement mental dans les états mentaux. Parler d'états mentaux, c'est simplement parler d'un ensemble neutre de relations causales ; et l'apparent « chauvinisme » des théories de l'identité de type à type – c'est-à-dire le chauvinisme consistant à supposer que seuls des systèmes dotés de cerveaux comme les nôtres peuvent avoir des états mentaux – est désormais évité grâce à cette conception bien plus « libérale [9] ». Tout système quel qu'il soit, quelle que soit sa composition, pourrait avoir des états mentaux à la seule condition d'avoir les bonnes relations entre ses entrées, son fonctionnement interne et ses sorties. Le fonctionnalisme de cette espèce ne dit rien sur la manière dont opère la croyance pour avoir les relations causales qu'elle a. Il traite simplement l'esprit comme une sorte de boîte noire dans laquelle se présentent ces diverses relations causales, et c'est pour cette raison qu'on l'a parfois intitulé « fonctionnalisme de la boîte noire ».

Les objections au fonctionnalisme de la boîte noire ont révélé le même mélange de sens commun et de technicité que celui que nous avons constaté plus haut. L'objection de sens commun était celle-ci : le fonctionnaliste semble laisser de côté le sentir subjectif qualitatif de certains au moins de nos états mentaux. Il y a certaines expériences qualitatives tout à fait spécifiques qui interviennent dans la vision d'un objet rouge ou dans le fait d'éprouver une douleur lombaire, et se contenter de décrire ces expériences en termes de leurs relations causales c'est laisser de côté ces *qualia* bien particuliers. Ce que l'on démontrait comme suit : supposons qu'une fraction de la population ait son spectre de couleurs inversé de telle manière que, par exemple, l'expérience que les gens appellent « voir rouge » une personne normale l'appellerait, elle, « voir vert » ; et que ce qu'ils appelleraient « voir vert » une personne normale l'appellerait, elle, « voir rouge » (Block et Fodor, 1972). On pourrait fort bien supposer

que cette « inversion du spectre » soit entièrement indétectable par le moindre test usuel de daltonisme, puisque le groupe anormal fait exactement les mêmes discriminations de couleur en réponse aux mêmes stimuli exactement que le reste de la population. Lorsqu'on leur demande de mettre les crayons rouges sur une pile et les verts sur une autre, ils font exactement ce que nous ferions tous ; cela leur *paraît différent* de l'intérieur, mais il n'y a aucun moyen de détecter cette différence de l'extérieur.

Or, si nous pouvons seulement comprendre cette possibilité – et nous le pouvons sûrement – alors le fonctionnalisme de la boîte noire doit se tromper lorsqu'il suppose que des relations causales spécifiées de façon neutre suffisent à rendre compte des phénomènes mentaux ; car de telles spécifications laissent de côté une caractéristique cruciale de bien des phénomènes mentaux, à savoir leur sentir qualitatif.

Une objection voisine était qu'une population énorme, comme celle de toute la Chine, pourrait se comporter de manière à imiter l'organisation fonctionnelle d'un cerveau humain au point d'avoir les bonnes relations d'entrée et de sortie et le bon schème de relations de cause à effet. Pour autant, le système ne sentirait toujours rien en tant que système. La population entière de la Chine ne sentirait pas de douleur en se contentant d'imiter l'organisation fonctionnelle appropriée à la douleur (Block, 1978).

Une autre objection d'allure plus technique au fonctionnalisme de la boîte noire concernait la partie « boîte noire » : le fonctionnalisme ainsi défini ne réussissait pas à énoncer en termes matériels ce qui dans les différents états physiques donne à différents phénomènes matériels les mêmes relations causales. Comment se fait-il que ces structures physiques tout à fait différentes soient causalement équivalentes ?

L'intelligence artificielle forte

C'est alors que s'est produit l'un des développements les plus passionnants de toute l'histoire du matérialisme depuis deux mille ans. La science en plein essor de l'intelligence artificielle apporta une réponse à cette question : différentes structures matérielles peuvent être mentalement équivalentes s'il s'agit de réalisations matérielles différentes du même programme d'ordinateur. En vérité, selon cette réponse, on peut voir que l'esprit n'est qu'un programme d'ordinateur et que le cerveau n'est que l'un parmi tant d'autres des différents « matériels » – qu'ils soient « durs » (*hardware*) ou « humides » (*wetware*) – d'un ordinateur capable d'avoir un esprit. L'esprit est au cerveau ce que le programme est au matériel (Johnson-Laird, 1988). L'intelligence artificielle et le fonctionnalisme firent cause commune, et l'un des aspects les plus stupéfiants de cette alliance fut de révéler que l'on pouvait adopter un matérialisme intégral sur l'esprit et continuer à croire, avec Descartes, que le cerveau n'a pas vraiment d'importance pour l'esprit. L'esprit étant un programme d'ordinateur, et ce programme pouvant se réaliser sur n'importe quel type de matériel (à la seule condition que le matériel soit assez puissant et stable pour effectuer les étapes du programme), on peut spécifier, étudier et comprendre les caractéristiques spécifiquement mentales de l'esprit sans rien connaître du fonctionnement du cerveau. Même si vous êtes matérialiste, vous n'avez pas besoin d'étudier le cerveau pour étudier l'esprit.

Cette idée a donné naissance à cette nouvelle discipline qu'est la « science cognitive ». J'y reviendrai un peu plus loin (aux chapitres VII, IX et X) ; je me contente ici de retracer l'histoire récente du matérialisme. La discipline de l'intelligence artificielle et la théorie philosophique du fonctionnalisme se rejoignirent toutes deux sur l'idée que l'esprit n'est qu'un programme d'ordinateur. J'ai baptisé cette conception l'« intelligence artificielle forte » (Searle, 1980 a), et on l'a aussi appelée le « fonctionnalisme de l'ordinateur » (Dennett, 1978).

Les objections faites à l'IA forte me paraissent révéler le même mélange d'objections de sens commun et d'objections plus ou moins techniques que nous avons pu constater dans les autres cas. Les difficultés et objections techniques à l'intelligence artificielle, que ce soit sous sa version forte ou sous sa version faible, sont nombreuses et complexes. Je n'essaierai pas de les résumer. En général, elles concernent toutes certaines difficultés que l'on a à programmer les ordinateurs de manière à leur permettre de satisfaire au test de Turing. Dans le camp même des partisans de l'IA, il y a toujours eu des difficultés telles que le « problème du cadre » ou l'impossibilité d'obtenir des analyses adéquates du « raisonnement non monotone » qui refléteraient le véritable comportement humain. À l'extérieur du camp de l'IA, il y a eu des objections comme celles de Hubert Dreyfus (1972) déclarant que le fonctionnement de l'esprit humain est tout à fait différent de celui de l'ordinateur.

L'objection de sens commun à l'IA forte était simplement que le modèle computationnel de l'esprit laissait de côté des choses cruciales concernant l'esprit telles que la conscience et l'intentionnalité. Je crois que l'argument le plus célèbre adressé à l'encontre de l'IA forte fut mon argument de la chambre chinoise (Searle, 1980 a) : il s'agissait de montrer qu'un système peut exemplifier un programme fournissant une simulation parfaite d'une capacité cognitive humaine quelconque, telle que la capacité de comprendre le chinois, même si ce système n'a aucune espèce de compréhension du chinois. Imaginons simplement que quelqu'un qui ne comprend pas du tout le chinois se trouve enfermé dans une chambre contenant plein de symboles chinois et un pro-

gramme d'ordinateur pour répondre aux questions en chinois. L'entrée au système est constituée par des symboles chinois sous la forme de questions ; la sortie du système, par des symboles chinois en réponse aux questions. On pourrait supposer que le programme est si bon qu'il soit impossible de distinguer les réponses qui sont faites aux questions de celles que pourrait faire un locuteur dont le chinois est la langue maternelle. Pour autant, ni la personne qui se trouve à l'intérieur ni quelque autre partie que ce soit du système ne comprend littéralement le chinois ; et comme l'ordinateur programmé ne possède rien que ne possède ce système, l'ordinateur programmé, en tant qu'ordinateur, ne comprend pas non plus le chinois. Comme le programme est purement formel ou syntaxique, et comme les esprits ont des contenus mentaux ou sémantiques, toute tentative visant à produire un esprit à l'aide uniquement de programmes d'ordinateur laisse de côté les caractéristiques essentielles de l'esprit.

Outre le behaviorisme, les théories de l'identité de type à type, les théories de l'identité de token à token, le fonctionnalisme et l'IA forte, il y a eu d'autres théories en philosophie de l'esprit dans le cadre général de la tradition matérialiste. L'une d'entre elles, qui remonte au début des années 60 et aux travaux de Paul Feyerabend (1963) et Richard Rorty (1965), a été récemment remise à l'honneur sous différentes formes par des auteurs tels que P. M. Churchland (1981) et S. Stich (1983). On y soutient que les états mentaux n'ont absolument aucune existence. C'est la thèse dite du « matérialisme éliminationniste », que je vais à présent évoquer.

Le matérialisme éliminationniste

Sous sa version la plus sophistiquée, le matérialisme éliminationniste défend l'argumentation suivante : nos croyances

de sens commun concernant l'esprit constituent une sorte de théorie primitive, une « psychologie populaire ». Mais, comme c'est le cas pour n'importe quelle théorie, les entités postulées par la théorie ne peuvent se justifier que dans la mesure où la théorie est vraie. Tout comme l'échec de la théorie phlogistique de la combustion a ôté toute justification à la croyance en l'existence du phlogistique, l'échec de la psychologie populaire enlève toute raison d'être aux entités psychologiques populaires. Partant, s'il se révélait que la psychologie populaire est fausse, alors nous ne serions absolument pas justifiés à croire en l'existence de croyances, de désirs, d'espoirs, de craintes, etc. Selon les matérialistes éliminationnistes, il semble très probable que la psychologie populaire se révélera fausse. Il semble probable qu'une « science cognitive mûre » montrera que la plupart de nos croyances de sens commun sur les états mentaux sont dénuées de tout fondement. Ce résultat aurait pour conséquence que les entités dont nous avons toujours supposé l'existence, nos entités mentales ordinaires, n'existent pas vraiment. Nous voici donc enfin en présence d'une théorie de l'esprit qui élimine purement et simplement l'esprit. D'où l'expression de « matérialisme éliminationniste ».

Un argument voisin utilisé en faveur du « matérialisme éliminationniste » me paraît être si épouvantablement mauvais que je crains de mal le comprendre. Pour autant que je le peux, voici comment il se présente :

> Imaginons que nous ayons une science neurobiologique parfaite. Imaginons que nous ayons une théorie qui ait réellement expliqué le fonctionnement du cerveau. Une telle théorie couvrirait le même domaine que la psychologie populaire, mais serait bien plus puissante. En outre, il paraît fort peu probable que nos concepts communs de psychologie populaire, tels que la croyance et le désir, l'espoir, la crainte, la dépression, l'allégresse, la douleur, etc., correspondent de près ou de loin à la taxinomie fournie par la neurobiologie parfaite que nous avons imaginée. Selon toute probabilité, il n'y aurait aucune

place dans cette neurobiologie pour des expressions telles que « croyance », « crainte », « espoir » et « désir », et aucune réduction aisée de ces phénomènes supposés ne serait possible.

Telle est la prémisse. Voici la conclusion :

Par conséquent, les entités prétendument nommées par les expressions de la psychologie populaire, croyances, espoirs, craintes, désirs, etc., n'existent pas vraiment.

Pour voir combien cet argument est mauvais, imaginons simplement un argument parallèle tiré de la physique :

Considérons la physique théorique dans son état présent. Ici, nous disposons d'une théorie qui explique le fonctionnement de la réalité physique, et qui est de très loin supérieure à nos théories de sens commun selon tous les critères usuels. La théorie physique couvre le même domaine que les théories de sens commun que nous avons des clubs de golf, des raquettes de tennis, des Renault-Espace et des ranchs à deux niveaux. En outre, nos concepts populaires ordinaires tels que « club de golf », « raquette de tennis », « Renault-Espace » et « ranch à deux niveaux » ne correspondent ni de près ni même de loin à la taxinomie de la physique théorique. Il n'y a tout simplement aucun usage en physique théorique pour l'une quelconque de ces expressions et aucune réduction aisée de ces phénomènes en termes de type n'est possible. La manière dont une physique idéale – en vérité la manière dont notre physique véritable – fait la taxinomie de la réalité est vraiment complètement différente de la manière dont notre physique populaire la fait.

Partant, les ranchs à deux niveaux, les raquettes de tennis, les clubs de golf, les Renault-Espace, etc., n'existent pas vraiment.

Je n'ai pas vu de discussion de cette erreur dans les écrits consacrés à ces sujets. Peut-être est-elle si énorme qu'on l'a simplement ignorée. Elle repose sur une prémisse manifestement fausse : pour toute théorie empirique et pour toute taxinomie correspondante, à moins d'avoir une réduction des entités dont on a fait la taxinomie en termes de type aux entités des théories meilleures qu'offre la science fondamentale, les entités n'existent pas. Si vous avez des doutes sur la fausseté de cette prémisse, essayez-la simplement sur quelque chose que vous voyez autour de vous – ou sur vous-mêmes [10] !

Avec le matérialisme éliminationniste, une fois encore, nous trouvons le même schème d'objections techniques et de sens commun que nous avons noté auparavant. Les objections techniques ont à voir avec le fait que la psychologie ordinaire, si c'est une théorie, n'est cependant pas un projet de recherche. Ce n'est pas en soi un domaine de recherche scientifique rival, et, en vérité, les matérialistes éliminationnistes qui attaquent la psychologie populaire, aux dires de leurs critiques, sont souvent injustes. Selon ses défenseurs, la psychologie populaire n'est somme toute pas une si mauvaise théorie ; il y a des chances pour que bon nombre de ses principes centraux se révèlent être vrais. L'objection de sens commun au matérialisme éliminationniste est simplement qu'il a l'air idiot. Il paraît idiot de dire que je n'ai jamais éprouvé de soif ou de désir, que je n'ai jamais eu de douleur, ou que je n'ai jamais vraiment eu de croyance, ou que mes croyances et désirs ne jouent aucun rôle dans mon comportement. Le matérialisme éliminationniste se distingue des théories matérialistes plus anciennes moins par son abandon de l'esprit que par sa négation, dès le départ, de l'existence de quoi que ce soit qui pourrait être laissé de côté. Lorsqu'ils ont à faire face à l'accusation selon laquelle le matérialisme éliminationniste paraît trop insensé pour mériter d'être pris au sérieux, ses défenseurs invoquent presque invariablement la manœuvre de l'âge héroïque de la science (P. S. Churchland, 1987). Selon eux, en un mot, cesser de croire que nous avons des croyances, c'est la même chose que cesser de croire que la terre est plate ou qu'il y a des couchers de soleil, par exemple.

Notons bien dans toute cette discussion qu'une certaine asymétrie paradoxale est survenue dans l'histoire du matérialisme. Les premières théories de l'identité de type à type soutenaient que nous pouvons nous débarrasser des mystérieux

états mentaux cartésiens parce que ces états ne sont *rien d'autre que* des états physiques (rien « de plus et d'autre que » des états physiques) ; et elles soutenaient cela sur la base du fait que l'on peut montrer que les types d'états mentaux sont identiques à des types d'états physiques, que l'on peut parvenir à une correspondance entre les résultats de la neurobiologie et les notions ordinaires que nous pouvons avoir de la douleur ou de la croyance. Or, dans le cas du matérialisme éliminationniste, c'est précisément l'échec supposé de n'importe quelle correspondance de ce genre qui est considéré comme la justification de l'élimination de ces états mentaux en faveur d'une neurobiologie intégrale. Les premiers matérialistes soutenaient qu'il n'y a pas de phénomènes mentaux distincts, parce que les phénomènes mentaux *sont identiques* à des états cérébraux. Les matérialistes plus récents soutiennent qu'il n'y a pas de phénomènes mentaux distincts parce qu'ils *ne sont pas identiques* à des états cérébraux. Ce schéma me semble très révélateur du besoin de se débarrasser des phénomènes mentaux à n'importe quel prix.

Naturaliser le contenu

Après un demi-siècle de ce schéma récurrent dans les débats sur le matérialisme, on aurait pu penser que les matérialistes et les dualistes se seraient aperçus que le problème résidait dans les termes du débat. Or, à ce jour, l'induction ne semble s'être produite ni d'un côté ni de l'autre. Au moment où j'écris ces lignes, on trouve la répétition du même schéma dans les tentatives actuelles de « naturalisation » du contenu intentionnel.

Stratégiquement, on s'emploie à détacher le problème de la conscience du problème de l'intentionnalité. Il se peut, admet-on, que la conscience soit irréductiblement mentale, et ne puisse donc se prêter à un traitement scientifique, mais

il se peut également, après tout, que la conscience soit moins importante qu'on ne le pense, et qu'on puisse s'en passer. Il nous faut seulement naturaliser l'intentionnalité, par quoi l'on entend : l'expliquer complètement en termes de (la réduire à des) phénomènes physiques, non mentaux. Le fonctionnalisme était une tentative de naturalisation du contenu intentionnel de ce genre, et il a été rajeuni par son alliance avec des théories causales externalistes de la référence. L'idée sous-jacente à ces conceptions est que le contenu sémantique, c'est-à-dire les significations, ne peut se trouver entièrement dans nos têtes ; en effet ce qui s'y trouve ne suffit pas pour déterminer comment le langage se rapporte à la réalité. En plus de ce qui se trouve dans nos têtes, le « contenu étroit », il nous faut un ensemble de relations causales physiques réelles avec les objets du monde, il nous faut un « contenu large ». Ces conceptions se sont à l'origine développées autour de problèmes de philosophie du langage (Putnam, 1975 b), mais il est facile de voir comment elles se généralisent aux contenus mentaux. Si la signification de la phrase « l'eau est humide » ne peut s'expliquer par ce qui se trouve à l'intérieur de la tête de locuteurs s'exprimant en français, alors la croyance que l'eau est humide ne relève pas, elle aussi, uniquement de ce qui se trouve dans leurs têtes. Dans l'idéal, on aimerait avoir une analyse du contenu intentionnel qui se fasse uniquement en termes de relations causales entre les gens, d'une part, et en termes d'objets et d'états de choses du monde, d'autre part.

Une analyse rivale, et, je crois, encore moins plausible de la tentative causale externaliste de naturalisation du contenu, est celle qui consiste à dire que les contenus intentionnels peuvent être individualisés par leur fonction téléologique, biologique darwinienne. Par exemple, mes désirs auront un contenu faisant référence à de l'eau ou à de la nourriture si et seulement s'ils fonctionnent pour m'aider à obtenir de l'eau ou de la nourriture (Millikan, 1984).

Pour l'heure, aucune tentative de naturalisation du contenu n'a produit d'explication (d'analyse, de réduction) du contenu intentionnel qui soit ne serait-ce que lointainement plausible. Considérons un type de croyance très simple : je crois, par exemple, que Flaubert était un meilleur romancier que Balzac. À quoi ressemblerait donc une analyse de ce contenu, énoncée en termes de causalité physique brute ou de sélection naturelle darwinienne, sans employer de termes mentaux ? Personne ne sera surpris d'apprendre que toutes ces tentatives avortent dès le départ.

Une fois encore de telles conceptions naturalisées du contenu se heurtent à des objections techniques comme à des objections de sens commun. Le plus célèbre des problèmes techniques est probablement celui de la disjonction (Fodor, 1987). Si un certain concept est causé par une certaine sorte d'objet, alors comment expliquons-nous les cas d'erreur sur l'identité ? Si « cheval » est causé par des chevaux ou par des vaches qui sont à tort identifiées comme étant des chevaux, nous faut-il alors dire que l'analyse de « cheval » est disjonctive, qu'elle veut dire soit cheval, soit certaines sortes de vaches ?

Au moment où j'écris ces lignes, des analyses naturalistes (externalistes, causales) du contenu font rage. Elles échoueront toutes pour des raisons qui, je l'espère, sont désormais évidentes. Elles laisseront de côté la subjectivité du contenu mental. Par le biais d'objections techniques on invoquera des contre-exemples, tels que les cas de disjonction, et l'on répondra aux contre-exemples par tel ou tel gadget – les relations nomologiques, et les contrefactuels, ou du moins c'est ce que je prédis – mais le plus qu'on puisse espérer de gadgets, même s'ils permettaient de parer aux contre-exemples, ce serait un parallélisme entre ce qui ressort du gadget et les intuitions que l'on peut avoir sur le contenu mental. On ne parviendrait toujours pas jusqu'à l'essence du contenu mental.

J'ignore si quelqu'un a déjà fait l'évidente objection de

sens commun au projet de naturalisation du contenu intentionnel, mais on voit bien, je l'espère, à partir de toute cette discussion en quoi elle consistera. Au cas où personne ne l'a jamais faite, la voici : toute tentative de réduction de l'intentionnalité à quelque chose de non mental échouera toujours parce qu'elle laisse de côté l'intentionnalité. Supposons, par exemple, que vous ayez une explication externaliste causale parfaite de la croyance que l'eau est humide. Cette analyse est fournie par l'établissement d'un ensemble de relations causales qu'entretient un système avec l'eau et l'humidité et ces relations sont entièrement spécifiées sans la moindre composante mentale. Le problème est évident : un système pourrait avoir chacune de ces relations et continuer à ne pas croire que l'eau est humide. Ce n'est qu'une extension de l'argument de la chambre chinoise, mais la morale qu'elle suggère est générale : vous ne pouvez réduire le contenu intentionnel (ou les douleurs ou les « *qualia* ») à quelque chose d'autre, parce que si vous le pouviez elles seraient quelque chose d'autre ; or, elles ne le sont pas. La position opposée à la mienne est celle qui est formulée par Fodor : « Si l'à-propos-de (*aboutness*) est réel, ce doit être vraiment quelque chose d'autre » (1987, p. 97). Tout au contraire : l'à-propos-de (i.e. l'intentionnalité) est réel, et ce n'est pas quelque chose d'autre.

Un bon indice du fait que le projet est radicalement vicié est que les notions intentionnelles sont fondamentalement normatives. Elles établissent des normes de vérité, de rationalité, de cohérence logique, etc., et, en aucune façon, ces normes ne peuvent être intrinsèques à un système entièrement constitué par des relations causales, brutes, aveugles et non intentionnelles. Il n'y a pas de composante normative dans l'action causale des boules de billard. Les tentatives biologiques darwiniennes de naturalisation du contenu essaient d'éviter ce problème en faisant appel à ce qu'elles supposent être le caractère normatif, fondamentalement téléologique de

l'évolution biologique. Mais c'est là une très grave erreur. Il n'y a rien de normatif ou de téléologique dans l'évolution darwinienne. En vérité, la contribution majeure de Darwin a été précisément de retrancher le but et la téléologie de l'évolution. L'analyse de Darwin montre que la téléologie apparente des processus biologiques est une illusion.

On ne fait que généraliser cette idée lorsqu'on souligne que des notions telles que celle de « but » ne sont jamais intrinsèques à des organismes biologiques (à moins bien sûr que ces organismes eux-mêmes n'aient des états et des processus intentionnels conscients). Et même des notions comme celle de « fonction biologique » se sont toujours construites relativement à un observateur qui attribue une valeur normative aux processus causaux. Il n'y a aucune différence *factuelle* à propos du cœur qui corresponde à celle qu'il y a entre le fait de dire :

1. Le cœur cause le pompage du sang.

et de dire :

2. La fonction du cœur est de pomper le sang.

Mais 2 attribue un statut normatif aux simples faits causaux bruts sur le cœur, et ce, en raison de notre intérêt pour la relation de ce fait à toute une série d'autres faits, tels que notre intérêt pour la survie. En bref, les mécanismes darwiniens, pour ne rien dire des fonctions biologiques elles-mêmes, sont entièrement dénués de but ou de téléologie. Chacune des caractéristiques téléologiques se trouve entièrement dans l'esprit de l'observateur [11].

Quelle morale en tirer pour le moment ?

J'ai visé jusqu'à présent dans ce chapitre à y illustrer un schéma récurrent dans l'histoire du matérialisme. Ce schéma est présenté de manière graphique dans le tableau 2.1. Je ne me suis pas tant soucié de défendre ou de réfuter le

matérialisme que d'examiner les vicissitudes qu'il a pu connaître en se trouvant confronté à certains faits de sens commun sur l'esprit, tel celui que la plupart d'entre nous sommes, pendant la majeure partie de notre vie, conscients. Ce que nous trouvons dans l'histoire du matérialisme, c'est une tension récurrente entre le besoin de donner une analyse de la réalité qui laisse de côté toute référence aux caractéristiques spécifiques du mental, telles que la conscience et la subjectivité, tout en rendant compte de nos « intuitions » sur l'esprit. Il est naturellement impossible de faire les deux à la fois. Aussi assiste-t-on à une série de tentatives, qui prennent un tour presque névrotique, pour dissimuler le fait que l'on a laissé de côté tel ou tel élément crucial sur les états mentaux. Et lorsque l'on fait remarquer que telle vérité évidente se voit niée par la philosophie matérialiste, les défenseurs de cette position ont presque invariablement recours à certaines stratégies rhétoriques censées montrer que le matérialisme doit avoir raison, et que le philosophe qui s'oppose au matérialisme doit souscrire à telle ou telle version du dualisme, du mysticisme, du mystérieux, ou à tel ou tel préjugé antiscientifique général. Mais la motivation inconsciente de tout cela, celle qui n'apparaît jamais en pleine lumière, est le postulat de l'incompatibilité nécessaire du matérialisme avec la réalité et l'efficacité causale de la conscience, de la subjectivité, etc. En d'autres termes, le postulat de base sous-jacent au matérialisme est le postulat essentiellement cartésien selon lequel le matérialisme implique l'antimentalisme et que le mentalisme implique l'antimatérialisme.

Il y a quelque chose d'extrêmement déprimant dans toute cette histoire : tout y paraît si inutile, si vain. Tout repose sur le postulat erroné que la conception de la réalité comme étant entièrement physique n'est pas compatible avec la conception selon laquelle le monde contient réellement des états conscients subjectifs (« qualitatifs », « privés », « tou-

chants-sentants », « immatériels », « non physiques ») tels que les pensées et les sentiments.

L'étrange dans toute cette discussion, c'est que le matérialisme hérite du pire postulat du dualisme. En niant la thèse du dualisme selon laquelle il y a deux sortes de substances dans le monde, ou en niant la thèse du dualisme des propriétés selon laquelle il y a deux sortes de propriétés dans le monde, le matérialisme accepte sans s'en apercevoir les catégories et le vocabulaire du dualisme. Il accepte les termes dans lesquels Descartes a formulé le débat. Il accepte, en un mot, l'idée que le vocabulaire du mental et du physique, du matériel et de l'immatériel, de l'esprit et du corps, est parfaitement adéquat, tel quel. Il accepte l'idée que si nous pensons que la conscience existe, nous acceptons le dualisme. Ce que je crois – comme le montre clairement toute la discussion – c'est que le vocabulaire et les catégories qui l'accompagnent sont la source de nos difficultés philosophiques les plus graves. Tant que nous employons des mots comme celui de « matérialisme », nous sommes presque invariablement contraints de supposer qu'ils impliquent quelque chose d'incompatible avec le mentalisme naïf. J'ai défendu l'idée que dans ce cas, on peut avoir le beurre et l'argent du beurre. On peut être un « matérialiste intégral » sans aucunement nier l'existence de phénomènes mentaux (subjectifs, internes, intrinsèques, souvent conscients). Néanmoins, comme mon usage de ces termes va à l'encontre de près de trois cents ans de tradition philosophique, il vaudrait sans doute mieux abandonner aussi ce vocabulaire.

Si l'on devait décrire la motivation la plus profonde en faveur du matérialisme, on pourrait dire que c'est simplement une terreur vis-à-vis de la conscience. Mais devrait-il en être ainsi ? Pourquoi les matérialistes devraient-ils redouter la conscience ? Pourquoi les matérialistes n'admettent-ils pas de bon cœur la conscience en n'y voyant qu'une propriété matérielle parmi tant d'autres ? Certains, en fait, tels que

TABLEAU 2.1
Le schéma général du matérialisme récent

Théorie	Objections du sens commun	Objections techniques
Behaviorisme logique	Laisse de côté l'esprit : objections du superspartiate/du superacteur	1. Circulaire : a besoin d'expliquer les croyances, et inversement 2. Ne peut pas effectuer les conditionnels 3. Laisse de côté la causalité
Théorie de l'identité de type à type	Laisse de côté l'esprit : ou alors conduit au dualisme des propriétés	1. Chauvinisme 2. Loi de Leibniz 3. Ne peut expliquer les propriétés mentales 4. Arguments modaux
Théorie de l'identité de token à token	Laisse de côté l'esprit : *qualia* absents	Ne peut identifier les traits mentaux du contenu mental
Le fonctionnalisme de la boîte noire	Laisse de côté l'esprit : *qualia* absents et inversion du spectre	La relation entre structure et fonction reste inexpliquée
L'IA forte (le fonctionnalisme de la machine de Turing)	Laisse de côté l'esprit : la chambre chinoise	La cognition humaine est non représentationnelle et donc non computationnelle
Le matérialisme éliminationniste (rejet de la psychologie populaire)	Nie l'existence de l'esprit : injuste envers la psychologie populaire	Défense de la psychologie populaire
La naturalisation de l'intentionnalité	Laisse de côté l'intentionnalité	Le problème de la disjonction

Armstrong et Dennett, prétendent le faire. Mais ils le font en redéfinissant la « conscience » de telle manière qu'ils en nient l'aspect central, à savoir sa qualité subjective. La raison la plus profonde de notre crainte de la conscience, c'est qu'elle possède une propriété fondamentalement terrifiante : la subjectivité. Les matérialistes répugnent à accepter cette propriété parce qu'ils croient qu'accepter l'existence de la conscience subjective contredirait leur conception de ce que doit être le monde. Beaucoup parmi eux pensent qu'au vu des découvertes de la physique, une conception de la réalité qui nie l'existence de la subjectivité est la seule possible. Une fois encore, comme dans le cas de la « conscience », une manière de s'en sortir est de redéfinir la « subjectivité » de telle sorte qu'elle ne signifie plus la subjectivité mais qu'elle signifie quelque chose d'objectif (pour un exemple, voir Lycan, 1990 a).

Je vois dans tout cela une monumentale erreur, et aux chapitres IV, V, et VI, j'examinerai en détail le caractère et le statut ontologique de la conscience.

Les idoles de la tribu

Il me reste à expliquer pourquoi une question qui pouvait paraître naturelle était vraiment incohérente. La question, évoquée au début de ce chapitre, est la suivante : comment des morceaux de matière dépourvus d'intelligence produisent-ils de l'intelligence ? On notera tout d'abord la forme de la question. Pourquoi ne posons-nous pas la question plus traditionnelle : comment des morceaux de matière inconscients produisent-ils de la conscience ? Cette question me paraît parfaitement cohérente. Elle concerne la manière dont le cerveau fonctionne pour causer des états mentaux conscients même si les neurones individuels (ou les synapses ou les récepteurs) du cerveau ne sont pas eux-mêmes conscients.

Mais, à l'époque actuelle, nous éprouvons des réticences à poser la question sous cette forme parce que nous manquons de critères « objectifs » de la conscience. La conscience a une ontologie subjective qu'il est impossible d'éliminer, aussi jugeons-nous plus scientifique de reformuler la question comme une question sur l'intelligence, parce que nous pensons que pour l'intelligence nous disposons de critères objectifs et impersonnels. Mais voici que nous rencontrons immédiatement une difficulté. Si par « intelligence » nous entendons quelque chose qui satisfait aux critères objectifs à la troisième personne de l'intelligence, alors la question contient un présupposé erroné. Parce que si l'intelligence est définie en termes behavioristes, alors en aucun cas les neurones ne sont intelligents. Les neurones, comme toutes les autres choses du monde, se comportent selon certaines trames régulières, et prévisibles. En outre, vus sous un certain angle, les neurones font du « traitement de l'information » extrêmement sophistiqué. Ils recueillent une quantité abondante de signaux d'autres neurones à leurs synapses dendritiques ; ils traitent cette information au niveau de leur corps cellulaire et renvoient l'information vers d'autres neurones à travers leurs synapses axonales. S'il faut définir l'intelligence en termes behavioristes, alors les neurones sont sacrément intelligents, quels que soient les critères que l'on adopte. En un mot, si nos critères d'intelligence sont entièrement objectifs et à la troisième personne – et le fait de poser la question de cette manière n'avait pour but que de parvenir à quelque chose qui satisfasse à ces conditions –, alors la question contient un présupposé qui en lui-même est faux. La question présuppose faussement que les morceaux ne répondent pas aux critères de l'intelligence.

La réponse à la question, cela n'est guère surprenant, hérite de la même ambiguïté. Il y a deux ensembles différents de critères qui permettent d'appliquer l'expression de « comportement intelligent ». L'un de ces ensembles consiste

en critères à la troisième personne ou « objectifs » qui n'ont pas nécessairement d'intérêt psychologique. En revanche, les critères de l'autre groupe sont essentiellement mentaux et impliquent le point de vue à la première personne. Le « comportement intelligent » selon le second groupe de critères implique la pensée, et la pensée est fondamentalement un processus mental. Or, si nous adoptons les critères à la troisième personne pour le comportement intelligent, alors bien sûr les ordinateurs – pour ne rien dire des calculatrices de poche, des voitures, des excavateurs, des thermostats, et, en vérité, à peu près tout dans le monde – font preuve de comportement intelligent. S'il est cohérent d'adopter le test de Turing ou tout autre critère « objectif » du comportement intelligent, alors la réponse à des questions telles que celle de savoir si des morceaux de matière dépourvus d'intelligence peuvent produire un comportement intelligent, et même celle de savoir comment exactement ils s'y prennent, sont d'une ridicule évidence. N'importe quel thermostat, n'importe quelle calculatrice de poche ou n'importe quelle chute d'eau produit un « comportement intelligent », et nous le savons à chaque fois qu'il fonctionne. Certains artefacts sont conçus pour se comporter comme s'ils étaient intelligents, et comme toutes les choses suivent les lois de la nature, elles auront toutes une description sous laquelle elles se comportent comme si elles étaient intelligentes. Mais ce sens de « comportement intelligent » n'a absolument aucune pertinence psychologique.

En un mot, nous avons tendance à entendre et la question et la réponse comme oscillant entre deux pôles différents : (*a*) comment des morceaux de matière dépourvus de conscience produisent-ils de la conscience ? (question parfaitement valable dont la réponse est : en vertu de caractéristiques neurobiologiques spécifiques – bien que largement ignorées – du cerveau) ; et (*b*) comment des morceaux de matière « dénués d'intelligence » (selon les critères à la première ou à la troisième personne ?) produisent-ils du comportement

« intelligent » (selon les critères à la première ou à la troisième personne ?) ? Mais dans la mesure où nous faisons des critères de l'intelligence des critères à la troisième personne, la question repose sur un présupposé erroné, ce que nous ne voyons pas en raison du fait que nous avons tendance à entendre la question au sens de l'interprétation (*a*).

CHAPITRE IV

La conscience et sa place dans la nature

La conscience et la conception « scientifique » du monde

Comme pour la majorité des mots, il n'est pas possible de donner une définition de la « conscience » en termes de conditions nécessaires et suffisantes, pas plus qu'il n'est possible de la définir à la mode aristotélicienne au moyen de genres et de différences. Néanmoins, et bien que nous ne puissions donner une définition verbale qui ne soit pas circulaire, il me paraît toujours capital de définir cette notion, tant nous la confondons souvent avec plusieurs autres. Par exemple, pour des raisons qui tiennent à la fois à l'étymologie et à l'usage, on confond souvent la « conscience » avec la « conscience morale », la « conscience de soi » et la « cognition ».

Des exemples illustreront le mieux ce que j'entends par « conscience ». Lorsque je me réveille au terme d'un sommeil sans rêves, j'entre dans un état de conscience, un état qui se poursuit tant que je suis éveillé. Lorsque je m'endors ou suis placé sous anesthésie générale ou que je meurs, mes états conscients cessent. Si durant mon sommeil j'ai des rêves, je deviens conscient, bien que les formes oniriques de la conscience en général soient de bien moindre intensité et vivacité que la conscience éveillée ordinaire. La conscience

peut varier en degré même durant nos heures de veille, comme par exemple lorsque d'éveillés et alertes que nous sommes, nous devenons endormis ou somnolents, ou simplement ennuyés et inattentifs. Certaines personnes introduisent des substances chimiques dans leur cerveau afin de produire des modifications de leurs états de conscience ; toutefois, même sans assistance chimique, il est possible dans la vie ordinaire de distinguer différents degrés et formes de conscience. La conscience est un interrupteur que l'on allume ou que l'on éteint : un système est conscient ou non. Mais une fois qu'il est conscient, le système est un rhéostat : il y a différents degrés de conscience.

Un proche synonyme de « conscience » (*consciousness*), à mon sens, est la « connaissance immédiate » (*awareness*), mais je ne pense pas qu'elles soient exactement équivalentes en signification parce que la « connaissance immédiate » se rattache de plus près à la cognition, à la connaissance, que la notion générale de conscience. En outre, il paraît possible d'envisager des cas dans lesquels inconsciemment on a une connaissance immédiate de quelque chose (cf. Weiskrantz *et al.*, 1974). On notera qu'il n'y a rien jusqu'à présent dans mon analyse de la conscience qui implique la conscience *de soi*. (Je discuterai au chapitre VI de la connexion qui existe entre conscience et conscience de soi.)

Certains philosophes (comme Block, « Deux concepts de la conscience ») soutiennent qu'il y a un sens du mot qui n'implique aucune espèce de sensitivité, un sens dans lequel on pourrait dire qu'un parfait zombie est « conscient ». Je ne vois pas de quel sens il s'agit, mais, en tout cas, ce n'est pas le sens que je donne à ce mot.

Les états conscients ont toujours un contenu. On ne peut jamais être simplement conscient ; quand on est conscient, c'est plutôt qu'il doit y avoir une réponse à la question de savoir de quoi l'on est conscient. Mais le « de » de « conscient de » n'est pas toujours le « de » de l'intentionnalité. Si je suis

conscient d'un frappement à la porte, mon état conscient est intentionnel parce qu'il fait référence à quelque chose qui est au-delà de lui, le frappement à la porte. Si je suis conscient d'une douleur, la douleur n'est pas intentionnelle, parce qu'elle ne représente rien au-delà d'elle-même [1].

Ce chapitre a pour but principal de localiser la conscience au sein de notre conception « scientifique » globale du monde. La raison pour laquelle on met l'accent sur la conscience dans une analyse de l'esprit est qu'il s'agit de la notion mentale centrale. D'une manière ou d'une autre, toutes les autres notions mentales – telles que l'intentionnalité, la subjectivité, la causalité mentale, l'intelligence, etc. – ne peuvent pleinement se comprendre comme *mentales* que par le biais des relations qu'elles ont avec la conscience (j'y reviendrai au chapitre VII). Comme à n'importe quelle étape de notre vie de veille, seule une minuscule fraction de nos états mentaux est consciente, il peut sembler paradoxal de penser que la conscience soit la notion mentale centrale, mais j'ai l'intention dans ce livre de résoudre ce paradoxe apparent. Une fois que nous avons localisé la place de la conscience dans notre conception globale du monde, nous voyons que les théories matérialistes de l'esprit que nous avons discutées au chapitre II sont tout aussi profondément antiscientifiques que le dualisme qu'elles croyaient attaquer. Nous nous apercevrons que lorsque nous essayons d'énoncer les faits, la pression exercée sur les catégories et la terminologie traditionnelles devient telle que notre vocabulaire commence à s'effondrer sous les coups de boutoir. J'aurai presque l'air de me contredire : d'un côté je vais soutenir que la conscience n'est qu'une caractéristique biologique ordinaire du monde, mais j'essaierai aussi de montrer pourquoi nous trouvons presque littéralement inconcevable qu'il en soit ainsi.

Notre conception actuelle du monde a commencé à se développer au XVII^e siècle, et son développement se poursuit

jusqu'en cette fin du XX^e^. Historiquement, l'une des clés de ce développement fut l'exclusion de la conscience hors du domaine scientifique par Descartes, Galilée, et d'autres au cours du XVII^e^ siècle. Selon la conception cartésienne, les sciences de la nature à proprement parler excluaient l'« esprit », la *res cogitans,* et ne s'occupaient que de la « matière », la *res extensa.* La séparation entre l'esprit et la matière fut un utile instrument heuristique au XVII^e^ siècle, un instrument qui facilita grandement le progrès accompli dans les sciences. Toutefois, la séparation est philosophiquement confuse, et au XX^e^ siècle, elle a fini par devenir un obstacle majeur à la compréhension scientifique de la place de la conscience au sein du monde naturel. J'entends aider à lever cet obstacle ; à réintroduire la conscience dans le domaine scientifique comme étant un phénomène biologique au même titre que n'importe quel autre. Pour ce faire, il nous faut répondre aux objections dualistes des cartésiens contemporains.

Il va sans dire que notre conception « scientifique » du monde est extrêmement complexe et inclut toutes nos théories généralement acceptées sur le genre de lieu qu'est l'univers et sur son fonctionnement. Elle inclut, en d'autres termes, des théories qui vont de la mécanique quantique et la théorie de la relativité à la tectonique des plaques de la géologie et la théorie ADN de la transmission héréditaire. Actuellement, par exemple, elle inclut la croyance aux trous noirs, la théorie microbienne de la maladie et l'explication héliocentrique du système solaire. Certains aspects de cette conception du monde sont encore non décidés, d'autres sont très bien établis. Au moins deux aspects de celle-ci sont si fondamentaux et si bien établis qu'ils ne se présentent plus comme des options possibles pour tout citoyen raisonnablement cultivé de notre époque ; en vérité, elles sont pour une bonne part constitutives de la conception moderne du monde. Il s'agit de la théorie atomique de la matière et de la théorie de l'évolution de la biologie. Bien sûr, comme toute autre théorie, il se pourrait

qu'elles soient réfutées par des recherches ultérieures, mais actuellement les preuves en leur faveur sont si écrasantes qu'elles ne sont pas simplement jetées en pâture au plus offrant. Pour situer la conscience au sein de notre compréhension du monde, il nous faut la situer par rapport à ces deux théories.

Selon la théorie atomique de la matière, l'univers est constitué de part en part de phénomènes physiques extrêmement petits que nous trouvons commodes, bien que ce ne soit pas entièrement exact, d'appeler « particules ». Toutes les entités du monde, de grande et de moyenne taille, telles que planètes, galaxies, voitures, pardessus, sont faites d'entités plus petites, elles-mêmes composées de particules subatomiques. Comme exemples de particules, il y a les électrons, les atomes d'hydrogène et les molécules d'eau. Comme l'illustrent ces exemples, des particules plus grandes sont faites de plus petites, et il y a encore beaucoup d'incertitude et de discussion quant à savoir comment identifier les particules qui sont ultimement les plus petites. Il est relativement embarrassant d'employer le terme de « particule » pour au moins deux raisons. La première est qu'il semble plus exact de décrire les plus fondamentales de ces entités comme des points de masse/énergie plutôt que comme des entités spatiales étendues. Et la seconde, plus radicale, est que, selon la mécanique quantique, tant qu'elles ne sont pas mesurées ou soumises à telle ou telle ingérence, les « particules » telles que les électrons se comportent davantage comme des ondes que comme des particules. Toutefois, par commodité, je m'en tiendrai au terme de « particule ».

Les particules, comme l'illustraient nos précédents exemples, s'organisent en de plus vastes *systèmes.* Il serait délicat d'essayer de définir la notion de système, mais l'idée intuitive simple est que les systèmes sont des collections de particules où les limites spatio-temporelles du système sont fixées par des relations causales. Ainsi, une goutte d'eau est

un système, mais un glacier aussi. Les bébés, les éléphants et les chaînes de montagne sont aussi des exemples de systèmes. Chacun aura compris, à la lumière de ces exemples, que les systèmes peuvent contenir des sous-systèmes.

Le point capital dans le dispositif explicatif de la théorie atomique n'est pas seulement l'idée que les grands systèmes sont faits de petits systèmes, mais que bien des aspects des grands peuvent *s'expliquer causalement* par le comportement des petits. Cette conception de l'explication nous donne la possibilité, en vérité l'obligation, d'expliquer quantité de macrophénomènes par des microphénomènes. D'où la conséquence, à son tour, qu'il y aura différents niveaux d'explication du même phénomène, selon que nous allons de gauche à droite – de macro à macro, de micro à micro –, ou de bas en haut – de micro à macro. Nous pouvons illustrer ces niveaux par un exemple.

Supposons que je souhaite expliquer pourquoi cette casserole d'eau est en train de bouillir. Une explication, une explication de gauche à droite et macro-macro, serait que je place la casserole sur la cuisinière et que j'allume la plaque en dessous. J'appelle cette explication « de gauche à droite » parce qu'elle cite un événement antérieur pour expliquer un événement postérieur [2] et je l'appelle « macro-macro » parce que l'*explanans* comme l'*explanandum* se situent au macroniveau. Une autre explication – de bas en haut et micro-macro – serait que l'eau est en train de bouillir parce que l'énergie cinétique transmise par l'oxydation des hydrocarbones aux molécules H_2O les a fait se mouvoir si rapidement que la pression interne des mouvements de molécules est égale à la pression de l'air extérieur, laquelle s'explique à son tour par le mouvement des molécules dont est composé l'air extérieur. J'appelle cette explication « de bas en haut et micro-macro » parce qu'elle explique les caractéristiques et le comportement de surface ou des macrophénomènes par des microphénomènes de niveau inférieur. Par quoi je ne prétends pas que ce soit là les seuls niveaux

possibles d'explication. Il y a aussi des explications de gauche à droite et micro-micro, et l'on peut faire encore d'autres subdivisions à chaque micro ou macroniveau.

Telle est donc l'une des principales leçons de la théorie atomique : bien des aspects des choses de grosse taille peuvent s'expliquer par le comportement des petites. Nous considérons la théorie microbienne de la maladie ou la théorie ADN de la transmission génétique comme des avancées majeures, précisément parce qu'elles cadrent avec ce modèle. Si quelqu'un donnait une explication des maladies en invoquant le mouvement des planètes, nous n'admettrions jamais qu'il s'agit d'une explication complète, même si elle marchait pour les diagnostics et les guérisons, à moins de comprendre comment des macrocauses et effets au niveau des planètes et des symptômes reposent sur des micro-macro structures causales de bas en haut.

À ces notions élémentaires de la théorie atomique ajoutons à présent les principes de la biologie de l'évolution. Sur de longues périodes de temps, certains *types* d'organismes vivants évoluent selon des modalités très particulières. Sur notre petite terre, les types de systèmes en question contiennent invariablement des molécules à base de carbone, et ils font une utilisation étendue d'hydrogène, de nitrogène et d'oxygène. Ils évoluent de manière compliquée, mais la procédure de base est que des tokens – c'est-à-dire des occurrences de types – font venir à l'existence des tokens similaires. Ainsi, une fois détruits les tokens originels, le type ou le schéma qu'ils exemplifient se perpétue dans d'autres tokens et continue à se reproduire au fur et à mesure que les générations suivantes de tokens en produisent encore d'autres. Des variations dans les caractères de surface, les phénotypes des tokens, donnent aux tokens de plus ou moins grandes chances de survie, en fonction de l'environnement spécifique dans lequel ils se trouvent. Ces tokens qui ont une plus grande probabilité de survie par rapport à leur environnement auront

donc une plus grande probabilité à produire d'autres tokens à leur image, des tokens dotés du même génotype. Et c'est ainsi qu'évolue le type.

Une partie de l'attrait exercé sur les esprits par la théorie de l'évolution, complétée par la génétique de Mendel et celle de l'ADN, est qu'elle s'adapte au modèle explicatif que nous avons dérivé de la théorie atomique. Plus particulièrement, le fondement des mécanismes génétiques en biologie moléculaire autorise différents niveaux d'explication des phénomènes biologiques correspondant aux différents niveaux d'explication que nous avons pour les phénomènes physiques. En biologie de l'évolution, il y a de façon caractéristique deux niveaux d'explication, un niveau « fonctionnel » où nous expliquons la survie de l'espèce en termes de « fitness globale », qui dépend des traits phénotypiques possédés par les membres de l'espèce, et un niveau « causal » où nous expliquons les mécanismes causaux par lesquels les traits en question mettent vraiment en rapport l'organisme et l'environnement. Prenons l'exemple des plantes vertes : pourquoi tournent-elles leurs feuilles vers le soleil ? Explication fonctionnelle [3] : ce trait a une valeur de survie. En augmentant la capacité de la plante à accomplir la photosynthèse, il augmente la capacité de la plante à survivre et à se reproduire. La plante ne se tourne pas vers le soleil pour survivre ; la plante a plutôt tendance à survivre parce qu'elle est prédisposée à se tourner vers le soleil de toute façon. Explication causale : la structure biochimique de la plante telle que la détermine sa constitution génétique cause en elle la sécrétion de l'hormone de croissance, auxine, et des concentrations variées d'auxine ont à leur tour pour effet que les feuilles se tournent en direction de la source de lumière.

Si vous réunissez ces deux niveaux d'explication, vous parvenez au résultat suivant : parce que le phénotype, tel qu'il est produit par l'interaction du génotype et de l'environnement, a une valeur de survie au regard de l'environnement, le génotype survit et se reproduit. Tels sont, sous une forme très brève, les mécanismes de la sélection naturelle.

Produits du processus évolutionnaire, les organismes sont faits de sous-systèmes appelés des « cellules », et certains de ces organismes développent des sous-systèmes de cellules nerveuses, que nous considérons comme des « systèmes nerveux ». En outre, et c'est le point crucial, certains systèmes nerveux extrêmement complexes sont capables de causer et de maintenir des états et des processus conscients. Nous ignorons comment en détail le cerveau cause la conscience, mais nous savons de fait que c'est ce qui se passe dans les cerveaux humains, et nous avons des preuves écrasantes que cela se passe aussi dans le cerveau de bon nombre d'espèces animales (Griffin, 1981). Nous ignorons à l'heure actuelle jusqu'à quel niveau au bas de l'échelle évolutionnaire s'étend la conscience.

À la base de notre conception du monde se trouve l'idée que les êtres humains et d'autres animaux supérieurs font partie de l'ordre biologique au même titre que n'importe quels autres organismes. Les humains sont en continuité avec le reste de la nature. Mais s'il en est ainsi, les caractéristiques biologiquement spécifiques de ces animaux – telles que le fait qu'ils possèdent un système riche de conscience, de même que leur plus grande intelligence, leur aptitude au langage, leur aptitude à des discriminations perceptuelles extrêmement fines, leur aptitude à la pensée rationnelle, etc. – sont des phénomènes biologiques au même titre que n'importe quels autres phénomènes biologiques. Qui plus est, ces caractères sont tous des phénotypes. Ils sont autant le résultat de l'évolution biologique que n'importe quel autre phénotype. *Bref, la conscience, est un trait biologique du cerveau humain et de certains cerveaux animaux. Elle est causée par des processus neurobiologiques et fait autant partie de l'ordre biologique naturel que n'importe quels autres traits biologiques tels que la photosynthèse, la digestion, la mitose.* Ce principe est le premier

stade dans la compréhension de la place qu'occupe la conscience dans notre conception du monde [4]. La thèse, développée jusqu'à maintenant, est la suivante : une fois que l'on a compris combien les théories atomiques et évolutionnistes sont centrales pour la conception scientifique contemporaine du monde, la conscience se met naturellement en place et apparaît comme un trait phénotypique évolué de certains types d'organismes dotés de systèmes nerveux éminemment développés. Toutefois, je ne défends pas cette conception du monde. À dire vrai, bon nombre de penseurs, dont je respecte les opinions, au premier rang desquels Wittgenstein, la jugent à divers degrés insoutenable, voire répugnante, dégradante. Elle leur paraît ne laisser aucune place – tout au plus un strapontin – à la religion, à l'art, au mysticisme, et aux valeurs « spirituelles » en général. Mais, que cela plaise ou non, c'est la conception contemporaine du monde. Étant donné ce que nous savons des détails du monde – de choses telles que la position des éléments dans le tableau périodique, le nombre de chromosomes dans des cellules de différentes espèces, et la nature du lien chimique – cette conception du monde n'est pas une option. Elle ne nous est pas simplement jetée en pâture avec une foule d'autres conceptions du monde rivales. Notre problème n'est pas que d'une manière ou d'une autre nous n'avons pas réussi à trouver une démonstration convaincante de l'existence de Dieu ou que cette hypothèse d'une vie après la mort reste sérieusement douteuse ; c'est plutôt que, dans nos réflexions les plus profondes, nous ne pouvons pas prendre au sérieux de telles opinions. Lorsque nous rencontrons des gens qui prétendent croire à de telles choses, nous pouvons leur envier le réconfort et la sécurité qu'ils affirment tirer de ces croyances, mais au fond, nous restons convaincus que ou bien ils n'ont pas entendu les nouvelles, ou bien ils sont sous l'emprise de la foi. Nous restons convaincus que, d'une manière ou d'une autre, il leur faut séparer leur esprit en

compartiments distincts pour croire de telles choses. Lorsque j'ai fait une conférence sur le problème des rapports du corps et de l'esprit en Inde, et que plusieurs membres du public m'ont assuré que mes idées ne pouvaient être qu'erronées, parce qu'ils avaient personnellement existé dans leurs vies antérieures sous forme de grenouilles, d'éléphants, etc., je n'ai pas pensé : « Voici la preuve en faveur d'une autre conception possible du monde », ni même : « Qui sait, peut-être ont-ils raison ? » Et mon insensibilité était bien plus que du pur et simple provincialisme culturel : étant donné ce que je sais de la marche du monde, je ne pouvais pas considérer que leurs conceptions fussent des candidats sérieux à la vérité.

Or, une fois que vous acceptez notre conception du monde, le seul obstacle qui interdit à la conscience d'avoir son statut de propriété biologique des organismes, c'est le postulat désuet dualiste/matérialiste selon lequel le caractère « mental » de la conscience l'empêche d'être une propriété « physique ».

J'ai seulement discuté de la relation qu'entretient la conscience avec les systèmes vivants à base de carbone du genre de ceux que nous connaissons sur terre, mais bien entendu, nous ne pouvons exclure la possibilité que la conscience ait pu évoluer sur d'autres planètes, dans d'autres systèmes solaires, dans d'autres parties de l'univers. Ne serait-ce que du fait de la taille de l'univers, il serait statistiquement sidérant que nous y soyons les seuls porteurs de conscience. Nous ne saurions davantage exclure la possibilité que la conscience ait pu évoluer dans des systèmes qui ne sont pas à base de carbone mais utilisent telle autre sorte de chimie. Pour autant qu'on le sache à l'heure actuelle, il se pourrait qu'il n'y ait aucun obstacle théorique au développement de la conscience dans des systèmes composés d'autres éléments. Nous sommes actuellement très loin d'avoir une théorie adéquate de la neurophysiologie de la conscience ; mais tant

que nous n'en n'aurons pas, il nous faut garder un esprit ouvert sur ses bases chimiques possibles. Mon pressentiment personnel serait que la neurobiologie de la conscience risque de se révéler au moins aussi limitée que, disons, la biochimie de la digestion. Il y a différentes sortes de digestion, mais n'importe quoi ne peut pas être digéré par n'importe quoi. Pareillement, il me semble que nous découvrirons probablement que, même s'il peut y avoir des variétés biochimiquement différentes de conscience, tout n'est pas permis.

En outre, étant donné que la conscience est entièrement l'effet du comportement de phénomènes biologiques de niveau inférieur, il serait en principe possible de la produire artificiellement en reproduisant exactement les pouvoirs causaux du cerveau en laboratoire. Nous savons que bien des phénomènes biologiques ont été créés artificiellement. Nous pouvons synthétiser certains composés organiques, et même créer artificiellement certains processus biologiques tels que la photosynthèse. Si nous pouvons artificiellement créer la photosynthèse, pourquoi pas aussi la conscience ? Pour la photosynthèse, la forme artificielle du phénomène a été créée en reproduisant exactement les processus chimiques en laboratoire. De même, si l'on devait créer artificiellement la conscience, la manière naturelle de procéder serait d'essayer de reproduire exactement la base neurobiologique réelle qu'a la conscience dans des organismes tels que nous. Et comme à ce jour nous ignorons exactement quelle est cette base neurobiologique, les perspectives d'une telle « intelligence artificielle » sont très lointaines. De surcroît, comme je l'ai suggéré auparavant, il se pourrait qu'on puisse produire de la conscience en utilisant un genre de chimie absolument différent de celui qu'utilisent en fait nos cerveaux. Néanmoins, une chose que nous savons avant même de commencer la recherche, c'est que *tout système capable de causer de la conscience doit être capable de reproduire exactement les pouvoirs causaux du cerveau.* Si, par exemple, cela se fait à l'aide de

pastilles de silicium, au lieu de neurones, ce doit être parce que la chimie des pastilles de silicium est capable de reproduire exactement les pouvoirs causaux spécifiques qu'ont les neurones de causer la conscience. Du fait que les cerveaux causent la conscience, n'importe quel autre système capable de causer la conscience, mais en utilisant des mécanismes complètement différents, devrait, en toute logique, avoir au moins le pouvoir équivalent des cerveaux pour le faire. (Comparez : les avions n'ont pas besoin de plumes pour voler, mais ils ont besoin de partager avec les oiseaux la capacité causale de triompher de la force de gravité dans l'atmosphère terrestre.)

En résumé : notre image du monde, bien qu'extrêmement compliquée dans le détail, fournit une analyse plutôt simple du mode d'existence de la conscience. D'après la théorie atomique, le monde est fait de particules. Ces particules s'organisent en systèmes. Certains de ces systèmes sont vivants, et ces types de systèmes vivants ont évolué sur de longues périodes de temps. Parmi eux, certains ont produit, par évolution, des cerveaux capables de causer et de maintenir la conscience. La conscience est donc un caractère biologique de certains organismes, dans l'acception même que l'on donne au terme « biologique » lorsque nous disons que la photosynthèse, la mitose, la digestion et la reproduction sont des caractères biologiques des organismes.

J'ai essayé de décrire la position de la conscience dans notre conception globale du monde en termes très simples, tant j'aimerais que cela paraisse absolument évident. Quiconque a eu, ne serait-ce qu'un minimum d'éducation « scientifique » après les années 1920, ne devrait rien trouver à redire dans les lignes qui précèdent. On me permettra d'ajouter que tout cela a été formulé sans recourir à aucune des catégories cartésiennes traditionnelles. Il n'a point été question de dualisme, de monisme, de matérialisme ou de quoi que ce soit de ce genre. En outre, il n'a pas été question

de « naturalisation de la conscience » ; elle est déjà complètement naturelle. La conscience, répétons-le, est un phénomène biologique naturel. L'exclusion de la conscience du monde naturel fut un procédé heuristique utile au XVII^e siècle, parce qu'il a permis aux savants de se concentrer sur des phénomènes qui étaient mesurables, objectifs et dénués de sens, c'est-à-dire dépourvus d'intentionnalité. Mais l'exclusion reposait sur une erreur. Elle reposait sur la croyance fausse que la conscience ne fait pas partie du monde naturel. Cette erreur, à elle seule, plus que toute autre chose, plus encore que la pure et simple difficulté qu'il y a à étudier la conscience avec les instruments scientifiques dont nous disposons, nous a empêchés de parvenir à une compréhension de la conscience.

La subjectivité

Les états et processus mentaux conscients ont une caractéristique bien particulière que ne possèdent pas d'autres phénomènes naturels, à savoir, la subjectivité. C'est cet aspect de la conscience qui dans son étude fait tant obstacle aux méthodes conventionnelles de la recherche biologique et psychologique, et qui est un problème pour l'analyse philosophique. Il y a différents sens de « subjectivité » et aucun d'entre eux n'est entièrement clair. Il me faut clarifier le sens que je donne à l'expression la conscience est subjective.

Nous parlons souvent de jugements en les qualifiant de « subjectifs » lorsque nous voulons dire par là que leur vérité ou leur fausseté ne peuvent être établies « objectivement », parce que la vérité ou la fausseté ne sont pas de simples questions de fait, mais dépendent de certaines attitudes, sentiments et points de vue de la part de ceux qui portent le jugement et de ceux qui l'entendent. Dire : « Van Gogh est un meilleur artiste que Matisse », c'est émettre un jugement subjectif. En ce sens

de « subjectivité », nous opposons de tels jugements subjectifs à des jugements tout à fait objectifs, tel le jugement : « Matisse a vécu à Nice durant l'année 1917. » Pour de tels jugements objectifs, nous pouvons établir quelles sortes de faits dans le monde les rendent vrais ou faux indépendamment des attitudes ou sentiments que l'on peut avoir à leur égard.

Or, cette acception du terme selon laquelle nous qualifions des jugements d'« objectifs » et de « subjectifs » n'est pas celle que je donne à « subjectif » quand je dis de la conscience qu'elle est subjective. Au sens où j'utilise ici ce terme, « subjectif » fait référence à une catégorie ontologique, et non à un mode épistémologique. Considérez par exemple l'énoncé : « J'ai en ce moment une douleur au bas du dos. » Cet énoncé est complètement objectif, au sens où il est rendu vrai par l'existence d'un fait réel et où il ne dépend pas d'une quelconque posture, des attitudes ou des opinions des observateurs. Pourtant, le phénomène en soi, la douleur réelle en soi, a un mode subjectif d'existence, et c'est en ce sens que je dis que la conscience est subjective.

Que pouvons-nous dire de plus sur ce mode subjectif d'existence ? D'abord, il est essentiel de voir que du fait de sa subjectivité, les observateurs n'ont pas tous pareillement accès à la douleur. Son existence, pourrions-nous dire, est une existence à la première personne. Pour que ce soit une douleur, il faut que ce soit la douleur de *quelqu'un* ; et cela en un sens bien plus fort que le sens où une jambe doit être la jambe de quelqu'un, par exemple : des transplantations de jambe sont possibles, pas des transplantations de douleurs. Et ce qui est vrai des douleurs l'est des états conscients en général. Tout état conscient est l'état conscient de *quelqu'un*. Et de même que j'ai une relation particulière avec mes états conscients, qui ne ressemble pas à la relation que j'ai avec les états conscients d'autres personnes, de même celles-ci, à leur tour, ont une relation avec leurs états conscients, qui ne ressemble pas à celle que j'ai avec les leurs [5]. Voici une autre

conséquence de la subjectivité : toutes mes formes conscientes d'intentionnalité qui me donnent une information sur le monde indépendamment de moi se font toujours à partir d'un point de vue particulier. Le monde lui-même n'a pas de point de vue ; en revanche, mon accès au monde par le biais de mes états conscients a toujours une perspective, il se fait toujours à partir de mon point de vue.

On aurait du mal à exagérer les effets désastreux qu'a eus sur le travail philosophique et psychologique du dernier demi-siècle l'incapacité à accepter la subjectivité de la conscience. La faillite d'une bonne partie des travaux menés en philosophie de l'esprit et la stérilité de la psychologie académique au cours des cinquante dernières années, et de l'ensemble de ma vie intellectuelle, sont, d'une manière qui d'emblée n'a rien d'évident, le fruit de l'incapacité persistante à reconnaître et à accepter le fait que l'ontologie du mental est une ontologie irréductiblement à la première personne. Il y a de très profondes raisons, dont beaucoup sont gravées dans notre histoire inconsciente, qui font que nous trouvons difficile, sinon impossible, d'accepter l'idée que le monde réel, le monde décrit par la physique et la chimie et la biologie, contient un élément subjectif inéliminable.

Comment pareille chose pourrait-elle être ? Comment pouvons-nous donc parvenir à une image cohérente du monde si le monde contient ces mystérieuses entités conscientes ? Pourtant nous savons tous que nous sommes la majeure partie de notre vie conscients, et que d'autres personnes autour de nous le sont aussi. Et, sauf à se laisser aveugler par de la mauvaise philosophie ou telle ou telle forme de psychologie académique, nous ne doutons pas sérieusement une seconde que les chiens, les chats, les singes et les petits enfants sont conscients, et que leur conscience est tout aussi subjective que la nôtre.

Essayons donc de décrire un peu plus en détail l'image

du monde qui contient la subjectivité à titre d'élément fondamental, et tentons alors de décrire certaines des difficultés que nous avons à accepter cette image du monde. Si nous pensons que le monde est constitué de particules, que ces particules s'organisent en systèmes, que certains de ces systèmes sont des systèmes biologiques, que certains de ces systèmes biologiques sont conscients, et que la conscience est fondamentalement subjective – alors que nous est-il au juste demandé d'imaginer lorsque nous imaginons la subjectivité de la conscience ? Somme toute chacune des autres choses que nous avons imaginées – particules, systèmes, organismes, etc. – était complètement objective. Elles sont donc pareillement accessibles à tous les observateurs compétents. Aussi, que nous est-il demandé d'imaginer si nous devons à présent jeter dans cette marmite métaphysique quelque chose d'irréductiblement subjectif ?

Il nous est, en réalité, demandé d'« imaginer » simplement le monde dont nous connaissons l'existence. Je sais, par exemple, que je suis à présent conscient, et que cet état conscient possède cette subjectivité à laquelle je faisais référence, et je sais qu'un très grand nombre d'autres organismes semblables à moi sont pareillement conscients et ont des états subjectifs similaires. Alors pourquoi cette impression de difficulté à s'imaginer quelque chose qui irait en quelque sorte à l'encontre de nos intuitions, alors que je ne fais que nous rappeler des faits que nous avons tout du long sous le nez ? La réponse tient en partie – mais seulement en partie – au fait que j'ai naïvement invoqué le terme d'« observateur » au paragraphe précédent. Lorsqu'il nous est demandé de former une *conception* ou une *image* du monde, nous les formons sur le modèle de la vision. Nous avons tendance littéralement à former une image de la réalité comme consistant en de minuscules morceaux de matière, les « particules », puis à imaginer celles-ci organisées en systèmes, dotés à nouveau de caractères visibles à l'œil nu. Mais quand

nous visualisons le monde, avec cet œil intérieur, nous ne voyons pas la conscience. En fait, c'est la subjectivité même de la conscience qui la rend invisible de manière décisive. *Si nous essayons de dessiner une image de la conscience de quelqu'un d'autre, nous finissons tout simplement par dessiner l'autre personne* (avec, peut-être, comme dans les bandes dessinées, un phylactère). *Si nous essayons de dessiner notre propre conscience, nous finissons par dessiner tout ce dont nous pouvons être conscients.* Si la conscience est la base épistémique fondamentale qui nous fait parvenir à la réalité, nous ne pouvons parvenir à la réalité de la conscience de cette manière. (Autre formulation possible : nous ne pouvons parvenir à la réalité de la conscience à la manière dont, en utilisant la conscience, nous pouvons parvenir à la réalité d'autres phénomènes.)

Il est important de s'arrêter sur ce point au lieu, comme à l'habitude, de passer à toute vitesse. Si j'essaie d'observer la conscience d'autrui, ce que j'observe ce n'est pas sa subjectivité, mais simplement son comportement conscient, sa structure et les relations causales qui existent entre la structure et le comportement, d'un côté, et l'environnement qui l'affecte et qu'il affecte à son tour, de l'autre. Ainsi ne puis-je en aucune manière observer la conscience de quelqu'un d'autre comme telle ; ce que j'observe plutôt, c'est lui et son comportement et les relations qui existent entre lui, le comportement, la structure et l'environnement. Et qu'en est-il de ce qui se passe à l'intérieur de moi ? Ne puis-je l'observer ? Le fait même de la subjectivité, que nous essayions d'observer, rend cette observation impossible. Pourquoi ? Parce que là où il est question de subjectivité consciente, il n'y a aucune distinction entre l'observation et la chose observée, entre la perception et l'objet perçu. Le modèle de la vision repose sur le présupposé qu'il y a une distinction entre la chose vue et le fait de la voir. Mais pour l'« introspection », il est tout bonnement impossible de faire cette

séparation. Toute introspection, quelle qu'elle soit, que j'ai de mon propre état conscient est elle-même cet état conscient. Ce n'est pas dire que mes phénomènes mentaux conscients ne se présentent pas à différents niveaux et sous différentes formes – nous aurons l'occasion d'examiner certains d'entre eux en détail plus loin –, c'est simplement dire que le modèle usuel d'observation ne marche tout bonnement pas pour la subjectivité consciente. Il ne marche pas pour la conscience d'autres personnes, et il ne marche pas pour la nôtre. Pour cette raison, l'idée qu'il puisse y avoir une méthode spécifique d'examen de la conscience, à savoir l'« introspection », qui serait prétendument une sorte d'observation intérieure, était vouée à l'échec dès le départ, et il n'est pas surprenant que la psychologie introspective ait fait faillite.

Nous avons du mal à accepter la subjectivité, non seulement parce que nous avons été élevés dans une idéologie qui dit qu'en dernière analyse, la réalité doit être complètement objective, mais parce que notre idée d'une réalité objectivement observable présuppose la notion d'observation, qui est en soi inéluctablement subjective, et qui ne peut elle-même devenir un objet d'observation à la manière dont le peuvent des objets et des états de choses existant objectivement dans le monde. Il n'y a pour nous, en d'autres termes, aucun moyen de dépeindre la subjectivité comme une partie de notre conception du monde parce que, pour ainsi dire, la subjectivité en question est l'acte de dépeindre. La solution n'est pas d'essayer de développer un mode spécifique d'acte de dépeindre, une sorte de superintrospection, mais plutôt de cesser à ce stade ni plus ni moins de dépeindre, et de reconnaître simplement les faits. Les faits sont que les processus biologiques produisent des phénomènes mentaux conscients, et que ceux-ci sont irréductiblement subjectifs.

Les philosophes ont inventé une autre métaphore pour décrire certains aspects de la subjectivité qui me paraît encore plus confuse que la métaphore de l'introspection développée

par le sens commun ; c'est la métaphore de l'« accès privilégié ». À la métaphore « visuelle » de l'introspection, nous sommes tentés de substituer la métaphore *spatiale* de l'accès privilégié, un modèle qui suggère que la conscience ressemble à une pièce privée dans laquelle nous sommes seuls autorisés à entrer. Je suis le seul à pouvoir pénétrer à l'intérieur de l'espace de ma propre conscience. Mais cette métaphore ne résout rien : pour qu'il y ait quelque chose à quoi j'aie un accès privilégié, il me faudrait, en effet, être différent de l'espace dans lequel je pénètre. De même que la métaphore de l'introspection s'est effondrée lorsque la seule chose à observer était le fait d'observer lui-même, la métaphore d'un espace intérieur privé s'effondre lorsque nous comprenons qu'il n'y a pas le moindre espace où je puisse pénétrer, parce que je suis incapable de faire les distinctions nécessaires entre les trois éléments que sont : moi-même, l'acte d'entrer et l'espace dans lequel je suis supposé entrer.

Nous pourrions résumer ces points en disant que notre modèle moderne de la réalité et de la relation entre réalité et observation ne peut pas accueillir le phénomène de la subjectivité. Le modèle est un modèle d'observateurs objectifs (au sens épistémique) observant une réalité objectivement existante (au sens ontologique). Mais en aucune manière ce modèle ne permet d'observer l'acte d'observer lui-même. Car l'acte d'observer est l'accès subjectif (sens ontologique) à la réalité objective. Même si je puis facilement observer une autre personne, je ne puis observer sa *subjectivité*. Et pire encore, je ne puis *observer* ma propre subjectivité, car n'importe quelle observation que je puisse envisager de faire est celle-là même qui était censée être observée. Ce que l'on a en tête quand on songe à une observation de la réalité, c'est précisément l'idée de représentations (ontologiquement) subjectives de la réalité. L'ontologie de l'observation – par opposition à son épistémologie – est précisément l'ontologie de la subjectivité. L'observation est toujours l'observation de

quelqu'un ; elle est généralement consciente ; elle s'effectue toujours à partir d'un point de vue ; il émane d'elle une impression de subjectivité, etc.

Je ne reprends pas le vieil argument confus suivant lequel l'étude de la subjectivité comporte un paradoxe autoréférentiel. De tels paradoxes ne me gênent pas du tout. Nous pouvons nous servir de l'œil pour étudier l'œil, du cerveau pour étudier le cerveau, de la conscience pour étudier la conscience, du langage pour étudier le langage, de l'observation pour étudier l'observation, et de la subjectivité pour étudier la subjectivité. Je ne vois en tout cela aucun problème. Ce qui est plutôt en cause, c'est le fait qu'en raison de l'ontologie de la subjectivité, nos modèles d'« étude » – des modèles qui reposent sur la distinction entre observation et chose observée – ne marchent pas pour la subjectivité elle-même.

Nous avons du mal à concevoir la subjectivité en un certain sens. Étant donné notre concept de ce à quoi doit ressembler la réalité et des découvertes de cette réalité qu'il serait possible de faire, il nous paraît inconcevable qu'il puisse y avoir quoi que ce soit d'irréductiblement subjectif dans l'univers. Pourtant nous savons tous que la subjectivité existe.

J'espère que nous gagnerons en clarté si nous essayons de décrire l'univers en laissant de côté la subjectivité. Supposons que nous insistions pour donner une analyse du monde qui soit complètement objective, pas seulement au sens épistémique où ses thèses sont indépendamment contrôlables, mais au sens ontologique où les phénomènes qu'elle décrit ont une existence indépendante de toute forme de subjectivité. Une fois que vous adoptez cette stratégie (la principale stratégie dans la philosophie de l'esprit des cinquante dernières années), il devient alors impossible de décrire la conscience, parce qu'il devient littéralement impossible de reconnaître la subjectivité de la conscience. Les exemples en sont vraiment trop nombreux pour qu'on puisse les mention-

ner ici, mais je citerai deux auteurs qui visent explicitement le problème de la conscience. Armstrong (1980) élimine tacitement la subjectivité en traitant simplement la conscience comme une capacité à faire des discriminations sur les propres états intérieurs que l'on peut avoir, et Jean-Pierre Changeux, pour sa part, définit simplement la conscience comme un « système de régulation global qui porte sur les objets mentaux et sur leurs calculs » (1985, p. 145). Ces deux analyses présupposent toutes deux une conception de la réalité à la troisième personne, une conception d'une réalité qui n'est pas objective de manière purement épistémique, mais qui est aussi ontologiquement objective ; et une réalité de ce genre n'accorde aucune place à la conscience, parce qu'elle n'accorde aucune place à la subjectivité ontologique.

La conscience et le problème des rapports du corps et de l'esprit

Je n'ai de cesse de répéter que le problème des rapports du corps et de l'esprit a une solution relativement simple, du moins dans ses grandes lignes, et il n'y a que deux obstacles qui nous empêchent d'avoir une pleine compréhension des rapports entre le corps et l'esprit : le préjugé philosophique qui nous fait supposer que le mental et le physique sont deux domaines distincts, et notre ignorance des opérations du cerveau. Si nous disposions d'une science adéquate du cerveau, d'une analyse du cerveau qui nous donnât des analyses causales de la conscience sous toutes ses formes et espèces, et si nous surmontions nos erreurs conceptuelles, il n'y aurait plus de problème concernant les rapports du corps et de l'esprit. Toutefois, la possibilité d'une solution quelconque à ce problème a été très fortement contestée au fil des années par Thomas Nagel (1974, 1986). Son argumentation est la suivante : à l'heure actuelle, nous ne

disposons tout bonnement pas de l'appareillage conceptuel qui nous permettrait ne serait-ce que de concevoir une solution au problème des rapports entre le corps et l'esprit. Parce que les explications causales fournies par les sciences de la nature ont, selon Nagel, une sorte de nécessité causale. Nous comprenons, par exemple, comment le comportement des molécules d'H_2O est la cause de l'état liquide de l'eau, parce que nous voyons que la liquidité est une conséquence nécessaire du comportement des molécules. La théorie moléculaire fait plus que montrer que les systèmes de molécules d'H_2O seront liquides dans certaines conditions ; elle montre plutôt pourquoi le système *doit être* sous forme liquide. Supposé que nous comprenions la physique en question, il est inconcevable que si les molécules se comportent ainsi, l'eau ne soit pas alors dans un état liquide. En bref, soutient Nagel, les explications scientifiques impliquent la nécessité, et la nécessité implique l'inconcevabilité du contraire.

Or, dit Nagel, nous ne pouvons parvenir à ce type de nécessité s'agissant des rapports entre la matière et la conscience. Aucune analyse possible du comportement neuronal n'expliquerait pourquoi, étant donné ce comportement, *il nous faut,* par exemple, avoir mal. Aucune analyse n'expliquerait pourquoi la douleur est une conséquence nécessaire de certaines sortes de déclenchements neuronaux. La preuve que l'analyse ne nous donne pas de nécessité causale, c'est que nous pouvons toujours concevoir le contraire. Nous pouvons toujours concevoir un état de choses dans lequel la neurophysiologie se comporte exactement comme on veut, et où, néanmoins, le système n'éprouve aucune douleur. Si l'explication scientifique adéquate implique la nécessité et que la nécessité implique que le contraire est inconcevable, alors, par contraposition, la possibilité de concevoir le contraire implique que nous n'avons pas de nécessité, ce qui implique à son tour que nous n'avons pas d'explication. La conclusion désespérante de Nagel est qu'il nous faudrait une refonte

majeure de notre appareillage conceptuel si nous devions un jour pouvoir résoudre le problème des rapports du corps et de l'esprit.

Je ne suis pas convaincu par cette argumentation. D'abord, il nous faut observer que les explications scientifiques n'ont pas toutes le genre de nécessité que nous avons trouvée dans la relation entre le mouvement des molécules et la liquidité. Ainsi, la loi du carré inverse est une analyse de la gravité, mais elle ne montre pas pourquoi les corps *doivent avoir* une attraction gravitationnelle. En second lieu, l'apparente « nécessité » de n'importe quelle explication scientifique peut être juste fonction du fait que nous trouvons l'explication si convaincante que nous ne pouvons, par exemple, concevoir que les molécules se déplacent d'une certaine manière et que les H_2O ne soient pas liquides. Il se pourrait qu'une personne vivant dans l'Antiquité ou au Moyen Âge n'ait pas trouvé que l'explication soit affaire de « nécessité ». Le « mystère » de la conscience aujourd'hui se présente grossièrement sous la même forme que le mystère de la vie avant le développement de la biologie moléculaire ou que le mystère de l'électromagnétisme avant les équations de Clerk-Maxwell. Le mystère vient de ce que nous ignorons comment fonctionne le système de la neurophysiologie/conscience, et une connaissance satisfaisante de son fonctionnement dissoudrait le mystère. De surcroît, il se pourrait que la thèse selon laquelle nous pourrions toujours concevoir la possibilité que certains états cérébraux *puissent ne pas* causer les états conscients appropriés dépende simplement de notre ignorance du fonctionnement du cerveau. Il me semble que si nous comprenions parfaitement la structure du cerveau, il nous paraîtrait probablement évident que, si le cerveau était dans telle sorte d'état, il serait alors nécessairement conscient. Notons que nous acceptons déjà cette forme de nécessité causale des états conscients pour des phénomènes molaires grossiers. Par exemple, si je vois un homme en train de

hurler, le pied pris dans une perforatrice, alors je sais que l'homme doit avoir horriblement mal. Il est, en un sens, inconcevable pour moi qu'un être humain normal se trouve dans une telle situation et n'éprouve pas une douleur horrible. Les causes physiques déterminent nécessairement la douleur.

Accordons toutefois ce point à Nagel, pour les besoins de l'argument. Rien n'en découle quant à la manière dont le monde fonctionne en fait. La limitation que souligne Nagel n'est qu'une limitation de nos pouvoirs de concevoir. Même en admettant qu'il ait raison, ce que montre son argumentation c'est seulement que, dans le cas des relations entre des phénomènes matériels et des phénomènes matériels, nous pouvons nous figurer subjectivement les deux côtés de la relation ; au contraire, dans le cas de relations entre des phénomènes matériels et des phénomènes mentaux, un côté de la relation est déjà subjectif, et nous ne pouvons donc nous figurer sa relation avec ce qui est matériel comme nous le pouvons des relations entre la liquidité et le mouvement des molécules, par exemple. L'argumentation de Nagel, en un mot, montre seulement que nous ne pouvons sortir de la subjectivité de la conscience, pour voir la relation nécessaire qu'elle a avec sa base matérielle. Nous formons une image de la nécessité fondée sur notre subjectivité, mais nous ne pouvons de cette façon former une image de la nécessité de la relation entre la subjectivité et les phénomènes neurophysiologiques, parce que nous sommes déjà dans la subjectivité, et que la relation d'acte de dépeindre exigerait que nous en sortions. (Si la solidité était consciente, il lui paraîtrait mystérieux qu'elle soit causée par les mouvements vibratoires de molécules dans des structures en treillis, et pourtant ces mouvements expliquent la solidité.)

Pour apprécier cette objection adressée à Nagel, il vous suffit d'imaginer d'autres façons de détecter des relations causalement nécessaires. Supposons que Dieu ou qu'une machine puisse simplement détecter des relations causalement

nécessaires ; pour Dieu ou la machine il n'y aurait alors aucune différence entre des formes matière/matière de nécessité et des formes matière/esprit de nécessité. En outre, même si nous admettons que nous ne pouvons dépeindre les deux côtés de la relation pour la conscience et le cerveau, comme nous pouvons le faire pour la liquidité et le mouvement des molécules, nous pourrions néanmoins parvenir aux relations causales impliquées dans la production de la conscience par des moyens indirects. Supposons que nous sachions réellement rendre compte des processus neurophysiologiques du cerveau qui causent la conscience. Il n'est pas du tout impossible que nous le puissions, puisque la mise en évidence habituelle de relations causales peut s'effectuer sur des relations cerveau/conscience comme pour n'importe quel phénomène naturel. La connaissance de relations causales nomologiques nous donnera toute la nécessité causale requise. En vérité, nous commençons déjà à comprendre des relations nomologiques de ce genre. Comme je l'ai indiqué au chapitre III, les manuels ordinaires de neurophysiologie nous expliquent régulièrement, par exemple, les ressemblances et les différences qu'il y a entre la manière dont les chats voient les choses et celle dont les humains les voient. Il est incontestable que certaines sortes de ressemblances et de différences neurophysiologiques sont causalement suffisantes pour certaines sortes de ressemblances et de différences qui sont à l'œuvre dans les expériences visuelles. En outre nous pouvons – et c'est du reste ce que nous allons faire – décomposer la grande question : comment le cerveau cause-t-il la conscience ? en un grand nombre de questions plus élémentaires (par exemple : comment la cocaïne produit-elle certaines expériences caractéristiques ?). Et les réponses détaillées que nous commençons déjà à donner (par exemple : la cocaïne empêche la capacité qu'ont certains récepteurs synaptiques de réabsorber la noradrénaline) permettent déjà les inférences caractéristiques qui vont de pair avec la nécessité causale (par exemple : si

vous augmentez la dose de cocaïne, vous augmentez l'effet). Je conclus que Nagel n'a pas montré que le problème des rapports du corps et de l'esprit est insoluble, même en restant dans le cadre de l'appareillage conceptuel et de la conception du monde qui sont communément les nôtres.

Colin McGinn (1991) porte l'argumentation de Nagel un pas plus loin et soutient qu'il est impossible *en principe* que nous puissions jamais être en mesure de comprendre la solution au problème des rapports du corps et de l'esprit. Son argumentation va au-delà de celle de Nagel et implique des postulats que Nagel ne fait pas, du moins pas explicitement. Comme les postulats de McGinn sont largement partagés dans la tradition philosophique du dualisme, et comme dans ce livre j'essaie – entre autres choses – de venir à bout de ces postulats, je les énoncerai explicitement et essaierai de montrer qu'ils sont faux. McGinn postule que :

1. La conscience est une sorte de « matériau » (*stuff*) [6].
2. Ce matériau est connu par la « faculté d'introspection ». La conscience est l'« objet » de la faculté introspective, de même que le monde physique est l'objet de la faculté perceptive (p. 14 *sq.* et p. 61 *sq.*).

En conséquence de 1 et de 2, bien que je ne sois pas sûr que McGinn souscrive à cela, la conscience, comme telle, telle qu'elle est connue par introspection, n'est pas spatiale ; à l'opposé du monde physique, qui, comme tel, tel qu'il est connu par la perception, est spatial.

3. Pour comprendre le problème des rapports du corps et de l'esprit, il nous faudrait comprendre le « lien » qui existe entre la conscience et le cerveau (*passim*).

McGinn ne doute pas qu'il y ait un tel « lien », mais il croit qu'il nous est impossible en principe de le comprendre. Il dit, reprenant le terme de Kant, que, pour nous, la relation est « nouménale ». Il nous est impossible de comprendre ce

lien, et partant, impossible de comprendre les relations du corps et de l'esprit. La conjecture que fait McGinn est que le lien est assuré par une structure cachée de la conscience qui est inaccessible à l'introspection.

Ces trois thèses sont des postulats cartésiens et la solution « proposée » est une solution de type cartésien (avec un inconvénient en plus, c'est que la structure cachée de la conscience est inconnaissable en principe. La glande pinéale au moins était accessible !). Toutefois, comme dans le cas de la glande pinéale, la solution n'en est pas une. Si vous avez besoin d'un lien entre la conscience et le cerveau, alors il vous faut un lien entre la structure cachée de la conscience et le cerveau. La postulation d'une structure cachée – même si elle était intelligible – ne nous mène nulle part.

Le vrai problème a trait aux trois postulats ; je crois, en effet, qu'ils incarnent la plupart des erreurs commises durant les trois cents dernières années par le dualisme. Plus particulièrement :

1. La conscience n'est pas un « matériau », c'est une *caractéristique* ou une *propriété* du cerveau au sens où, par exemple, la liquidité est une caractéristique de l'eau.
2. La conscience n'est pas connue par introspection d'une manière analogue à celle dont les objets du monde sont connus par perception. Je développe ce point au chapitre suivant, et j'ai déjà commencé à le discuter dans celui-ci, aussi le formulerai-je très simplement : le modèle consistant à « *inspecter intro* », c'est-à-dire le modèle d'une inspection intérieure, requiert une distinction entre l'acte d'inspecter et l'objet inspecté ; or, nous ne pouvons faire une telle distinction pour la conscience. La doctrine de l'introspection est un bon exemple de ce que Wittgenstein appelle l'ensorcellement de notre intelligence par l'intermédiaire du langage.

En outre, une fois que vous vous êtes débarrassé de l'idée que la conscience est un matériau qui est l'« objet » de l'introspection, il est facile de voir qu'elle est spatiale, parce qu'elle est localisée dans le cerveau. Nous n'avons aucune connaissance immédiate dans l'expérience consciente ni de la localisation spatiale ni des dimensions de notre expérience consciente, mais pourquoi le faudrait-il ? C'est une question neurophysiologique extrêmement délicate, une question que nous sommes loin de résoudre, que de se figurer exactement quel est le lieu de l'expérience consciente dans notre cerveau. Il se pourrait fort bien, pour autant qu'on le sache, qu'il soit distribué sur de très larges portions du cerveau.

3. Il n'y a pas plus de « lien » entre la conscience et le cerveau qu'il n'y en a entre la liquidité de l'eau et les molécules d'H_2O. Si la conscience est une caractéristique du cerveau de niveau supérieur, alors il ne peut y avoir en aucun cas de lien entre la caractéristique et le système dont elle est une caractéristique.

La conscience et l'avantage sélectif

Mon approche en philosophie de l'esprit, le naturalisme biologique, est parfois confrontée à la difficulté suivante : si nous avons pu imaginer qu'un zombie inconscient ait pu produire le même comportement ou un comportement similaire, pourquoi, dans ces conditions, l'évolution a-t-elle tout simplement produit la conscience ? En vérité, on présente souvent cela en suggérant que la conscience n'existe peut-être même pas. Je ne vais naturellement pas essayer de démontrer l'existence de la conscience. Si quelqu'un n'est pas conscient, il n'y a aucune façon dont je puisse lui démontrer l'existence de la conscience ; s'il est conscient, il est relativement inconcevable qu'il puisse sérieusement douter qu'il le soit. Je ne dis pas qu'il n'existe pas de gens qui sont en proie à une telle confusion philosophique qu'ils *disent* douter qu'ils soient conscients, mais j'ai du mal à prendre de telles affirmations au sérieux.

En réponse à la question sur le rôle de l'évolution dans la production de la conscience, je voudrais rejeter le présupposé implicite selon lequel tout trait biologiquement hérité doit avoir un avantage évolutionniste pour l'organisme. Cela me paraît être du darwinisme très grossier, et nous avons à

présent toutes sortes de bonnes raisons de l'abandonner. S'il était vrai que toute prédisposition innée d'un organisme fût le résultat de quelque pression sélective, alors il me faudrait conclure que mon chien a été sélectionné pour aller chercher des balles de tennis. Il a une passion pour aller chercher les balles de tennis, et ce n'est de toute évidence pas quelque chose qu'il a appris, mais ce n'est pas une raison de supposer que cela provienne d'une quelconque rétribution biologique. Ou, plus proche de nous, la passion que les êtres humains ont pour le ski alpin a, je crois, une base biologique qui n'est pas le résultat de l'apprentissage ou du conditionnement. L'expansion qu'a connue le ski a été simplement phénoménale ; et les sacrifices que sont prêts à faire les gens en argent, en confort, en temps, pour pouvoir simplement passer quelques heures sur une piste de ski sont une assez bonne preuve qu'ils en tirent des satisfactions qui sont inhérentes à leur nature biologique. Mais cela ne veut pas dire du tout que nous avons été sélectionnés par l'évolution pour notre prédilection à faire du ski alpin [7].

Ces précisions étant données, nous pouvons toujours aborder la question : « Quel est l'avantage évolutionniste de la conscience ? » Et la réponse est que la conscience fait toutes sortes de choses. Pour commencer, il y a toutes sortes de formes de conscience telles que la vision, l'audition, le goût, l'odorat, la soif, les douleurs, les chatouillements, les démangeaisons et les actions volontaires. En second lieu, dans chacun de ces domaines, il peut y avoir une variété de fonctions assurées par les formes conscientes de ces différentes modalités. Néanmoins, pour parler en termes généraux, il paraît clair que la conscience sert à organiser un certain ensemble de relations entre l'organisme et son environnement comme ses propres états. Et, une fois encore, pour parler en termes très généraux, la forme d'organisation pourrait être décrite comme de la « représentation ». C'est ainsi que par l'entremise des modalités sensorielles, l'organisme acquiert

une information consciente sur l'état du monde. Il entend des sons à proximité ; il voit des objets et des états de choses dans son champ visuel ; il sent les odeurs spécifiques de traits distincts de son environnement, etc. En plus de son expérience sensorielle consciente, l'organisme aura aussi, de façon caractéristique, des expériences d'action. Il courra, marchera, mangera, combattra, etc. Ces formes de conscience n'ont pas pour but premier d'acquérir de l'information sur le monde ; ce sont plutôt des cas où la conscience permet à l'organisme d'agir sur le monde, de produire des effets dans le monde. Pour parler encore de façon très approximative – nous discuterons plus loin de ces points en termes plus précis – nous pouvons dire que, dans la perception consciente, l'organisme a des représentations causées par des états de choses du monde, et dans le cas d'actions intentionnelles, l'organisme cause des états de choses dans le monde au moyen de ses représentations conscientes.

Si cette hypothèse est correcte, nous pouvons émettre une thèse générale concernant l'avantage sélectif de la conscience. La conscience nous donne de bien plus grands pouvoirs de discrimination que n'en auraient des mécanismes inconscients. C'est ce que mettent en lumière les cas étudiés par Penfield (1975). Certains des patients de Penfield souffraient d'une forme d'épilepsie connue sous le nom de *petit mal*. Dans certains de ces cas, la crise d'épilepsie rendait le patient totalement inconscient, et pourtant le patient continuait à manifester ce qu'on appellerait normalement un comportement finalisé. Voici quelques exemples :

> Un patient, que j'appellerai A, était un bon étudiant de piano, et sujet à des crises du type que l'on nomme *petit mal*. Il lui arrivait souvent de faire une légère interruption alors qu'il était en train de jouer, que sa mère reconnaissait comme le début d'une « absence ». Il continuait alors à jouer pendant quelque temps avec une considérable dextérité. Le patient B était sujet à des crises d'épilepsie qui débutaient par une décharge dans le

> lobe temporal. Parfois l'attaque lui tombait dessus alors qu'il rentrait chez lui du travail. Il continuait à marcher et à se frayer son chemin à travers les rues bondées qu'il devait prendre pour rentrer à la maison. Il pouvait se rendre compte plus tard qu'il avait eu une attaque parce qu'il avait un trou de mémoire concernant une partie de son itinéraire, tel que de l'avenue X à la rue Y. Si le patient C conduisait une voiture, il continuait à conduire, bien qu'il se rendît compte plus tard qu'il avait brûlé un ou plusieurs feux rouges (p. 39).

Dans tous ces cas, nous avons affaire à des formes complexes de comportement apparemment finalisé sans la moindre conscience. Or pourquoi les comportements ne pourraient-ils pas être tous ainsi ? Qu'est-ce que la conscience ajoute ? On notera que, dans les cas en question, les patients étaient en train d'effectuer des types d'actions qui étaient habituelles, routinières et mémorisées. Il y avait probablement des voies neuronales bien établies dans le cerveau de l'individu correspondant à sa connaissance du chemin à prendre pour rentrer chez lui ; et de même, le pianiste avait probablement, réalisée dans les voies neuronales de son cerveau, la connaissance de la manière de jouer le morceau de piano particulier. Le comportement complexe peut être préprogrammé dans la structure du cerveau, pour autant du moins que nous sachions quelque chose du fonctionnement du cerveau en de pareils cas. Apparemment, une fois commencée, l'activité peut suivre son cours même lors d'une attaque de *petit mal*. En revanche, le comportement conscient normal, humain, a un degré de flexibilité et de créativité que l'on ne retrouve pas dans les cas étudiés par Penfield du conducteur inconscient et du pianiste inconscient. La conscience ajoute des pouvoirs de discrimination et de flexibilité même aux activités de routine mémorisées.

Selon toute apparence, c'est juste un fait de la biologie que les organismes qui ont une conscience ont, en général, de bien plus grands pouvoirs de discrimination que ceux qui n'en ont pas. Les tropismes des plantes, par exemple, qui

sont sensibles à la lumière, sont bien moins capables de faire des discriminations fines et bien moins flexibles que, par exemple, le système visuel humain. L'hypothèse que j'avance donc est que l'un des avantages de l'évolution, qui nous a été conféré par la conscience, est la bien plus grande flexibilité, sensibilité et créativité qui nous vient du fait d'être conscient.

Les traditions behavioriste et mécaniste dont nous avons hérité nous rendent aveugles à ces faits ; en vérité, elles rendent impossible ne serait-ce que de poser les questions de façon satisfaisante, parce qu'elles cherchent constamment des formes d'explication qui traitent le mental-neurophysiologique comme fournissant simplement un mécanisme entrée-sortie, une fonction de mise en correspondance des stimuli d'entrée avec les comportements de sortie. Les termes mêmes dans lesquels sont posées les questions empêchent d'introduire des sujets qui sont décisifs si l'on veut comprendre la conscience, tels que la créativité, par exemple.