in A. Barberousse, D. Bonnay, M. Cozic, dir., Précis de philosophie des sciences, Paris : Vuibert, 2011, pp. 519-571.

Chapitre XV

Philosophie des sciences cognitives

Les sciences cognitives se présentent comme un ensemble articulé de recherches visant à constituer une science de l'esprit. À certains égards, elles sont des sciences « comme les autres », et la philosophie des sciences cognitives ressemble à la philosophie d'autres sciences particulières. Mais par d'autres côtés les sciences cognitives sont très différentes de la plupart des disciplines ou groupes de disciplines, et en conséquence la philosophie des sciences cognitives diffère notablement de branches telles que la philosophie de la physique, la philosophie de la biologie ou la philosophie de l'économie.

On pourrait penser que la principale différence vient de la *pluralité* que recouvre le pluriel grammatical de « sciences cognitives ». Cette différence joue en effet un certain rôle, et explique que la philosophie des sciences cognitives ressemble un peu, par exemple, à la philosophie des sciences sociales. Mais l'unité des différentes disciplines est une affaire de degré, et il n'est pas possible d'en donner une évaluation qui n'engage pas des hypothèses théoriques conséquentes. On peut dire, en première approximation, que la physique est plus unifiée que la biologie (encore souvent désignée, précisément, par la locution plurielle « sciences de la vie »), que les sciences sociales sont nettement moins unifiées que les sciences de la vie, et que les sciences cognitives occupent entre ces deux derniers groupes une position intermédiaire. Ce n'est donc pas leur relative absence d'unité qui confère aux sciences cognitives leur singularité philosophique, même si cette pluralité interne est pour le philosophe un sujet de réflexion.

La *jeunesse* des sciences cognitives est un autre aspect qui semble les distinguer, expliquant notamment qu'elles soient largement méconnues, et qu'elles semblent fragiles : ainsi la philosophie peut-elle jouer à leur endroit un rôle – explicatif et défensif – qui fut le sien vis-à-vis des naissantes sciences physiques lors de la Révolution scientifique, avec tous les changements dus à la distance qui nous sépare de cette époque.

Mais la différence essentielle gît (selon moi, il y a là déjà matière à discussion) dans l'*incertitude persistante quant à leur objet* (qu'on désignera, conventionnellement, par le terme *cognition*) et, de manière concomitante, dans *l'interpénétration des sciences cognitives et de la philosophie*[1].

1. De la philosophie tout court, et non (simplement) de la philosophie des sciences cognitives (comme c'est le cas, *mutatis mutandis*, de toute science particulière et de la philosophie de cette science).

Quoi qu'il en soit, la philosophie des sciences cognitives est un domaine proliférant, immense, aux frontières mal définies et poreuses : il est souvent difficile de dire si l'on y est encore, ou si l'on a gagné une autre branche de la philosophie, ou bien une province de la science positive. Ces questions de démarcation sont d'importance relative, car ce sont les problèmes, et leurs interrelations, qui structurent la recherche, bien davantage que les étiquettes qu'on leur accole pour organiser les institutions et le travail des étudiants. Cependant, la conception que les philosophes se font du rôle qui peut ou doit être le leur à l'égard des sciences cognitives fait l'objet de divergences doctrinales. Il est donc utile d'avoir au moins une idée approximative de la position relative des grandes aires d'activité philosophique liées à la cognition, d'autant qu'elles occupent (sous des dénominations diverses, on y reviendra) des bataillons de philosophes plus nombreux que n'importe quelle autre branche de la philosophie des sciences, et dont la production, en diversité et en quantité, défie littéralement l'entendement.

Cette géographie sera cependant esquissée seulement à la fin du présent chapitre, car il vaut mieux se faire d'abord une idée un tant soit peu précise de ce qui se fait effectivement dans le domaine. Disons seulement que ce chapitre sera essentiellement consacré à des questions relevant sans ambiguïté et de manière spécifique de la philosophie des sciences cognitives (des questions qui sont dans le même rapport aux sciences cognitives que, par exemple, des questions classiques de philosophie de la biologie, telles que la notion d'organisme, le concept de fonction ou la réduction moléculaire, à la biologie), et ne fera qu'évoquer des problèmes philosophiques plus généraux que ces sciences soulèvent. Il est à peine utile de préciser qu'il ne s'agira pas de « faire le tour » de la philosophie des sciences cognitives : il s'agira d'un échantillon qu'on voudrait représentatif.

On ne trouvera pas non plus ici une mini-encyclopédie des sciences cognitives. Ce n'est pas ce qu'on attend d'un chapitre de philosophie des mathématiques, ou de philosophie de la biologie. Les sciences cognitives ont beau être jeunes, elles n'en sont pas moins dotées aujourd'hui d'une vaste bibliothèque d'ouvrages introductifs ou avancés, généralistes ou spécialisés, qui dispense le philosophe des sciences cognitives du rôle (qu'il s'est parfois senti l'obligation d'assumer au début) de vulgarisateur et d'historien.

1. *La structure de l'esprit : un programme de recherche*

1.1 De Gall à Fodor

1.1.1 *L'idée même d'une architecture de l'esprit et le projet d'une psychologie des facultés*

Notre point de départ est une question qui taraude les sciences cognitives depuis qu'un philosophe, Jerry Fodor, l'un des principaux théoriciens du domaine, formula

il y a un quart de siècle l'hypothèse d'une « architecture » modulaire de l'esprit (Fodor, 1983). L'intuition initiale, dont on attribue la première formulation scientifique à Franz Gall, au début du XIX^e siècle, est simple : l'esprit serait une collection de facultés spécialisées. L'idée de Gall le conduisit, avec l'aide de son disciple Spurzheim, à ce qu'on considère aujourd'hui comme un épisode calamiteux de pseudo-science, la phrénologie ou théorie des « bosses du crâne » : l'aptitude aux mathématiques, par exemple, était expliquée par le sur-développement d'une aire spécialisée du cortex cérébral, causant à l'endroit correspondant de la boîte crânienne une déformation anatomique qui méritait le nom de « bosse des maths » ; et ainsi de suite pour toute une série de « facultés » (27 exactement, dont 19 partagées avec les animaux, et 8 propres à l'homme) suggérées par une psychologie largement spéculative et entachée des préjugés anthropologiques de l'époque (Gall & Spurzheim, 1810-1819).

Ce qui doit nous intéresser aujourd'hui, ce ne sont pas les erreurs de Gall et Spurzheim, mais le grain de vérité sur lequel ils avaient peut-être mis le doigt. En réalité, on peut rétrospectivement leur attribuer la formulation d'un programme de recherche dans lequel s'inscrivent une bonne partie des sciences cognitives contemporaines. Ce programme s'articule autour de trois grandes questions :

(1) Sachant que l'esprit humain est capable d'accomplir des tâches d'une variété et d'une complexité considérables, est-il composé de parties, et quelles sont-elles?

(2) Si l'on admet que l'esprit est produit par un système dédié de notre organisme tel que le cerveau ou, plus précisément, le système nerveux central (SNC), quels sont les rapports entre, d'une part, l'esprit (vu comme l'ensemble des fonctions mentales ou psychiques) et le SNC et, d'autre part, les facultés (les composantes de l'esprit) et les parties du SNC ?

(3) S'il se confirme que l'esprit est composé de parties, correspondant à une composition en parties du SNC, comment s'explique la capacité, réelle ou apparente, de l'esprit à faire face à une variété indéfinie de situations qui ne peuvent chacune relever de la seule compétence d'une faculté fondamentale, et de manière concomitante, comment peut-on rendre compte du sentiment introspectif de l'unité essentielle de l'esprit ?

Dans leur généralité, on pourrait craindre que ces questions ne se révèlent à la réflexion rhétoriques ou excessivement vagues, ou bien, mises à l'épreuve de l'enquête scientifique, stériles. Nous allons voir que ce n'est pas le cas, mais auparavant il est utile de s'interroger sur le cadre dans lequel ces questions peuvent prendre sens.

1.1.2 *L'esprit ouvrier*

La première nous invite à considérer l'esprit, au premier chef, comme une entité accomplissant des tâches. La machine à coudre coud, le soc fend les mottes, le cœur fait circuler le sang dans le corps, l'abeille récolte le miel, l'élève multiplie 13 par 17 et l'esprit, de même, vaque à de nombreuses tâches. Pourtant, quoi qu'on puisse

entendre exactement par « esprit », et sans lui attribuer des qualités mystérieuses[1] en considération desquelles il faudrait rendre le mot français par l'anglais « *spirit* » et non pas, comme on le fait dans le présent contexte, par « *mind* », on ne saurait dire que l'esprit se présente à nous sous ce jour. Il se présente plutôt comme un « flux » (flux mental, flux de pensées) et comme le siège de la conscience, ou bien encore comme un œil interne, ou ce que le philosophe Daniel Dennett appelle (avec dérision) le « théâtre cartésien ». William James disait de la psychologie qu'elle avait pour objet la « *conscious mental life* » (« la vie mentale consciente »).

Cette observation élémentaire appelle à son tour plusieurs remarques. La première est que les deux conceptions ne sont antinomiques qu'en tant qu'elles prétendent saisir l'essence ou le cœur de la notion d'esprit. En revanche, on peut subordonner l'une à l'autre : l'esprit comme flux conscient peut être mis au service d'une tâche, comme lorsque l'esprit de l'élève (et non son foie ou ses jambes) est mis à contribution pour déterminer le produit de 13 par 17 ; inversement, on peut facilement imaginer que l'esprit vu comme potentialité d'accomplissement de tâches (nous proposerons bientôt une expression moins gauche) donne lieu à des phénomènes secondaires se manifestant dans notre expérience personnelle sous la forme de « flux » de pensées conscientes ou d'un « théâtre intérieur » où se succèdent des « apparitions ». Cependant, et c'est la deuxième remarque, la conception « accomplissement de tâches » semble à première vue plus restrictive, et correspondre aux épisodes purement délibératifs de notre vie mentale : en faire le cœur de l'esprit est prendre une option forte, qui n'est pas sans rappeler d'autres moments dans l'émergence d'une science, tels que la conception galiléo-cartésienne du mouvement inaugurant une science « pauvre » de la dynamique dégagée de la conception « riche » du mouvement hérité d'Aristote. Une telle option jouit d'une légitimité initiale, à titre de conjecture ou de pari, et gagne en crédibilité à mesure que se développe, à partir d'elle, un programme de recherche fécond ou progressif. En troisième lieu, il faut s'attendre (comme dans le cas du mouvement en physique) à ce que le sens en lequel l'esprit accomplit des tâches subisse des modifications considérables. Initialement, des exemples caractéristiques de tâches sont la résolution d'un problème formel simple, la détermination de la cause ou de l'agent responsable d'un événement courant, la traduction d'un texte simple, la planification d'une action ; et les voies typiquement suivies par l'esprit pour accomplir ces tâches relèvent de la logique (entendue en un sens suffisamment large). Mais les sciences cognitives ne sont nullement tenues de se conformer ou de se limiter à ce paradigme ; nous verrons, de fait, qu'elles s'en sont affranchies. Ce double mouvement de restriction puis d'affranchissement des conceptions de sens commun, ou d'un héritage métaphysique, est à l'œuvre dans la genèse de toute science, et c'est une banalité. Dans le cas des sciences cognitives, à cause de leur

1. Je n'ai pas dit : « imaginaires » ! Il ne s'agit pas d'éliminer le *spirit*, mais de délimiter, dans la mesure du possible, un domaine d'investigation, et *mind* fournit un périmètre déjà fort large.

jeunesse et de la porosité des frontières qu'elles partagent avec la philosophie et avec le sens commun, ce geste de constitution de son objet doit être souligné et rappelé autant de fois que nécessaire, car en dehors des sciences cognitives, il est mal compris, et donne lieu à des contestations qui sont le plus souvent des malentendus. Enfin, c'est la dernière remarque, le philosophe, sans en contester la légitimité en tant que conjecture ou pari, ne doit pas accepter cette option sans examen. On s'en rend compte d'autant mieux qu'elle est précisément mise en question aujourd'hui, non pas de l'extérieur, mais de l'intérieur, par des scientifiques et des philosophes qui estiment que les sciences cognitives doivent faire éclater, d'une manière ou d'une autre, le cadre conceptuel dans lequel elles ont pris leur essor (il en est brièvement question en 2.2.4 *infra*).

1.1.3 *Le cerveau et l'esprit*

Passons à la deuxième question issue de la problématique de Gall. Elle reposait pour Gall déjà (comme pour ses prédécesseurs et ses contemporains matérialistes) sur l'idée que les productions de l'esprit sont, en un sens, également des productions du cerveau. Quel peut être ce sens ? Les médecins, et les philosophes derrière eux, se sont longtemps satisfaits de la métaphore du « siège » : le cerveau est le siège de la pensée. Cela signifiait que sans cerveau, la pensée est impossible, et qu'une lésion du cerveau conduit généralement à une altération de la pensée. Il était néanmoins clair que le cerveau ne « produit », au sens causal, que des événements ou épisodes cérébraux, de nature biologique, électrique et chimique, susceptibles de déclencher à leur tour des événements moteurs. Mais la pensée (les productions ou manifestations caractéristiques de l'esprit) n'est de nature ni biologique, ni chimique, ni électrique, ni motrice… On reconnaît là l'une des formes du problème corps-esprit, auquel les philosophes et les premiers représentants de la psychologie scientifique s'efforçaient d'apporter une solution. Or l'idée de Gall semblait promettre non pas une solution, mais un contournement de ce problème dont aucune solution proposée ne semblait susceptible de rallier l'opinion. Cela peut paraître surprenant, puisque la notion de correspondance d'une faculté particulière de l'esprit avec une aire spécifique du cerveau semble dépendre logiquement de la notion de correspondance entre l'esprit (dans sa totalité) et le cerveau (entier). Comment comprendre ce que signifie que telle partie du cortex produit la pensée mathématique tant que l'on ne comprend pas ce que peut vouloir signifier que le cerveau produit la pensée ? Mais voici comment on peut espérer surmonter la difficulté. Supposons que nous réussissions à montrer (i) que tout processus mental est un « geste » élémentaire relevant d'une faculté particulière, ou une combinaison réglée de tels gestes, ou encore une combinaison réglée de gestes relevant de diverses facultés ; (ii) qu'à chaque faculté correspond une zone dédiée du cerveau ; (iii) qu'à chaque combinaison de processus mentaux élémentaires correspond une transformation spécifique du substrat cérébral. Alors on pourrait considérer que (iv) il existe entre les pensées, l'ensemble des productions

de l'esprit, d'une part, et les états et transformations du cerveau, de l'autre, une sorte d'isomorphisme et que (v) sur le plan strictement scientifique, cette correspondance empirique suffit pour les besoins de l'explication et de la prédiction, rendant superflues les conceptions métaphysiques irrémédiablement diverses qui sont et seront proposées pour rendre raison de cette correspondance. Remarquons la parenté entre cette manière de traiter par les moyens de la science un problème métaphysique avec la solution proposée par le réalisme structural à la question générale du réalisme scientifique : en suivant une piste ouverte par Poincaré (et dans une certaine mesure anticipée par Comte – voir par exemple Comte, 1948 –), les partisans contemporains du réalisme structural tels que John Vorrall (1989) estiment que la science ne peut identifier que le système des relations entre les entités du monde, et qu'elle doit renoncer à déterminer la nature profonde ou l'essence des entités elles-mêmes. On peut parler d'un « structuralisme » inhérent à une neuropsychologie des facultés telle que l'ébauche Gall, et qui trouvera une expression à la fois plus générale et plus précise, comme nous allons le voir, dans la conception fonctionnaliste qui demeure le cadre de référence des sciences cognitives.

Mais en même temps, cette esquisse de solution, ou de dissolution modulariste du problème corps-esprit, est peut-être une victoire à la Pyrrhus : car si l'esprit n'est manifestement contenu dans aucune fonction ou faculté suffisamment restreinte pour être « mise en correspondance » avec une aire du cerveau (qui peut raisonnablement penser que tout ce que l'esprit accomplit se laisse distribuer dans un nombre fini raisonnable de catégories ?), qu'est-ce qui nous permet de considérer qu'il est contenu dans leur réunion ? L'esprit ne serait-il pas précisément ce qui échappe à la spécialisation ? Ou encore, ce qui mobilise à bon escient les facultés spécialisées ? Nous en arrivons ainsi à notre troisième question (p. 521). Elle peut conduire à trois attitudes : ou bien l'on s'en tiendra à l'idée d'une combinatoire de processus spécialisés, en soulignant qu'une combinatoire peut précisément engendrer une variété infinie de pensées hybrides (mêlant plusieurs composantes spécialisées) – mais alors il faudra pouvoir expliquer ce qui reste de la modularité si l'on autorise toute combinaison entre les productions des différents modules – ; ou bien l'on admettra qu'une partie de la pensée échappe à la modularité, fût-elle enrichie par un jeu de combinaisons permises ; ou bien enfin on estimera le problème suffisamment grave pour revenir sur les hypothèses cadres sur lesquelles on s'est appuyé jusqu'ici pour donner sens aux questions que pose le programme de Gall.

Voilà donc déjà toute une série d'interrogations que l'on peut rétrospectivement poser à propos du projet gallien d'une psychologie des facultés ou, en termes contemporains, d'une conception modulariste de l'architecture fonctionnelle de l'esprit, sans avoir même commencé à déployer les concepts fondamentaux des sciences cognitives. Nous allons prendre conscience graduellement au cours de ce chapitre combien la problématique gagne en précision, et en contenu assignable, grâce à ces concepts.

1.1.4 *Les deux étages de l'esprit selon Fodor*

Revenons donc à Fodor. L'esprit, selon lui[1], serait constitué de deux sortes de processus : d'un côté, des facultés autonomes spécialisées, appelées « systèmes d'entrée » (*input systems*) ; de l'autre, des « systèmes centraux » assurant la « fixation des croyances », c'est-à-dire l'aboutissement des processus cognitifs sous la forme d'une adhésion consciente à une proposition telle que « Un livre rouge est posé sur la table ». Il existe, en réalité, une grande variété d'états mentaux conscients caractérisés par une « attitude propositionnelle » : acceptation (éventuellement graduée), doute ou rejet, crainte ou espoir... d'un état de fait, réel ou supposé, lequel est exprimé dans un langage, par exemple – nous y reviendrons – notre langue maternelle. Les processus centraux postulés par Fodor conduisent l'esprit à un état de ce type, sur la base de données fournies par les systèmes d'entrée (dont la fonction est, selon Fodor, de « présenter le monde à la pensée » ; ils comprennent en effet les processus perceptifs, ainsi que le langage, du moins la ou les composantes automatiques du traitement et de la production du langage parlé). Ceux-ci sont locaux, spécialisés, ne traitant que certains types d'informations ; ils sont « isolés » au sens où, par construction, ils ne peuvent exploiter d'informations extérieures à leur base propre ; ils sont automatiques et rapides ; ils présentent des profils caractéristiques d'apprentissage et de dégradation en cas de lésion ou d'affection ; ils sont au moins approximativement localisés dans le cerveau et ont une dimension innée. Ces propriétés rendent les modules accessibles à l'enquête scientifique, et de fait les sciences cognitives progressent dans la théorisation des processus cognitifs modulaires. Au contraire, la science rencontre des obstacles dirimants lorsqu'elle aborde les processus centraux. Selon Fodor, les sciences cognitives n'ont fait aucun progrès dans ce domaine, et il prédisait à l'époque qu'elles n'en feraient pas (il n'est pas plus optimiste aujourd'hui : voir Fodor, 2000). L'argument repose sur une comparaison avec la théorie de la confirmation scientifique : d'une part, rien ne limite ce qui en droit doit être pris en compte pour déterminer la valeur de vérité d'une croyance ; d'autre part, toute croyance s'insère dans un système de croyances, dont le degré de confirmation ne peut s'évaluer que collectivement.

Fodor propose ainsi des réponses aux questions que posait la théorie de Gall, réponses qui appelleront de nouvelles questions dont certaines seront abordées dans un instant :

---

1. Fodor n'a pas inventé, ni même réinventé à lui seul dans le contexte contemporain la notion et l'hypothèse modularistes. Il en a fait la théorie systématique, mobilisant les ressources des sciences cognitives et de l'analyse conceptuelle, et s'est risqué à proposer une explication du bilan contrasté des sciences cognitives, allant jusqu'à leur assigner une limite de principe. Je le précise pour deux raisons : d'une part, ce chapitre ne vise pas à l'exactitude historique, et les noms cités ne le sont qu'à titre de grands repères ; d'autre part, la contribution de Fodor à la question de la modularité est un exemple caractéristique de « philosophie cognitive », au sens qui sera précisé dans la conclusion.

(1F) Oui, l'esprit est composé de parties, et nous avons une idée relativement précise de ce que sont ces parties et comment elles se caractérisent. Cependant, cette division en parties ne concerne qu'un secteur de l'activité mentale, laissant échapper une province importante du mental. (Bien entendu, les modules conjecturés par Fodor n'ont pratiquement aucun rapport avec les vingt-sept facultés de Gall ; la notion même de faculté, qui recouvre chez ce dernier aussi bien des instincts et des traits de caractère que des talents intellectuels particuliers ou différentes formes de mémoire, revêt chez Fodor un sens précis, qui s'articule avec les autres postulats de sa psychologie[1].)

(2F) Les parties de l'esprit identifiées par Fodor, qu'elles soient ou non modulaires, sont décrites comme des systèmes de traitement de l'information. On peut concevoir (mais il faut le rendre explicite, ce que fait Fodor dans la première partie du livre dans lequel il rappelle le cadre général que se sont donné les sciences cognitives depuis leur naissance, nous y revenons au § 2) que le cerveau soit le système matériel qui exécute ce traitement, et que les modules de l'esprit soient associés à (aient pour siège, ou pour « substrat neural » comme on tend à dire aujourd'hui) des sous-systèmes du cerveau dédiés à l'exécution des tâches spécialisées qui échoient au module correspondant.

(3F) La capacité de l'esprit à faire face à une variété indéfinie de situations dont la plupart ne peuvent logiquement pas relever d'une faculté particulière est un mystère que les sciences cognitives ne sont pas prêtes d'expliquer.

1.2 L'idée d'intelligence générale et ses difficultés

Quand Fodor publie son livre, l'un des plus influents dans l'histoire des sciences cognitives, il prend à contre-pied l'une des principales intuitions qui avaient présidé à la première phase de cette histoire, tout en s'inscrivant, sur un autre plan, dans le droit fil de cette tradition de recherche. Dans un article fondateur paru en 1950, le logicien Alan Turing, l'inventeur du concept abstrait d'ordinateur, défendait l'hypothèse que certaines machines pourraient être capables de « penser », c'est-à-dire d'accomplir toutes les tâches que l'homme doit à son intelligence de pouvoir accomplir. Précisé et amplifié par Herbert Simon, Alan Newell et d'autres (Newell & Simon, 1972[2]), ce projet prit bientôt le nom d'« intelligence artificielle » (IA) et constitua

1. Les modules de Fodor se distinguent plus généralement des composantes qu'a recherchées tout au long du XIX^e siècle la « psychologie des facultés » : celles-ci étaient « horizontales », c'est-à-dire qu'elles désignaient des « opérations », telles que l'attention, la mémoire, l'observation, la précision, la rapidité, la discrimination sensorielle, etc., applicables à tous les domaines ; les modules de Fodor sont, au contraire, « verticaux » : chacun a une compétence limitée qui n'empiète pas sur celle des autres. La psychologie des facultés, qui avait des conséquences importantes en matière de pédagogie, a été définitivement discréditée au début du XX^e siècle (Thorndike & Woodworth, 1901).

2. Cette date de publication est trompeuse : la naissance de l'IA se situe vers le milieu des années 1950 (voir Buchanan, 2005 ; McCorduck, 2004 ; Bowden, 1953 ; Hook, 1960).

(avant la lettre) la première grande figure des sciences cognitives[1]. Ce que Fodor reprend du cadre de l'IA, et qu'il contribuera d'ailleurs à préciser, c'est l'idée que les processus mentaux sont essentiellement des transformations réglées d'informations. Ce que Fodor rejette en revanche, c'est la conséquence que l'IA a tirée de la découverte, pourtant très frappante, d'un fait de nature essentiellement logique, à savoir l'existence d'une « machine de Turing » (un calculateur symbolique) possédant la propriété d'*universalité* : une telle machine est capable de calculer, à partir du schéma de construction (techniquement : de la table) de n'importe quelle autre machine de Turing, ce que cette machine calcule (Turing, 1937). Ainsi le néomécanisme turingien semble-t-il capable de surmonter la limitation essentielle du concept classique de mécanisme, qui est de ne pouvoir rendre raison que de machines *dédiées* : une tâche, une machine[2]. Une machine de Turing universelle (MTU) accomplit, dans le domaine qui est le sien (le traitement de l'information), toute tâche concevable[3]. Notre troisième question recevait ainsi une réponse satisfaisante : si notre esprit possède les fonctionnalités d'une MTU, alors on s'explique qu'il puisse accomplir n'importe quelle tâche cognitive, et dans la mesure où il est « réalisé » dans cet organe à nous qu'est le cerveau, on peut comprendre le sentiment que nous avons d'une unité de l'esprit, un peu à la façon dont nous comprenons intuitivement que notre main puisse exécuter, dans certaines limites, tout geste manuel concevable.

Pourquoi Fodor et les partisans de la modularité renoncent-ils à cette solution ? Pour deux raisons principales. La première est l'argument de l'explosion combinatoire : le nombre d'opérations à effectuer, lors d'une tâche cognitive, est une fonction exponentielle du nombre d'informations susceptibles d'être pertinentes. Si ce dernier est très grand, les opérations nécessaires « explosent » et dépassent les capacités nécessairement finies de tout système matériel. Une « intelligence » ou système cognitif universel aurait par définition affaire à une base de données d'une taille quasiment infinie, ce qui l'empêcherait d'exécuter la plupart de ses tâches, en tout cas dans un délai raisonnable (l'exemple favori des modularistes est celui du tigre : face à un signe de présence probable d'un tigre, tel qu'une perception visuelle

1. Dans le présent contexte, on assimile volontiers « intelligence » à « esprit » (ou du moins à « capacités cognitives »), et on peut corrélativement voir dans l'intelligence artificielle un modèle abstrait de l'intelligence humaine. Il y a là un ensemble de décisions pour partie terminologiques, pour partie doctrinales, qui seront abordées plus loin dans le chapitre. Il existe un autre usage du mot « intelligence », commandant un autre concept d'intelligence générale, lié à la question de la comparaison et de la mesure qualitative de degrés d'intelligence ou de qualité des performances cognitives. C'est là un autre domaine, celui du QI, qui ne recoupe que partiellement, dans l'état actuel des connaissances, celui des sciences cognitives, même si à terme la question du QI devrait s'y intégrer pleinement. L'intelligence au sens du QI pose des problèmes de philosophie des sciences du plus haut intérêt (voir par exemple Sternberg, 1988 ; Flynn, 2007 ; Nisbett, 2009) qui ne pourront être abordés ici.
2. Rappelons que pour Aristote, c'est parce que l'esprit peut recevoir toutes les formes possibles (c'est-à-dire penser n'importe quel objet) qu'il ne saurait être matériel (*De anima*, III, 4 ; 429a10-b9) (voir Robinson, 2007).
3. Quelle que soit sa signification exacte pour les sciences cognitives, la portée conceptuelle générale de la notion de MTU est considérable (Herken, 1988).

ayant l'apparence d'un tigre, il est crucial de pouvoir prendre une décision rapide). L'hypothèse de la modularité, en limitant drastiquement, pour certaines familles de tâches, la base de données, les rend matériellement faisables dans un système matériel de traitement de l'information.

La seconde raison de renoncer au modèle de la MTU est l'argument dit de la pauvreté du stimulus. Le premier cas de modularité a été défendu par Chomsky (Chomsky, 1957 ; Piatelli-Palmarini, 1979) : l'apprentissage de la langue maternelle est une tâche particulièrement importante et complexe qu'accomplissent sans faillir tous les enfants normaux de la terre. Si c'était, comme on a pu longtemps le penser, l'œuvre d'une capacité générale d'apprentissage appliquée à l'environnement linguistique du jeune enfant, ce succès serait (selon Chomsky) impossible, pour des raisons essentiellement logiques : ce que l'expérience fournit à l'enfant (le « stimulus[1] ») serait, affirme-t-il, beaucoup trop ténu (« pauvre ») pour lui permettre d'identifier la « grammaire » de sa langue, c'est-à-dire l'ensemble articulé des connaissances (tacites) qui lui permettent de comprendre et de parler. L'induction en vertu de laquelle l'enfant passe des informations que lui fournit son environnement à la maîtrise de la grammaire (en ce sens étendu, qui va bien au-delà de la grammaire traditionnelle) ne peut réussir que dans un cadre contraint, comparable au chemin développemental suivi par un organe ou un membre d'animal. Le « système d'acquisition du langage » serait donc un module essentiellement indépendant de facultés générales de l'esprit. Les arguments de l'école chomskyienne, qui restent à ce jour contestés mais conservent non seulement, aux yeux de cette école, leur validité, mais également, pour les sceptiques et les adversaires déclarés, un défi, sont de nature à la fois linguistique, logique, psychologique, physiologique et, plus largement, biologique. Plus encore, le cas du langage a valeur paradigmatique pour l'ensemble des processus cognitifs : le modèle chomskyien, on vient de le voir avec Fodor, s'étend à d'autres aptitudes cognitives et soulève, *mutatis mutandis*, la même série de questions, à la clarification desquelles les philosophes ont très largement contribué. Nous allons à présent en examiner quelques aspects.

1.3 Développement et innéisme

1.3.1 *Le mystère de l'*infans

Depuis Platon, les philosophes s'interrogent sur l'origine de nos connaissances. L'*infans,* celui qui ne parle pas (et qui, a-t-il longtemps semblé, pense, s'il est possible, encore moins), se développe physiquement et mentalement. Mais alors que l'on peut observer, à l'œil nu, bien des aspects de la transformation du corps, en ayant l'impression de les comprendre, ce qu'on observe de la transformation de l'esprit

1. La terminologie provient de la psychologie béhavioriste, dont la théorie du langage a suscité de la part de Chomsky une critique souvent jugée comme décisive (voir sa recension de l'ouvrage *Verbal Behavior* de B.F. Skinner : Chomsky, 1959).

nous plonge dans la perplexité. Si l'idée de croissance, à partir d'Aristote, constitue un socle d'évidence qui nous rassure et qui assoit conjointement une conception de sens commun et un programme de recherche en biologie largement couronné de succès, nous restons dans une profonde incertitude s'agissant du développement mental.

Que ce mystère ait longtemps été pratiquement ignoré, relégué en tout cas loin derrière les problèmes de l'origine du cosmos, de la nature de la matière, ou de l'essence de la vie constitue en soi un mystère philosophique. J'y vois, pour ma part, l'effet d'un renoncement rationnel, à l'image de la parabole des raisins trop verts : autant l'on a très tôt trouvé des prises pour aborder ces trois derniers problèmes, et qu'ils sont aujourd'hui, sinon pleinement résolus, du moins profondément attaqués, jusqu'à tout récemment le premier mystère a paru offrir à notre regard une paroi verticale parfaitement lisse. Nous sommes restés paralysés, pris en tenaille entre une conception naturaliste et organique du développement mental (l'enfant croît mentalement *comme* il croît physiquement) et une métaphore scripturale de l'esprit, selon laquelle il reçoit des inscriptions qui l'informent progressivement et le mettent en état d'effectuer les opérations qui caractérisent la cognition adulte. Former l'esprit, c'est l'informer (lui fournir ce qu'au XVII^e^ siècle on appelait des idées, qu'on appellera plus tard des représentations). Ces inscriptions sont ou bien présentes (en totalité ou en partie) dès la naissance, comme le croit l'innéisme (parfois également appelé, dans ce contexte, rationalisme), ou bien, comme le soutient l'empirisme, proviennent intégralement de l'expérience, à partir des premiers jours de la vie. Pour un camp (où se range Descartes) comme pour l'autre (avec Locke), l'esprit est sans structure (sans « architecture » au sens expliqué plus haut) : il est un récipient essentiellement passif, doué seulement, contrairement à tous les autres systèmes naturels, d'une aptitude à se laisser « impressionner » d'une infinité de façons, aptitude caractérisée comme apprentissage ou mémoire. L'enfant se développe mentalement parce qu'il acquiert des connaissances, de même qu'il se développe physiquement parce qu'il acquiert de la matière organique, du muscle, de l'os, d'autres tissus, qui viennent seulement renforcer des structures déjà présentes (dans l'ensemble, les organes et segments visibles du corps adulte sont présents dans le corps du nouveau-né).

1.3.2 *L'idée moderne de développement*

Les fondateurs de la conception moderne du développement cognitif (Piaget, Vygotsky, Chomsky, Bruner, Carey…), s'ils s'opposent fortement sur certaines questions centrales, ont en commun d'avoir su se déprendre de ces conceptions traditionnelles, tout en en conservant certains éléments :

(i) Ils ont admis la possibilité que l'architecture de l'esprit soit complexe et différenciée.

(ii) Ils ont admis que cette architecture puisse varier au cours du développement.

(iii) Ils ont admis que l'évolution des capacités cognitives de l'enfant résulte conjointement d'un développement organique de l'architecture de l'esprit et de la modification (par acquisition et révision) des connaissances (idées, représentations, croyances…) qu'il détient, étant entendu que ces connaissances n'ont pas nécessairement (et n'ont de fait en général pas) le caractère explicite et conscient des connaissances de l'adulte en situation de délibération (dont le scientifique au travail est le paradigme).

Ce qui est conservé, c'est l'idée que l'acquisition des connaissances (en un sens qui s'éloigne progressivement à la fois du sens habituel et des conceptions développées au XVII^e siècle) joue un rôle dans l'épigenèse des capacités cognitives, et en particulier qu'elles peuvent être ou bien innées, c'est-à-dire présentes dès l'origine (il s'agit alors d'une acquisition de l'espèce, plutôt que de l'individu), ou bien acquises au cours du développement individuel. Ce qui est rejeté, c'est l'axiome de l'homogénéité, ou indifférenciation initiale de l'esprit, l'axiome d'invariance structurelle ou organique au cours du développement, et enfin l'idée que le développement cognitif est exclusivement imputable à l'accumulation des connaissances. Désormais, la problématique du développement s'articule en trois moments : l'état initial, la transition ou développement, l'état final, l'accent étant mis sur les caractéristiques invariantes d'un individu à l'autre, et la recherche portant sur la distinction et les interactions entre les processus de changement structurel (parfois appelés maturation) et les processus d'acquisition des connaissances (parfois appelés apprentissage).

Ces hypothèses sont essentiellement indépendantes, mais leur conjonction forme un cadre théorique cohérent et jugé productif par beaucoup de chercheurs. Aucune d'elles n'a l'évidence de son côté. Au contraire, elles sont toutes hasardeuses et comportent une part d'obscurité, qu'il va falloir essayer de réduire, en mêlant l'enquête empirique et l'analyse conceptuelle. On a rapidement évoqué la difficulté inhérente à l'hypothèse d'une architecture de l'esprit (et on va y revenir). Tant que cette notion n'a pas été pleinement clarifiée, celle d'une évolution de l'architecture est également frappée d'obscurité. Provisoirement, on peut se contenter de l'idée gallienne d'une structuration fonctionnelle calquée sur une topographie anatomique, mais cette stratégie, on le verra, soulève des objections. Ces difficultés obèrent également la distinction entre maturation et apprentissage, ou entre évolution de l'architecture et acquisition des connaissances. Nous verrons néanmoins qu'il existe des manières de les lever sur le plan théorique, la nouvelle question étant alors celle de l'adéquation globale du cadre proposé avec l'ensemble des données empiriques.

1.3.3 *Qu'est-ce qu'une capacité innée ?*

Mais l'hypothèse qui appelle une clarification de manière peut-être la plus urgente est celle de capacité (connaissance ou aptitude) *innée*. L'innéisme joue un rôle crucial dans les sciences cognitives, car de nombreux courants de recherche concluent au caractère inné non seulement du langage, mais aussi d'autres facultés ou (pour

employer le terme le plus inclusif possible) de structures cognitives, conclusions qui sont contestées par d'autres courants. Des réponses inspirées de l'anatomie et de la physiologie viennent spontanément à l'esprit. Mais même dans ce domaine, le caractère inné de certaines structures ou traits organiques soulève une série de questions centrales pour la philosophie de la biologie. S'agissant de fonctions mentales, le problème est encore plus difficile, et l'issue des débats en cours est incertaine.

La première observation est que la définition la plus naturelle de l'inné est privative : l'inné est ce qui n'est pas acquis, que ce soit pour des raisons empiriques ou pour des raisons conceptuelles. On peut concevoir, en effet, que certains concepts ou capacités pourraient être acquis, mais qu'en fait ils ne le sont pas ; d'autres, au contraire, peuvent sembler difficiles, voire impossibles à acquérir.

Mais en quoi consiste la possession d'une structure cognitive innée ? La réponse dépend-elle de la structure en question ? La capacité de sourire, de déglutir, de cligner des paupières est innée : il s'agit de réflexes moteurs. La capacité de servir au tennis est acquise : c'est une habileté qu'on apprend peu à peu par imitation intelligente. Mais comment comprendre que le concept de temps ou que le concept d'objet solide sont innés, alors que le concept de mariage morganatique ou celui de société à responsabilité limitée sont acquis ?

D'autre part, qu'entend-on au juste en excluant l'acquisition ? Veut-on dire que l'environnement ne joue aucun rôle ? C'est évidemment trop demander : bon nombre de traits anatomiques et fonctionnels de l'organisme adulte dépendent de l'environnement pour se développer, et très souvent aussi pour prendre une forme spécifique parmi plusieurs possibles. On peut du moins parler, comme le propose le philosophe de la biologie Paul Griffiths (Griffiths, 2002), d'invariance développementale, ce qui signifie que la structure en question émerge au cours du développement indépendamment des différences environnementales, dans les limites d'un large spectre d'environnements naturels.

Ou bien veut-on dire que la structure en question reste essentiellement la même au cours de la vie de l'organisme (tels le sexe [chez l'homme, hors intervention humaine et abstraction faite de certaines formes d'hermaphroditisme], la couleur des yeux ou le nombre de doigts) ? C'est une autre propriété que la précédente. Ou encore, troisième possibilité, qu'elle est présente à la naissance ?

Une deuxième observation est que l'inné n'est pas, comme le montre l'exemple du sexe, ce qui est propre à l'espèce. Pour autant, le concept d'innéité, et son usage, le rapprochent des idées conjointes d'hérédité et d'universalité au sein d'une espèce – en d'autres termes, par « inné », on entendrait souvent ce qui est « codé » dans le patrimoine génétique de l'espèce. C'est certainement ce que beaucoup entendent lorsqu'ils affirment, par exemple, que le langage est « le propre de l'homme » ou que l'on constate *a contrario* que certaines espèces animales (mais pas toutes) possèdent des capacités numériques élémentaires ou sont capables de conduites altruistes. Une difficulté propre à cette conception est que la notion de codage par ou dans le

patrimoine génétique donne lieu à des difficultés bien connues par les philosophes de la biologie.

Une troisième observation est que l'inné semble matérialiser une norme propre à l'espèce : est inné ce qui normalement conduit à un trait universellement partagé par les membres normaux de l'espèce. Les seins féminins sont innés en ce sens, sans d'ailleurs être présents à la naissance. Il en va de même d'innombrables systèmes métaboliques, de structures cérébrales, etc. Ces traits sont normatifs aussi en ce qu'ils sont fonctionnels, donc résultent probablement, directement ou indirectement, de la sélection naturelle.

Indépendamment des questions que soulèvent ces caractérisations, prises une à une, on peut se demander si elles sont, conceptuellement ou empiriquement, co-extensives ou si du moins elles coïncident largement. Sur le plan conceptuel, en première analyse, la réponse est clairement négative : les définitions fondées sur la non-apprenabilité ou l'indépendance à l'égard de l'environnement, les définitions fondées sur le patrimoine génétique de l'espèce et l'universalité intraspécifique, les définitions fondées sur la normativité fonctionnelle et adaptative ne sont pas conceptuellement équivalentes. Et de fait, en se plaçant sur le plan empirique, les biologistes ont exhumé quantité de contre-exemples à la thèse d'une coïncidence même approximative. Certains auteurs en sont venus à recommander l'abandon pur et simple de la notion. D'autres préconisent un emploi différencié selon les contextes et les fins théoriques (une solution souvent préconisée, par exemple, pour le concept de gène). La plupart s'en tiennent toutefois à l'idée que ces différentes caractérisations renvoient à des propriétés qui sont de fait souvent associées, et qu'il est utile de considérer les structures qui les possèdent toutes. En d'autres termes, l'innéité serait une propriété « en grappe » (*cluster property*) faite de traits généralement associés, mais qui ne sont pas nécessairement tous présents ; on sait que la vie (en tant que propriété pour un système matériel d'être vivant) est souvent considérée aujourd'hui comme une propriété de ce genre.

Dans le cas de structures cognitives, on l'a dit, la difficulté est redoublée par l'incertitude quant à la nature de ces structures. Plutôt que d'essayer d'en parler de manière générale, revenons au cas précis du langage. Observons d'abord qu'un argument important invoqué par les innéistes est que la progression suivie dans l'acquisition du langage est largement indépendante de l'individu et de sa langue maternelle, qu'elle est rapide et n'exige aucun apprentissage volontaire. C'est là l'indice d'un développement organique, comparable à celui d'un organe ou d'un segment corporel. C'est aussi le signe que le rythme du processus serait imposé par la maturation plutôt que par l'acquisition d'informations (on peut présumer que celle-ci donnerait lieu à des variations importantes d'un individu et d'une langue à l'autre). Il est ensuite bien clair que ce qui est inné ne saurait être la langue particulière parlée par l'enfant : non seulement les enfants apprennent des langues différentes, mais *tout* enfant plongé dans un milieu linguistique donné apprend la langue de ce milieu, indépendamment de ses origines, exactement de la même façon

(mêmes étapes, même rythme, même résultat final) que tous les autres enfants. Ce qui est inné ne peut donc être que la capacité d'apprentissage de la langue, qui en vertu de l'argument de la pauvreté du stimulus est dédiée au langage, au sens de ne pouvoir servir à l'apprentissage d'autre chose ; Chomsky l'appelle selon le contexte *grammaire universelle* ou *mécanisme d'acquisition du langage* (*language acquisition device* [LAD]). L'apprentissage consiste alors en la détermination, sur la base des indices disponibles dans l'environnement linguistique, de la grammaire particulière de la langue ambiante. Dire que la grammaire universelle est innée reviendrait à dire, selon une interprétation discutée actuellement, qu'elle est une « primitive » cognitive. En d'autres termes, elle ne relève pas de la psychologie mais de la biologie. En ce sens, elle serait véritablement un organe (plus exactement, une structure fonctionnelle cérébrale) susceptible d'accueillir et de traiter des informations linguistiques et de produire *in fine* une structure informationnelle ou psychologique constituée de représentations engendrant par combinaison la totalité des phrases de la langue, c'est-à-dire de phrases acceptables aux oreilles de ses locuteurs.

La même série de questions se pose chaque fois qu'on fait l'hypothèse qu'une structure ou capacité cognitive est innée, où l'on a le plus souvent à l'esprit l'une ou l'autre des trois grandes familles de propriétés évoquées à l'instant. On pourra, par exemple, être amené à conjecturer que tel concept (celui de temps ou d'espace, celui de nombre entier, celui d'itération, celui d'objet matériel, celui de mouvement, celui de cause, celui de relation, celui de conséquence logique, voire celui de concept) est inné ; il faudra essayer alors de comprendre à quoi cela revient, c'est-à-dire de passer d'une propriété diagnostique (le concept n'est [apparemment] pas appris, voire pas apprenable) à une caractérisation intrinsèque (que signifie pour un concept d'être inné ?) (Samuels, 2002 ; Carruthers, Laurence & Stich, 2005 ; Khalidi, 2007).

1.3.4 *La question empirique : quelles capacités sont-elles innées ?*

Mais, à supposer que les incertitudes ontologiques quant au concept d'innéité soient levées, ou bien que l'on puisse s'accorder provisoirement sur une caractérisation opérationnelle du caractère inné d'une structure cognitive donnée, il reste encore à peser les arguments *pro* et *contra*. Dans le cas du langage, outre les propriétés indiquées ci-dessus, l'étude des enfants aveugles ou sourds de naissance, qui ne bénéficient pas de tout l'apport informationnel dont disposent les enfants entendants et voyants, renforce considérablement l'hypothèse innéiste. Dans le cas des concepts, c'est l'apparente impossibilité d'induire l'extension d'un concept à partir d'un échantillon d'instances qui motive l'innéisme (Fodor,1975, 1981). Les sceptiques quant à eux (Elman *et al.*, 1996 ; Cowie, 1999) contestent notamment l'argument de la pauvreté du stimulus : d'une part, le stimulus est moins pauvre que ce qui est affirmé, une partie de l'information nécessaire pouvant provenir de sources non linguistiques ; d'autre part, en admettant que l'identification de la grammaire exige des contraintes supplémentaires, ils contestent que ces contraintes doivent nécessairement prendre la forme de connaissances (tacites), conçues généralement

comme des règles ou des paramétrages de règles universelles. Ils doutent également que les myriades de régularités propres à chaque langue puissent être toutes déduites d'un nombre raisonnable de règles ou de paramètres. Des modèles connexionnistes (voir § 3.a), apparemment incompatibles avec les conceptions chomskyennes de la compétence linguistique, semblent montrer que les impossibilités inductives postulées par les innéistes résultent en fait d'un manque d'imagination de leur part : ne pas voir comment un système S pourrait apprendre X sur la base d'un certain ensemble d'informations n'implique pas que X soit inné chez S, mais seulement que le chercheur n'a pas trouvé de solution (qu'elle existe ou pas) (Elman *et al.*, 1996). C'est pour écarter ce genre d'objection qu'une théorie formelle (logique) de l'apprentissage a été développée ; elle permet de formuler des résultats d'impossibilité : sous certaines hypothèses, on démontre (mathématiquement) qu'un système S doté de telles et telles ressources ne peut identifier la grammaire d'une langue sur la base d'une information empirique présentant certaines caractéristiques (Jain *et al.*, 1999). Ces résultats doivent néanmoins être jaugés à l'aune de la pertinence des idéalisations initiales et de la plausibilité des hypothèses formelles, ce qui explique qu'ils n'aient pas mis fin au débat (Stainton, 2006, p. 57-112). Celui auquel donne lieu la question de l'innéité des concepts (parmi les sceptiques : Prinz, 2002 ; Laurence & Margolis, 2002) n'est pas davantage tranché.

1.4 L'idée même de base neurale

Revenons à la modularité (sans nous éloigner beaucoup de la question de l'innéité). Pour Gall, on l'a vu, les facultés ont des « sièges » distincts, qui sont autant d'aires délimitées du cerveau (en général, mais pas toujours, du cortex). Fodor est beaucoup plus prudent, considérant d'une part que les modules ne sont pas nécessairement localisés *anatomiquement*, et qu'ils peuvent l'être seulement fonctionnellement (correspondre donc à des modes de fonctionnement neurophysiologique qui ne se ramènent pas simplement à l'ensemble des activités d'une aire particulière), d'autre part que la localisation n'est pas strictement nécessaire à la modularité, en tout cas sur le plan conceptuel. Il n'en est pas moins vrai qu'une interprétation neurodynamique est une manière assez naturelle de préciser l'hypothèse modulariste. La neuropsychologie, issue des découvertes de neurologues tels que Broca (Broca, 1861) et Vernicke, se donnait pour objectif d'établir une correspondance entre déficits cognitifs et lésions cérébrales. L'existence de patients présentant des déficits très spécifiques a constitué le principal argument empirique en faveur de l'idée générale de différenciation du système nerveux central, dont la modularité est une formulation plus précise adaptée au cadre informationnel des sciences cognitives contemporaines.

La neuropsychologie a rejoint aujourd'hui les neurosciences cognitives, qui recherchent les « bases neurales » des fonctions cognitives chez l'être humain normal. La contribution spécifique de la neuropsychologie consiste à exploiter des

comparaisons entre tableaux cliniques pour formuler des hypothèses sur l'organisation cérébrale « responsable » de certaines fonctions cognitives. La situation caractéristique à cet égard est la « double dissociation » : un patient X présentant un déficit grave dans une capacité A (telle que l'identification d'artefacts courants – peigne, marteau, ciseaux, etc. – ; ou bien, autre exemple, la lecture de mots concrets) mais aucun dans une capacité B (telle que l'utilisation d'artefacts ; dans l'autre exemple, la lecture de mots abstraits), un patient Y présentant un déficit grave en B, aucun en A. Un tel couple de tableaux cliniques inspire au chercheur, en l'absence d'indications contraires, une hypothèse modulaire attribuant des bases neurales distinctes à A et à B. Bien entendu, il s'agit non d'une déduction, mais au mieux d'une inférence à la meilleure explication (ou abduction) : si les bases neurales de A et de B étaient effectivement localisées dans des composantes distinctes, alors cela expliquerait très directement que des tableaux cliniques tels que X et Y soient possibles. *A contrario*, le fait qu'on trouve invariablement associés deux déficits accrédite (sans l'établir fermement) l'hypothèse d'un large recouvrement des bases neurales de A et de B.

Cette démarche soulève toute une série de questions conceptuelles, méthodologiques et empiriques. Ainsi, on doit s'interroger sur la notion de différence, s'agissant de fonctions ou processus cognitifs. En un sens, toute différence compte : chacun admet que des processus cognitifs différents sont « pris en charge » par des circuits cérébraux qui diffèrent, ne serait-ce que légèrement (en vertu du principe de survenance selon lequel toute différence assignable au niveau mental implique une différence au niveau cérébral). En un autre sens, seules certaines différences présentent un intérêt théorique : autant nous aurions beaucoup à apprendre d'un lien de dépendance entre certaines fonctions apparemment distantes (par exemple, la navigation spatiale et la mémoire autobiographique, ou la perception de la direction du regard et la compréhension des mobiles d'autrui), ou inversement d'une autonomie mutuelle entre deux fonctions que le sens commun tend à confondre (prononciation des noms concrets et des noms abstraits), autant rien ne semble découler de la considération des liens entre la mémorisation des marques de voiture et la mémorisation des marques de lave-linge. Le double danger qui semblerait menacer la recherche de dissociations en neuropsychologie est donc la trivialité, d'une part, la fragmentation, d'autre part : les lésions cérébrales n'étant jamais « pures » (au sens de n'affecter exactement qu'un système fonctionnel), il est à craindre que des doubles dissociations finissent par être mises au jour pour des couples de processus ne présentant que des différences minimes sans portée théorique. En pratique, ce sont le bon sens clinique ainsi qu'un cadre théorique déjà esquissé, qui permettent d'éviter ces obstacles.

Mais d'autres difficultés surgissent. La manière la plus simple dont une base neurale peut se différencier d'une autre, on l'a dit, c'est spatialement. Au-delà, on peut imaginer des circuits distincts, mais pas nécessairement disjoints. Mais un troisième genre de rapport, bien plus exotique, est concevable. Les modèles connexionnistes, et plus généralement les modèles dérivés de la théorie des systèmes

dynamiques, prouvent que des fonctions distinctes peuvent être produites par un seul système complexe fonctionnant sous des régimes distincts. La conséquence de cette possibilité est de saper à la base l'intuition fondamentale de la modularité, qui est d'expliquer la structure de la pensée par l'organisation du système matériel dont elle procède (causalement ou métaphysiquement).

Une autre question est celle de la part de stabilité et de la part de la plasticité dans l'architecture cérébrale. Personne ne conteste que le système nerveux central soit capable de se réorganiser à plusieurs échelles de temps et d'espace. Les chauffeurs de taxi londoniens présentent un surdéveloppement sensible de l'hippocampe, structure essentielle pour la navigation spatiale (Maguire *et al.*, 1997). Un certain nombre d'enfants, victimes d'une épilepsie gravissime, ont été soumis très jeunes à l'ablation d'un hémisphère cérébral entier, et présentent un profil cognitif essentiellement normal (Battro, 2001). Mais la question est de savoir dans quelle mesure le cerveau se « construit » lui-même au cours de son existence, sous l'effet de l'expérience et des tâches qu'il accomplit. Pour les partisans du « constructivisme neuronal », la plasticité cérébrale rend vaines les tentatives pour dégager une architecture qui soit à la fois celle du cerveau et celle de l'esprit (Quartz & Sejnowski, 1997).

C'est donc le concept même de « base neurale » qui se trouve mis en question, du moins dans la version qui semble s'ajointer le plus naturellement à l'idée d'une correspondance terme à terme des primitives cognitives et des structures neurales fondamentales. Cette idée sous-tend le principe méthodologique simple selon lequel un même phénomène cognitif (mémoire, raisonnement, reconnaissance des visages, planification, etc.) peut être étudié à deux niveaux : le niveau informationnel et le niveau cérébral ou neural, les deux approches étant directement liées et pouvant ainsi s'appuyer mutuellement.

1.5 La distinction entre fonctions inférieures et supérieures et l'hypothèse de la modularité massive

Revenons cette fois à la modularité selon Fodor. Autant sa conception de processus modulaire et d'organisation modulaire de la cognition s'inscrivait dans le droit fil d'un courant de recherche séculaire, autant la coupure franche qu'il introduisait entre systèmes modulaires et systèmes centraux, assortie d'un principe d'inaccessibilité des seconds à l'enquête scientifique, heurtait de front les présupposés, et les espoirs, de bon nombre de chercheurs.

Les processus modulaires, on l'a vu, sont liés pour l'essentiel (l'exception étant certaines fonctions linguistiques) à la perception et à la motricité. Ce sont donc les processus « inférieurs », qui possèdent des analogues chez les animaux non humains. Notons au passage que tout en reprenant la distinction traditionnelle entre processus inférieurs et processus supérieurs, Fodor, en représentant des sciences cognitives contemporaines, la transforme profondément. La différence ontologique entre des systèmes psychophysiques, capteurs ou effecteurs, pures machines biologiques, et

des processus intellectuels, purement mentaux ou idéels, disparaît dans le cadre contemporain au profit d'une distinction structurelle entre deux grandes catégories de systèmes biologiques de traitement de l'information.

Une élucidation des processus « inférieurs » chez l'homme et l'animal n'a, dans ces conditions, rien de trivial. Elle pose des problèmes scientifiques et philosophiques considérables, elle offre des perspectives comparatistes essentielles pour la compréhension des processus chez l'homme, elle est indispensable pour la compréhension des processus « supérieurs », enfin elle peut proposer des pistes ou des modèles pour l'étude de ces derniers. Cependant, il est vrai que les sciences cognitives ont pour ambition première de rendre compte de la cognition dans son ensemble, et qu'une exclusion de principe des processus « supérieurs » constituerait, si elle était fondée, une terrible déception (ainsi du reste que la confirmation du point de vue sceptique à l'égard des prétentions des sciences psychologiques qui reste majoritaire dans tout un secteur de l'opinion, notamment chez beaucoup de philosophes et de spécialistes des sciences de l'homme).

L'une des ripostes possibles au pronostic de Fodor consiste à rejeter tout ou partie de ses hypothèses fondamentales : l'existence de modules, leur caractère largement inné, la distinction entre processus inférieurs et processus supérieurs… Nous n'en parlerons pas, mais nous dirons quelques mots d'une réaction différente, qui a consisté à accepter l'analyse de Fodor, tout en rejetant l'une de ses deux principales conclusions, à savoir la non-modularité des processus supérieurs. Les partisans de la « modularité massive » (Tooby & Cosmides, 1992 ; Hirschfeld & Gelman, 1994 ; Sperber, 2005 ; Carruthers, 2006) défendent l'idée que ces processus sont, en tout ou en partie, également modulaires. La modularité dont ils jouissent est comprise de manière un peu plus souple que le sens fodorien. L'accent est mis sur (i) la « spécificité de domaine » ou « domanialité » (en anglais : *domain specificity*) : un module supérieur ne traite que les informations relatives à un secteur bien délimité du monde naturel, conceptuel ou social ; (ii) l'isolement informationnel (*encapsulation*) : un module n'a accès qu'à un stock limité d'informations, qui lui est propre ; (iii) l'innéité ; (iv) le caractère adaptatif. Les arguments généraux en faveur de la modularité massive sont exactement les mêmes que les arguments généraux en faveur de la modularité tout court. S'y ajoutent des arguments relatifs à différents modules supérieurs conjecturés en particulier par des psychologues du développement, parmi lesquels on cite souvent certaines « théories naïves », corpus de connaissances tacites spécialisées, présentes très tôt dans le développement, présentant peu de différences interindividuelles, universelles dans toutes les cultures, et ayant une fonctionnalité dont on peut conjecturer qu'elle était importante dans l'environnement adaptatif d'*Homo sapiens*, fonctionnalité qui conserve souvent de l'importance aujourd'hui. Des exemples de tels corpus, qui constituent ce qu'on appelle aussi parfois le « savoir-noyau » (*core knowledge* ; Spelke, 2000), sont : un ou plusieurs systèmes numériques, une physique naïve, une psychologie naïve, une biologie naïve, une sociologie naïve, un système de gestion de la coopération…

Remarquons une ambiguïté : s'agit-il seulement de corpus de connaissances (en quelque sens précis que ce puisse être) relatives à des domaines particuliers, et permettant au jeune enfant, et plus tard à l'adulte, d'agir de manière rapide et adaptée dans les situations relevant de chaque domaine, ou bien s'agit-il de systèmes cognitifs, comprenant non seulement des connaissances mais également des mécanismes particuliers de mise en œuvre de ces connaissances ? Dans le premier cas, l'hypothèse de modularité se vide de tout contenu spécifique, au-delà de l'idée triviale de connaissance spécialisée, et de l'idée hautement non triviale, mais différente, d'innéité. Seule la seconde lecture donne à la modularité son véritable sens « architectural » et sa fécondité éventuelle (conditionnée par son degré de fidélité aux faits).

Mais le problème principal que soulève l'hypothèse de la modularité massive est celui que nous avions signalé dès le début de la discussion. Une fois retiré les modules, supérieurs et inférieurs, reste-t-il quelque chose de l'esprit ? Les deux réponses possibles sont données par différents défenseurs de la modularité massive. La réponse positive risque de priver l'hypothèse d'une partie de son intérêt, car elle ménage la possibilité, très plausible comme on va le voir, qu'une part essentielle des propriétés de l'esprit humain, tout particulièrement ses vertus exceptionnelles dans le règne vivant, réside dans la partie non modulaire. Il ne faut pourtant pas tomber dans l'excès inverse : une architecture partiellement modulaire des processus supérieurs aurait des conséquences théoriques et pratiques importantes (pour l'éducation, par exemple).

Quant à la réponse négative, la plus audacieuse, elle appelle toute une série d'objections. L'une des sources de puissance de l'esprit humain semble justement résider dans sa capacité à appliquer à une variété très grande de situations, y compris des situations entièrement nouvelles, un certain nombre de procédures générales ne relevant d'aucun domaine en particulier. Ensuite, si les modules ne sont compétents que dans leur domaine propre, comment fait-on face à des situations qui relèvent pour partie du domaine d'un premier module, pour partie du domaine d'un second module ? De manière plus générale, la flexibilité et l'inventivité ne sont-elles pas la marque de l'intelligence, et ne confèrent-elles pas à l'esprit une part de sa stupéfiante efficacité ? Un esprit entièrement modulaire ne serait-il pas *a contrario* réduit à réagir de manière réflexe aux problèmes qu'il rencontre, en les catégorisant selon le module compétent ? N'est-ce pas précisément de cette manière que fonctionne une société bureaucratique ossifiée, avec les résultats que l'on sait ? Les habitudes limitent certainement en pratique notre capacité à déployer avec souplesse et promptitude des stratégies nouvelles, mais elles ne semblent pas, contrairement à une architecture massivement modulaire, l'interdire absolument. Ce dernier argument renvoie à la notion, problématique mais résistante, d'intelligence *générale*, que nous avons déjà rencontrée dans le contexte de la première IA, et corrélativement à l'existence de syndromes de handicap mental *général*.

À quoi les partisans de la modularité massive répondent de deux manières. Ils contestent, d'une part, le sérieux des arguments de leurs adversaires : après tout,

s'agit-il d'autre chose que de constatations de sens commun, appuyées sur rien d'autre que nos intuitions ? Ce sentiment de flexibilité, de fluidité, de mobilité, accompagné d'une conviction introspective d'homogénéité des processus supérieurs, tout cela a-t-il des chances de résister davantage à l'enquête scientifique que le sentiment que nous avons de l'homogénéité de notre vision, de l'isotropie de notre champ visuel et de la connexité de notre image rétinienne (thèses que l'on peut considérer aujourd'hui comme définitivement réfutées) ? Pour les processus supérieurs comme pour la perception, ces questions sont de nature empirique et les évidences introspectives sont dépourvues de poids. Une deuxième riposte, plus ciblée, a été proposée par Dan Sperber (Sperber, 2001). D'une part, il rappelle qu'il faut concevoir les modules à l'image du système d'acquisition du langage selon Chomsky : ces modules sont des systèmes spécialisés d'apprentissage qui permettent à l'organisme de façonner des composants modulaires adaptés à l'environnement et, en ce sens, acquis (la grammaire universelle est un module inné, mais qui sert à acquérir, au contact avec un environnement linguistique particulier, la maîtrise d'une langue particulière parmi les cinq à six mille qui existent aujourd'hui encore). Enfin, Sperber conjecture l'existence d'un module supérieur particulier, dit « métareprésentationnel », dont le domaine est constitué par les représentations issues de tous les autres modules. Ce module peut ainsi « croiser » et combiner les informations collectées par les différents modules, et assurer ainsi les fonctions de transfert, de généralisation, etc., qui confèrent au système cognitif les qualités que lui attribuent les adversaires de la modularité massive. Cette hypothèse métareprésentationnelle fait écho à la conception très ancienne selon laquelle c'est le langage qui permet à l'esprit humain d'accéder aux plus hautes performances cognitives : travailler sur des termes et des phrases, c'est traiter non pas directement les objets et états de fait du monde, mais leurs représentations linguistiques. Il y a cependant un fossé entre la conception traditionnelle et l'hypothèse de Sperber : celle-là prend l'esprit comme un donné, celle-ci prétend l'expliquer par un principe de réflexion en vertu duquel une propriété de l'esprit est reflétée au niveau de son fonctionnement interne. Nous reviendrons sur ce principe.

1.6 La perspective évolutionniste en sciences cognitives

Les défenseurs de la modularité massive accordent une importance primordiale à la nature biologique de l'esprit. La théorie de l'évolution constitue donc à leurs yeux une ressource théorique essentielle : elle commande, comme pour l'ensemble de la biologie, un registre explicatif spécifique et premier. De plus, il ne s'agit pas d'une simple position de principe, comme cela reste le cas dans nombre de secteurs de la biologie : les modularistes ne peuvent pas se passer de la perspective évolutionniste. En cela aussi, ils s'opposent à Fodor (Fodor, 2000, 2008) et rejoignent Daniel Dennett (Dennett, 1995), qui fut le premier philosophe des sciences cognitives à placer l'évolution à l'origine de la cognition, et à faire, par conséquent, de la théorie de l'évolution le fondement même des sciences cognitives.

Historiquement, l'émergence du thème évolutionniste est un fait frappant : il est difficile d'imaginer aujourd'hui que les sciences cognitives sont nées et ont longtemps grandi dans une ignorance complète de la théorie de l'évolution. Chomsky lui-même, l'un des pères de la « révolution cognitive », insista très tôt sur le caractère fondamentalement biologique de la cognition, mais résista longtemps à l'idée que la théorie de l'évolution pourrait contribuer à en rendre compte scientifiquement. Ce cheminement illustre, ironiquement, l'une des principales raisons avancées par Fodor pour nier que les processus supérieurs puissent devenir un jour objet d'une science de la nature : comme on l'a vu, une leçon capitale de la philosophie des sciences, selon lui, est que rien n'exclut qu'un fait si éloigné en apparence que ce soit se révèle pertinent dans l'évaluation d'une croyance.

Le « tournant évolutionniste » des sciences cognitives se manifeste de manière diffuse dans tout le domaine : même lorsqu'on est incapable de dire précisément *comment* tel phénomène que l'on étudie a pu se mettre en place au cours de l'évolution, il est communément admis qu'il faudrait, dans le meilleur des mondes scientifiques, pouvoir l'expliquer, car ce phénomène était d'abord absent, puis a émergé au cours de l'évolution, et que pour en rendre compte de manière complète, il faut pouvoir au moins montrer comment cette émergence est théoriquement possible (Bickhard, 2002).

Mais la théorie de l'évolution intervient de manière beaucoup plus constructive et précise dans les sciences cognitives, en nourrissant deux branches nouvelles (qui n'en font en réalité qu'une, tant elles sont intriquées) : la psychologie évolutionniste et l'anthropologie évolutionniste, ou théorie évolutionniste de la culture. Les questions que soulèvent ces programmes de recherche sont multiples, mais on peut les classer en trois grandes familles.

Il y a d'abord les questions de *méthode*. Aux difficultés générales de l'application de la théorie de l'évolution s'ajoutent dans le cas de la cognition (i) l'absence quasi totale d'archives fossiles, l'essentiel de la structure pertinente étant composée de parties molles, et les parties dures (anatomie crânienne, cavité pharyngienne…) ne fournissant que des indices très partiels, difficiles à interpréter ; (ii) la paucité des informations solides dont on dispose concernant l'environnement évolutionnaire adaptatif (EEA) dans lequel a émergé notre espèce ; (iii) le caractère encore très fragmentaire de nos hypothèses concernant l'architecture de l'esprit : les composantes élémentaires du système cognitif sont très loin d'avoir été identifiées avec le même degré de certitude et de précision que les organes, systèmes ou structures corporelles auxquels elles sont comparées. La situation s'améliore néanmoins avec le développement de la paléogénétique et de l'éthologie cognitive et avec les progrès des neurosciences cognitives. Les problèmes méthodologiques n'en restent pas moins nombreux et complexes.

Les questions de *fondements* ne sont pas moins pressantes. On peut parfaitement admettre que les bases matérielles de l'esprit, son « siège », sont un système biologique, comparable à cet égard au système cardio-vasculaire, au système digestif ou au

système locomoteur, dont la forme actuelle et les fonctions ont été conjointement façonnées par l'évolution. L'esprit a toutefois cette particularité essentielle d'être doté de dispositions qui vont bien au-delà de toute spécification initiale à laquelle la sélection naturelle a pu donner satisfaction. Contrairement aux autres systèmes biologiques, le système nerveux central humain soutient non seulement des fonctions spécialisées ou dédiées, mais également des « métafonctions » capables de produire des processus et des entités qui ne gardent aucune ou presque aucune trace des mécanismes évolués[1] présents dans le système[2]. La culture, prise au sens le plus large possible, comprend des processus et entités de ce genre, mêlés certes à des entités évoluées, mais en proportion telle qu'il se pourrait qu'au total les ressources explicatives de la théorie de l'évolution soient d'une utilité marginale pour une science de la culture. Dans la mesure où l'esprit, parmi ses « métafonctions », possède la capacité d'absorber et d'incorporer une vaste quantité de matériaux externes fournis par l'expérience individuelle et plus encore par la culture, la psychologie elle-même est « contaminée » par la culture : les déterminations biologiques, en particulier évolutionnistes, prennent peut-être la seconde place derrière les déterminations culturelles.

Se manifeste ici, bien évidemment, la méfiance éternelle des sciences de l'homme culturalistes, historicistes, interprétativistes à l'égard du naturalisme, et il est généralement admis que les sciences cognitives doivent poursuivre leur petit bonhomme de chemin sans prêter trop attention à ces inquiétudes : tant qu'on ne leur conteste pas toute espèce de pertinence, donc le droit à l'existence, les sciences cognitives doivent poursuivre leur objectif, qui est la mise au jour des contraintes naturelles. Selon la plupart des chercheurs du domaine, l'importance respective de ces contraintes et des déterminations culturelles fera l'objet d'un arbitrage ultérieur, qu'il serait très prématuré de prononcer maintenant. Les termes mêmes de l'arbitrage restent à déterminer, dans la mesure où le ou les modes d'interaction entre « nature » et « culture » sont l'objet d'une part importante des recherches actuelles. En particulier, l'un des thèmes principaux de l'anthropologie évolutionniste est la « co-évolution » des gènes et de la culture : comme l'illustrent de nombreux exemples, la culture contribue à sélectionner des gènes, en favorisant leurs porteurs par des dispositifs coutumiers, institutionnels ou matériels (Richerson & Boyd, 2004 ; Diamond, 1997 ; Sterelny, 2004). Il y a là une tentative intéressante pour apporter une réponse scientifique à un problème de fondement, source de querelles philosophiques sans fin.

La troisième catégorie de questions porte sur la *fécondité* des approches évolutionnistes. Pendant longtemps, elles ont porté, pour l'essentiel, sur les fonctions les plus directement liées à la sélection naturelle, à savoir les fonctions reproductives. Elles

1. Ce terme revêt dans le présent contexte un sens technique : est « évolué » (en anglais : *evolved*) un mécanisme, système ou processus, qui résulte de l'évolution biologique.
2. Le système locomoteur occupe à cet égard une position intermédiaire : il n'a pas été sélectionné « pour » la danse ou l'acrobatie, mais ses « métafonctions » restent très limitées, et les traces de l'évolution restent visibles dans toutes ses productions.

s'étendent aujourd'hui aux fonctions cognitives supérieures, en particulier dans le cadre offert par la modularité massive, au langage notamment, et même aux structures cognitives les plus générales qui rendent possibles la socialité humaine (et la socialité d'autres espèces) et la culture, en particulier les systèmes normatifs, sur laquelle repose toute société humaine. Cette nouvelle phase des recherches fait perdre aux controverses de la phase précédente beaucoup de leur tranchant. Elle soulève, en revanche, la question de la portée des résultats qu'on peut en attendre. Quelles sont les découvertes, ou les arbitrages, que l'on peut attendre de l'approche évolutionniste ? C'est l'objet d'un débat très vif, qu'on ne peut aborder ici.

Avant de clore cette première partie, il faut insister sur le fait que la modularité nous a servi à la fois d'exemple caractéristique d'une question de philosophie des sciences cognitives et de fil conducteur. Nous avons ainsi pu rencontrer toute une série d'autres questions et hypothèses qui sont probablement plus centrales et durables que la modularité elle-même. Il n'est pas exclu, en effet, que la modularité, en tant que telle, cesse de faire l'objet de discussions d'ici quelques années (même si depuis un quart de siècle elle figure sur la liste des « questions vives » de la discipline, et compte parmi les sujets favoris des philosophes des sciences cognitives), alors que les autres thèmes semblent appartenir à un socle beaucoup plus durable d'interrogations. Cette évolution est d'ailleurs esquissée : plus que de modularité, les chercheurs débattent aujourd'hui, à propos de raisonnement mais aussi de manière plus générale, de théories « duales » de la cognition (*dual process theories* : Evans, 2003 ; Egidi, 2007). Ce qui est proposé sous ce terme est l'idée que deux sortes de processus sont concurremment ou successivement à l'œuvre dans beaucoup de processus cognitifs : des processus automatiques, échappant au contrôle volontaire, rapides, rigides, généralement non conscients, et des processus volontaires, délibératifs, conscients, lents, faillibles. On retrouve là certaines des propriétés invoquées dans le débat sur la modularité, mais le thème des facultés quitte l'avant-scène au profit d'une organisation assez différente du travail mental. Ce qui reste néanmoins de la problématique modulariste, c'est l'idée d'une architecture de l'esprit, structuré en composantes stables.

2. L'esprit comme objet de science : fondements et domaine des sciences cognitives

2.1 Qu'est-ce que fonder les sciences cognitives ?

Une mission traditionnelle de la philosophie des sciences, reconnue dans la plupart de ses écoles de pensée, est la mise au jour des fondements, que ce soit ceux des sciences en général (ou de *la* science), ou ceux d'une discipline particulière. Mais qu'est-ce que les fondements, en quoi consiste leur mise au jour, et quel est l'apport

de la philosophie, sachant que la science elle-même peut sembler dans son mouvement même se charger de la tâche? C'est sur ces questions que les écoles divergent.

Pour nous limiter au contexte présent et aux fondements d'une discipline particulière, on peut discerner deux attitudes principales. L'objectif du philosophe, pour certains, doit être de construire un cadre métaphysique cohérent et complet dans lequel la science ait sa place. Pour d'autres, cet objectif doit être de dégager la cohérence de la discipline, en explicitant ses présupposés et en exhibant la structure logique de ses concepts fondamentaux. Pour le dire brièvement, le contraste oppose une conception globale ou externe de l'intelligibilité recherchée, et une conception locale ou interne. Enfin, un philosophe peut refuser de choisir, et faire siens tous ces objectifs, voire refuser de tracer entre eux une frontière nette.

Cette distinction en croise une autre, qui porte sur la troisième question, celle des rôles respectifs de la philosophie et de la science. Pour le philosophe naturaliste, les deux entreprises sont dans un rapport de continuité, la philosophie se situant aux marches de la science, dans sa zone de plus grande abstraction. La question d'une répartition des rôles ne se pose donc pas (elle n'admet en tout cas pas de réponse stable, puisque les fruits de l'activité philosophique sont rapidement intégrés au foyer actif de la science). Selon le philosophe naturaliste, si l'objectif est de dresser le tableau métaphysique, la science y contribue au même titre que la philosophie, et dans le même mouvement. De même, si l'objectif est la « grammaire » conceptuelle de la discipline, l'intrication de la philosophie et de la science est complète.

S'il n'épouse pas, ou pas complètement, le naturalisme, le philosophe voit les choses différemment. Il tend à rejeter l'idée que la science puisse contribuer notablement à dresser le tableau métaphysique ; pour autant, il peut estimer que la tâche ne concerne pas davantage la philosophie des sciences, dont à ses yeux l'unique mission, qui n'est pas celle de la science, est de mener à bien l'explicitation du cadre conceptuel de la science étudiée.

S'agissant des sciences cognitives, ces questions sont rendues particulièrement délicates du fait de leur objet. L'option métaphysique consiste à inclure dans le champ de la philosophie des sciences cognitives le problème corps-esprit, le problème de l'intentionnalité, la nature des représentations mentales et de la perception, la conscience, le libre arbitre… ; et, selon que l'on est naturaliste ou non, s'y intéresser en tant qu'objet des sciences cognitives elles-mêmes, ou en tant que parties constitutives du cadre philosophique général dont la cohérence avec les résultats scientifiques doit être assurée.

Nous reviendrons, dans la conclusion, sur le partage des tâches, au sein même de la philosophie, entre les différentes branches concernées par les sciences cognitives. Ici, nous prendrons le parti de la modestie et placerons au cœur de la philosophie des sciences cognitives l'étude de ses concepts les plus généraux. Prenons par exemple, le problème corps-esprit, qui désigne en réalité plusieurs énigmes distinctes quoique liées, mais dont nous ne considérerons ici qu'une formulation simple : comment

rendre compte de la place des entités mentales dans l'ordre matériel. Certains estiment qu'il sera résolu *par* les sciences cognitives (dont ce serait d'ailleurs le but premier), de la même manière que la biologie a (peut-on penser) résolu le problème vie-matière, ou que la physique a dessaisi Zeus du tonnerre au profit de l'électromagnétisme. D'autres pensent qu'il faut lui trouver une solution pour que les sciences cognitives acquièrent un fondement solide. Mais le philosophe des sciences « modeste », pour sa part, constate que les sciences cognitives ont justement développé une stratégie qui leur permet de contourner ce problème[1]. Nous avions évoqué au tout début de ce chapitre le « structuralisme » inhérent au projet des sciences cognitives. Nous sommes en mesure, enfin, d'en parler de manière plus précise.

2.2 Représentation et computation : le cadre fonctionnaliste et le langage de la pensée

2.2.1 *Le fonctionnalisme*

Les sciences cognitives ont pris leur essor dans un cadre théorique relativement précis, qui a non seulement historiquement constitué leur point d'appui initial, mais qui demeure aussi, par-delà les critiques qui lui sont adressées, et les ajustements qui lui sont actuellement apportés dans l'espoir (vain selon certains, raisonnable selon d'autres) de le sauver, le point de départ de toute discussion de leurs fondements. Ce cadre, nous l'appellerons « fonctionnalisme », nous conformant à un usage répandu, en dépit de l'ambiguïté du terme[2].

Le fonctionnalisme est une forme de structuralisme appliqué aux entités mentales. Il consiste à substituer à la question de la nature de ces entités une description de leurs rapports mutuels. Plus exactement, tout ce que nous avons à connaître d'états tels que les douleurs, les croyances, les désirs, les souvenirs, les regrets, les intentions, les projets, etc., ce sont les rapports qui existent entre eux, ainsi que les rapports qu'ils entretiennent avec les stimulations sensorielles et les mouvements. Les rapports de cette seconde espèce constituent quelque chose comme des conditions aux limites observables : remarquons, en effet, que les états internes que sont les croyances et autres ne sont pas observables, sinon (peut-être : beaucoup en doutent) par l'agent lui-même. Pour le scientifique, ce sont des entités théoriques qui jouent au sein des théories de la cognition le rôle qu'ont, par exemple, les forces dans la dynamique newtonienne, les quarks dans la physique des particules, l'utilité espérée en économie, la pression sélective en théorie de l'évolution, etc.

---

1. Cette possibilité avait été entrevue par certains psychologues dès le XVIII^e siècle (*cf.* Hatfield, 1995).
2. Nous allons voir que dans le contexte des sciences cognitives, il y a plusieurs conceptions du fonctionnalisme. Mais le terme recouvre également des positions prises dans d'autres champs, notamment la linguistique, l'anthropologie et la sociologie, les sciences de la vie, etc. Ces autres emplois sont sans rapport (en tout cas direct) avec le fonctionnalisme dans les sciences cognitives et la philosophie de l'esprit.

Les rapports qu'entretiennent les états mentaux internes entre eux et avec les stimulations et la motricité sont de nature causale, et engendrent la dynamique mentale (avec des antécédents et des conséquences physiques[1]). Le système cognitif passe ainsi d'un état complexe à l'autre, sous l'effet de forces qui sont fonction des rapports constants existant entre les différents types d'états mentaux. Pour prendre un exemple, ma croyance que j'ai mal à la tête depuis un moment est appréhendée (sur le plan théorique) par le biais des rapports que cette croyance entretient avec des stimuli sensoriels (ces stimuli ont contribué à causer cette croyance, et ce genre de stimuli tendent à causer, *mutatis mutandis,* une croyance du type « j'ai mal à la tête depuis un moment »), avec des désirs tels que celui de mettre un terme à mon mal de tête, lequel se combine avec une autre croyance, portant sur l'efficacité de l'aspirine, pour tendre à causer une intention de prendre de l'aspirine, intention qui à son tour provoque, en conjonction avec d'autres croyances, intentions et désirs, un plan de recherche d'aspirine dans l'armoire à pharmacie, etc.

L'intuition fonctionnaliste fondamentale est donc celle-ci : s'il s'agit de mettre au jour les déterminations de la dynamique mentale, ou encore, pour reprendre une expression d'une autre époque, les « lois de la pensée », il n'est pas nécessaire de se prononcer sur l'étoffe dans laquelle les états mentaux, les pensées, sont découpés ; il suffit de mettre au jour les liens constants qui existent entre eux. Ces liens sont dispositionnels : en présence de certaines conditions, un enchaînement causal spécifique est déclenché (rappelons l'exemple type de propriété dispositionnelle : plongé dans l'eau, le sucre fond, sauf situation exceptionnelle : sa solubilité est une propriété dispositionnelle). Mais cette causalité doit être mise au jour. Elle appelle en fait deux explications : l'une vise le phénomène général, l'autre sa distribution. Il s'agit de comprendre, d'une part, comment une pensée peut causer quelque événement que ce soit ; et d'autre part, ce qui fait que la pensée que j'ai mal à la tête, contrairement au projet de mettre fin à mes jours, ne me conduit (normalement) pas à l'intention d'avaler de la strychnine.

Pour cela, il faut en dire un peu plus sur les états mentaux. Leur « opérationnalisation » reste abstraite tant qu'on n'a pas précisé la manière dont ils sont individués. C'est ici que se séparent plusieurs conceptions du fonctionnalisme. Pour le fonctionnalisme *analytique*, chaque état mental est *défini* par sa place dans le réseau des

1. On se heurte ici à une difficulté terminologique bien connue : tout partisan du naturalisme, fût-ce à titre seulement méthodologique et non métaphysique, attribue aux états et processus mentaux une nature physique : une croyance ou une douleur particulière n'est pas considérée comme moins physique qu'une stimulation rétinienne ou qu'un mouvement de la main. La différence pertinente est que la croyance est entendue en tant qu'elle possède un contenu sémantique ; elle est un événement physique, certes, mais saisi sous une description particulière qui ne l'est pas. Nous y revenons dans un instant, mais un exemple tiré d'un autre domaine peut aider le lecteur : quand je parle d'un billet de 20 euros, je parle bien d'un objet matériel, mais j'en parle via sa valeur nominale, et je choisis cette description car c'est celle dont j'ai besoin pour rendre compte de ce qui se passe à la boulangerie quand je paie ma baguette. Cet exemple n'est pas sans poser à son tour des problèmes, mais il n'est proposé ici qu'à titre d'éclaircissement provisoire.

dispositions exprimées par les platitudes de sens commun dans lequel il figure (la croyance que l'on a mal à la tête, en présence de la croyance que l'aspirine soulage le mal de tête, déclenche, en l'absence de la crainte d'être allergique à l'aspirine, l'intention d'absorber de l'aspirine : la croyance que l'on a mal à la tête n'est rien d'autre que le rôle fonctionnel occupé dans le réseau de toutes les platitudes de ce genre). Pour le fonctionnalisme *empirique* (ou *psychofonctionnalisme*), le réseau des platitudes sert seulement à désigner les entités mentales, et c'est la science qui est chargée de déterminer leurs véritables propriétés ; de la même manière, le sens commun *désigne* l'eau (il donne le *sens* du mot ou du concept), mais c'est la physico-chimie qui *découvre* ce que l'eau est réellement[1] (qui en fixe l'*extension*). Enfin, le fonctionnalisme *turingien* ou *mécaniste* (*machine functionalism* en anglais) assimile les états mentaux aux états internes d'une machine de Turing (ou, plus généralement, d'un système computationnel).

2.2.2 *La théorie computationnelle de l'esprit*

Le fonctionnalisme turingien se place sur un autre plan que les deux précédents, et il n'est incompatible ni avec l'un ni avec l'autre. C'est une hypothèse d'un très haut degré d'abstraction et, il faut bien le dire, difficilement compréhensible en dehors du contexte plus général de la théorie psychologique dans lequel il prend place, et qu'il faut maintenant rapidement exposer. Il s'agit de la théorie computationnelle de l'esprit (TCE, en anglais CTM pour *computational theory of mind*)[2].

Les fonctionnalismes analytique et empirique ont pour motivation première une analyse conceptuelle des *états* mentaux. Ils dérivent, d'autre part, du « béhaviorisme logique », dont ils rejettent en partie l'héritage mais conservent le souci d'économie ontologique, et la vive conscience de la difficulté de donner une définition essentialiste des entités mentales. Cette forme philosophique de béhaviorisme était l'aboutissement d'une réflexion d'origine largement wittgensteinienne sur les réifications abusives auxquelles conduit une confiance excessive dans la forme superficielle des expressions du langage *commun*.

Le fonctionnalisme turingien, quant à lui, part d'une réflexion sur les *processus* mentaux, et puise par ailleurs à la longue réflexion, amorcée avec Frege, qui aboutit dans les années 1930 à la notion de langage (ou système) *formel*. L'arithmétique fournit des exemples caractéristiques de ces langages[3] : on se donne des symboles pour des nombres particuliers tels que 0 ou 1, des symboles pour certaines opérations telles que le passage d'un entier au suivant, l'addition ou la multiplication, des symboles pour des nombres quelconques, des symboles logiques, et des règles

1. Soit dit en passant, la réponse n'est pas « H_2O » ; elle est bien plus complexe que cela (Weisberg, 2006). Mais c'est une réponse de ce genre que la science a pour rôle de fournir.
2. C'est cette théorie plus complète que certains auteurs (par exemple, Putnam lui-même : Putnam, 1988) appellent « fonctionnalisme ».
3. Il existe plusieurs langages formels qui s'adaptent naturellement à l'arithmétique.

morphologiques de combinaison de ces symboles. Un langage de ce genre peut planer dans la sphère des idéalités ou concepts abstraits, ou bien être matériellement « réalisé » de diverses manières. Toute calculatrice, depuis la pascaline jusqu'aux calculateurs analogiques de Zuse et aux appareils mécaniques ou électromécaniques qui précèdent l'électronique, puis aux calculettes et ordinateurs contemporains, réalise ou « implémente » un langage formel de l'arithmétique. Ces réalisations sont multiples : les enchaînements causaux qu'elles impliquent sont profondément différents, et ne soutiennent entre eux aucune sorte d'isomorphisme exprimable dans le langage de la physique[1]. Ce qu'ils ont en commun n'est visible que d'un point de vue abstrait, celui des spécifications formelles qui ont présidé à leur construction. L'intuition fondamentale du fonctionnalisme turingien est que les opérations mentales sont formelles, et qu'elles peuvent être physiquement réalisées de différentes manières, en sorte que la théorie de ces opérations ne relève pas de la physique, mais d'une science formelle qu'on pourrait appeler science de l'information (cette expression n'est en fait pas employée en ce sens). Plus concrètement, une loi de la pensée telle que le *modus ponens* (passage de l'ensemble formé des deux pensées que A et que A implique B à la pensée que B) doit être comprise comme une forme abstraite de *calcul* et qu'un système matériel obéit à cette loi dans la mesure où il effectue concrètement ce calcul (comme l'élève qui *écrit* « B » à la craie sur le tableau noir sous les inscriptions « A » et « $A \supset B$ »). Il en est de même, *mutatis mutandis*[2], du passage d'une pensée de migraine et d'une croyance quant à l'efficacité de l'aspirine à une intention d'absorber de l'aspirine. Il est de fait que ce genre d'opération abstraite peut être réalisé par des mécanismes différents sur le plan physique. Cet argument, dit de la « réalisabilité multiple », est au fondement du fonctionnalisme turingien et de la théorie computationnelle de l'esprit qui en constitue le développement.

Un langage formel a deux visages : il est, d'une part, une combinatoire de symboles, d'autre part, le support d'une « interprétation » qui attribue aux symboles, termes et énoncés du langage un sens. Interprété, le langage désigne des objets, relations et états de fait dans un « univers » qui peut être abstrait (l'ensemble des nombres entiers par exemple, ou un échiquier muni de ses pièces, compris non comme objet matériel mais comme système de relations) ou concret, réel ou imaginaire. Ces deux visages sont corrélés de la manière suivante : une opération sur les symboles correspond à une mise en relation des entités interprétées, en sorte que le tableau changeant des configurations symboliques reflète les aspects pertinents du domaine d'interprétation. Ainsi, le contrôleur aérien suit à la trace et guide les avions à partir de symboles qui en indiquent l'identité, la position, la destination ; les opérations du contrôleur portent sur les symboles, mais la correspondance assure que ces opérations

1. Pour faire ressortir encore plus clairement cette idée, on propose parfois d'imaginer des calculateurs constitués de poules pondant des œufs reliées par des tubes, d'enfants qui se transmettent des cris (pas des mots) dans la cour de récréation, de canettes de bière connectées par des jeux de ficelles, etc.
2. Les deux cas diffèrent notablement par certains aspects : on y revient sous peu.

renvoient de manière fiable aux trajectoires des avions eux-mêmes, en sorte que, sauf accident, les avions arrivent à bon port, selon les intentions du contrôleur[1].

Il manque deux éléments essentiels à ce schéma pour qu'il puisse constituer, serait-ce à l'état d'ébauche, une théorie de l'esprit. Le premier porte sur l'interprétation des symboles : en vertu de quoi représentent-ils ce qu'ils représentent, et que signifie concrètement qu'ils représentent quoi que ce soit ? La TCE est une théorie *représentationnelle*, en un sens familier en théorie de la connaissance depuis le XVII^e^ siècle : l'esprit est peuplé de représentations, que Descartes et Locke appellent en général des *idées*. C'est d'ailleurs la raison pour laquelle elle est parfois appelée théorie computo-représentationnelle de l'esprit. Il ne suffit pas cependant de lui accoler une étiquette supplémentaire : il faut montrer comment une théorie représentationnelle de l'esprit peut être aussi une théorie naturaliste de l'esprit.

L'exemple du contrôle aérien nous met sur la voie (sans nous mener au but) : ce qui confère aux inscriptions lues par le contrôleur sur ses écrans et ses *strips* leur valeur représentative, ce sont les connexions causales complexes qui vont des entités représentées (par exemple, un avion immatriculé N à l'endroit (x, y, z) de l'espace à l'instant t) aux inscriptions représentantes (ici, le positionnement d'un point étiqueté N à tel endroit de l'écran, associé aux coordonnées (x, y) plus la valeur z du paramètre *altitude*). Les symboles postulés par la TCE sont de même supposés être naturellement dotés de signification, mais ce qu'il faut entendre par là est très loin d'aller de soi, et nous évoquerons cette question sous l'intitulé « intentionalité » un peu plus loin. Notons dès à présent qu'à la différence des indicateurs dont dispose le contrôleur aérien, le système cognitif n'est pas occupé en son centre par un « contrôleur » disposant lui-même des principaux attributs de l'esprit : les symboles internes ne peuvent être « lus ». La solution à cette difficulté-ci est à rechercher du côté de l'idée fonctionnaliste : le sens d'un symbole pourrait être défini fonctionnellement par l'ensemble des effets que ce symbole peut (dispositionnellement) exercer sur le reste du système.

Le second vide à combler concerne les différentes catégories de pensée. Nous avons fait comme s'il n'y en avait qu'une : la croyance ou l'assertion. Or l'esprit entretient, on l'a noté plus haut, d'autres types d'états, par exemple des désirs qui sont précisément tout autre chose que des croyances sur l'état du monde : si je veux acheter une voiture, autrement dit si je veux que le monde soit tel que je sois propriétaire d'une voiture, c'est (normalement) que le monde n'est actuellement

1. Par souci de simplification, mais au risque de causer une confusion, je ne distingue pas dans cet exemple deux types de transformation en réalité très différents. Dans un cas, l'univers est fixe et ce sont les représentations de cet univers qui sont modifiées (par exemple, lorsque certaines conclusions inédites sont tirées d'informations déjà présentes). Dans l'autre, l'univers lui-même change, notamment en raison de l'intervention de l'agent. Les deux processus sont souvent à l'œuvre simultanément ; c'est le cas du contrôle aérien : à partir de données valables à un instant t, le contrôleur est amené à déduire (calculer) certaines informations supplémentaires valables au même instant ; mais il infère également, à partir d'informations valables au temps t et de connaissances sur l'évolution du système (sous l'effet de causes soit endogènes soit exogènes, dont sa propre intervention), des informations valables à un instant t' postérieur à t.

pas tel. L'esprit forme également, pour les considérer, toutes sortes de pensées qui ne sont ni des croyances ni nécessairement des désirs, mais des hypothèses : s'il avait fait beau hier, nous aurions pu rentrer le foin ; s'il fait beau demain, nous le ferons. L'esprit doit donc maintenir des listes séparées pour ses croyances, ses désirs, ses intentions, ses craintes, ses regrets… Il reste à préciser comment ces listes sont connectées : comme on l'a vu, certaines conjonctions de désirs et de croyances, par exemple, produisent des intentions ; mais tout désir ne se conjoint pas à n'importe quelle croyance pour produire une intention. L'esprit ne peut donc fonctionner ni si les listes sont étanches, ni s'il est impossible d'assortir leurs éléments de manière différenciée.

2.2.3 Le langage de la pensée

La TCE peut à son tour être immergée dans une théorie plus riche. Pour la présenter, nous avons utilisé l'exemple des langages formels de la logique, et parlé d'opérations ou calculs logiques. Mais rien dans la TCE n'oblige à postuler que le système symbolique qui est au cœur des opérations du système soit réellement un langage formel et que les opérations soient des calculs syntaxiques effectifs au sens de la logique. On pourrait très bien imaginer d'autres systèmes, et d'autres notions de computation que celles de la logique[1] ; on verra d'ailleurs bientôt (2.2.2) que de telles conceptions sont effectivement proposées.

L'hypothèse du langage de la pensée (HLP, LOTH en anglais pour *language of thought hypothesis*) est néanmoins, pour un esprit formé à la logique moderne, une extension apparemment naturelle de la TCE. Elle est que le médium représentationnel est précisément constitué par un langage formel du type de ceux que construit la logique, médium qu'on appelle « langage de la pensée » ou parfois « mentalais ». Cette hypothèse a toute une série de conséquences qui sont autant d'arguments en sa faveur :

1. Elle donne une forme parfaitement précise à la nature duale des états et processus mentaux. Les énoncés du mentalais ont une forme matérielle, qui leur confère des dispositions à se transformer sous l'effet de processus causaux dont la forme est donnée par la syntaxe. Ils ont aussi une sémantique, c'est-à-dire qu'ils renvoient à des entités, relations et états de fait de l'univers d'interprétation (qui est en général le monde matériel auquel l'organisme a accès via la perception et sur lequel il peut agir via la motricité). Syntaxe et sémantique sont indépendantes, mais sont comme le miroir l'une de l'autre. Cette conformité explique en particulier la compositionnalité, une propriété que beaucoup attribuent à la pensée, à savoir

1. Cette affirmation risque de faire bondir le lecteur qui a appris qu'il n'y a, en réalité, qu'une seule notion mathématique de computation (ce qui peut se discuter d'ailleurs). Mais dans le contexte présent, le concept est plus élastique, et peut désigner en réalité presque toute procédure mécanisable, même si elle fait intervenir des opérations ou des dispositifs qui ne respectent pas le cahier des charges de la computation au sens logique strict (c'est-à-dire la calculabilité).

le fait qu'une pensée complexe est entièrement caractérisée par sa structure et par les pensées qui la composent. Elle explique aussi que les transformations syntaxiques conservent la vérité : une pensée déduite formellement de pensées vraies (vérifiées dans l'univers d'interprétation) est vraie – pour le dire rapidement, en suivant la syntaxe, on ne quitte pas le chemin de la vérité.

2. Elle offre une solution élégante à la nécessité de séparer les pensées en listes distinctes, conformément à ce qui vient d'être exposé, tout en rendant possible certaines combinaisons. La croyance que P peut être vue comme une relation de la forme C(<P>), où C est un prédicat associé à la croyance et <P> une phrase de mentalais exprimant P. La croyance que P est ce que les philosophes appellent, à la suite d'un célèbre article de Russell de 1905, une *attitude propositionnelle*, et l'HLP en propose une théorie relationnelle très naturelle. De même, le désir que P serait une relation D(<P>), D étant un autre prédicat. Schématiquement, le fait pour un individu de croire que P serait réalisé par la présence de <P> dans un secteur de son esprit (ou de son cerveau) dédié aux croyances (sa « boîte à croyances » pour reprendre une expression imagée courante inventée par Schiffer, 1981) ; et désirer (que) P consisterait pour l' individu à avoir <P> dans sa « boîte à désirs ». Cette façon de réaliser croyances et désirs (ainsi que les autres attitudes propositionnelles) rend possible des appariements spécifiques : si je crois que P entraîne Q et que je désire Q, alors je forme l'intention de faire en sorte que P.
3. Elle permet d'expliquer l'indépendance relative, du moins apparente, d'une partie de la pensée vis-à-vis du langage (naturel, celui de la personne), En d'autres termes, si l'HLP est vraie, on comprend qu'une pensée sans langage soit possible (par exemple, celle des enfants préverbaux, dont on mesure de plus en plus l'étendue, et celle de diverses espèces animales). Corrélativement, elle ouvre la possibilité d'envisager l'acquisition du langage, conformément du reste à une intuition commune, comme un processus ancré dans une pensée déjà structurée : si cette structuration est procurée par le déploiement interne du mentalais, on échappe au risque de circularité.
4. Elle rend compte naturellement de l'intuition que différentes expressions linguistiques expriment une même pensée. « It's raining », « Piove », « Il pleut » ont le même sens, ce dont l'HLP rend compte de manière très simple : c'est la même phrase de mentalais qui est pensée, ou activée par le système cognitif ; de même, du reste, dans une même langue, pour des phrases telles que « Marie a tué Pierre » et « Pierre a été tué par Marie »[1]. On peut espérer expliquer de même le caractère universel de certains schémas de pensée (tels que des règles d'inférence), qui se traduisent très différemment dans différentes langues naturelles, et même dans différents idiolectes d'une même langue.

1. L'exemple ne vaut que sous la condition d'une forte idéalisation : il est clair qu'il existe des contextes d'énonciation dans lesquels on ne substituerait pas normalement un énoncé à l'autre.

5. La pensée semble, en première analyse, jouir des propriétés de « productivité » : une infinité de pensées peuvent être engendrées à partir d'un stock initial fini de pensées, et de « systématicité » : si une pensée telle que « Marie a tué Pierre » est pensable, les pensées « Pierre a tué Marie », « Quelqu'un a tué Pierre », « Marie a tué quelqu'un » sont nécessairement pensables[1]. Ces propriétés sont partagées par les langues, naturelles (au moins idéalement) et formelles, et l'HLP en rend compte aisément.

Pour autant, l'HLP n'a rien d'évident, et elle s'expose de fait à de fortes objections. Son apparente trivialité procède d'une illusion. La pensée comme *produit* peut bien être décrite à l'aide d'un langage formel (en admettant ici que les objections bien connues à l'idée que le langage naturel ait, moyennant certaines idéalisations, la structure d'un langage formel puissent être contournées en considérant que la pensée correspond au contenu, ou à la structure profonde, des énoncés du langage naturel, et non à leur forme de surface). Mais pourquoi *ce qui produit* la pensée, à savoir l'esprit, aurait-il précisément la même structure ? Une chose est de décrire la structure de la pensée, qui est l'objet de la logique (entendue de manière très large) ; autre chose est de décrire la genèse de la pensée, qui est l'objet de la psychologie. L'HLP est donc une hypothèse audacieuse, et non la formulation savante d'un truisme ; elle affirme que l'esprit, quelle que soit la tâche qu'il accomplit, procède comme un système formel autopropulsé : il applique des règles de composition et d'inférence formelles à des ensembles d'énoncés de mentalais. La version truistique serait d'expliquer que pour multiplier 31 par 12 (pour passer de la pensée composite (<multiplier>, <31>, <12>) à la pensée <372>, l'esprit applique une table interne de multiplication aux symboles signifiant en mentalais 31 et 12, et produit le symbole de mentalais signifiant 372. Cette interprétation conduirait, en réalité, à une régression : comment rendrait-on compte de cette opération interne ? Faudrait-il postuler, à l'intérieur de l'esprit, un sub-esprit qui lui permette d'effectuer la manœuvre ?

C'est l'erreur de l'homunculus. Comment l'HLP y échappe-t-elle ? Elle postule que lorsque *je* multiplie 31 par 12, mon système cognitif suit une trajectoire qu'on peut décrire comme l'application de certaines opérations à certains symboles complexes de mentalais. Mais ce qui distingue le système cognitif de moi, l'être conscient dont il s'agit d'expliquer le flux de pensées, c'est que le système cognitif est un mécanisme « aveugle », sans pensée, intelligence ni conscience. D'une part, tel un robot sur une chaîne de montage, il ne fait que déplacer des entités matérielles : ce qui est chez moi de l'ordre des raisons est dans le système de l'ordre des causes ; d'autre part,

1. Pour expliquer cette idée, Fodor, qui la propose, établit un parallèle avec les guides de conversation pour touristes, qui peuvent fort bien contenir la phrase « Le métro de Londres est-il plus cher que celui de Paris ? », mais pas la question « Le métro de Paris est-il plus cher que celui de Londres ? ». Pour un lecteur qui n'a aucune notion de la syntaxe du français, la première phrase, grâce au guide, devient dicible, la seconde demeure indicible. En remplaçant « dicible » par « pensable », on obtient une illustration de non-systématicité de la pensée.

ce qui procède chez moi de la saisie du sens des symboles correspond dans le système à une position nodale dans un réseau de dispositions.

Cette explication appelle trois remarques. La première, d'ordre pédagogique, est qu'il y a quelque chose de trompeur dans le choix de l'exemple : il se trouve que multiplier 31 par 12 est une opération formelle gouvernée par des règles, et que pour trouver le résultat, nous appliquons un algorithme à peu près comme le fait une calculette ou un ordinateur (et ce n'est pas fortuit : les machines ici *imitent* l'esprit de celui qui calcule[1]). Mais c'est là un cas limite : dans leur immense majorité, les processus cognitifs n'ont pas ce caractère. La force de l'HLP est d'affirmer que la perception, la mémoire, la compréhension des mobiles d'autrui, la communication linguistique, l'apprentissage du piano, la recherche scientifique, la navigation dans le métro de Tokyo, toutes tâches qui n'ont pas l'apparence de procédures algorithmiques effectuées par le sujet conscient, s'accomplissent grâce à des processus cognitifs de même nature que ceux qui sous-tendent la multiplication de 31 par 12. Contrairement à ce qu'on lit souvent, l'HLP ne prétend donc pas que les processus mentaux sont formels, mais que les mécanismes qui rendent compte de ces processus le sont.

En deuxième lieu, il faut reconnaître que la manière dont la saisie du sens est expliquée reste obscure. La difficulté est double. Il faut, d'une part, comprendre l'intentionalité comme un phénomène naturel ; or c'est un problème qui, de l'avis général, demeure largement ouvert. Il faut, d'autre part, comprendre d'où viennent les concepts, qui sont, dans l'HLP, les sens des symboles du mentalais (dans notre exemple, le concept de multiplication, les concepts de 31, de 12 et de 372 ; il faut aussi considérer les symboles logiques). Pour des raisons qu'il n'est pas possible de développer ici, l'HLP incline fortement vers l'innéisme : les concepts primitifs du mentalais seraient innés. Toute raison de rejeter l'innéisme met en cause l'HLP et amène à se demander dans quelle mesure il est possible d'en conserver une partie sans s'engager en faveur d'une forme franche d'innéisme.

Enfin, une question importante est celle des rapports qui existent entre les concepts primitifs du mentalais (ou, de manière plus générale, les unités sémantiques de base) et les concepts du sens commun, et plus généralement ceux qui s'expriment dans la langue naturelle. Nous allons voir que c'est là un point nodal sur lequel les chercheurs se divisent.

2.2.4 *Sous la vie mentale consciente*

Les premiers exemples de processus mentaux qui viennent sous la plume de celui qui veut présenter la TCE et l'HLP font intervenir, on l'a vu et on peut comprendre

1. Ce que Turing, dans l'article princeps de 1937 où il pose les bases de la théorie des ordinateurs, appelle le « computor ». Un autre exemple qui est souvent choisi est celui du jeu d'échecs, où sont mis en scène, d'une part, le joueur humain, d'autre part, le programme informatique. Il présente le même caractère d'évidence trompeuse.

pourquoi, des notions de sens commun, en suggérant que ces processus peuvent être expliqués par des opérations du système cognitif mettant en jeu des « pré-concepts », termes de mentalais, qui reflètent fidèlement les concepts présents et consciemment déployés au cours de l'épisode considéré de la vie mentale du sujet.

Ce choix d'exemples comporte un double inconvénient. On a dit un mot du premier : il encourage le paralogisme de l'homuncule ; c'est là un problème conceptuel. Le second inconvénient est d'ordre plus empirique : il détourne l'attention d'une possibilité cruciale. Avant de l'exposer, précisons que le problème n'est pas seulement pédagogique. La première phase de l'IA et de la psychologie cognitive ont beaucoup fait pour accréditer le projet d'une explication de la vie mentale par des processus se situant au même niveau sémantique, et il demeure, au sein des sciences cognitives, une tension entre une conception « homosémantique » et une conception « hétérosémantique ». La terminologie n'est pas standard, mais voici ce dont il s'agit.

L'idée remonte à loin, et elle est périodiquement oubliée puis redécouverte en philosophie et en psychologie. Déjà pour Leibniz, par exemple, les mouvements visibles de l'esprit s'expliquaient par une dynamique de « petites perceptions » ; les philosophes écossais William Hamilton et Alexander Bain, le grand physicien, physiologiste et psychologue allemand Helmholtz, le neuropsychologue américain Karl Lashley, ont chacun à leur manière compris qu'une bonne partie des processus cognitifs ne sont ni conscients ni aisément décrits dans le vocabulaire conceptuel ordinaire, fût-ce au prix de raffinements. Comme l'écrit Bain en 1893 : « L'expression manifeste, si serrée et consécutive qu'elle puisse paraître, n'en est pas moins une succession de bonds, de glissés et de sauts. Elle ne fournit pas la suite complète des mouvements mentaux[1]. » Qu'on se place sur le plan temporel et causal ou sur le plan rationnel de l'enchaînement des idées, la suite des pensées consciences est incomplète ; il semble nécessaire de postuler, à un niveau plus profond, une trajectoire connexe dont certains « pics » émergent pour former l'« expression manifeste » de Bain.

Cette intuition ne vaut pas une analyse, moins encore une théorie appuyée sur l'expérimentation. Elle s'exprime, on le voit, sous une forme métaphorique. Elle n'en est pas moins à mon sens la source de la troisième idée fondamentale des sciences cognitives (les deux premières étant celle d'information ou de représentation comme propriété relationnelle de composants d'un système matériel, et celle de computation comme modalité mécanique abstraite). C'est peut-être la plus originale et la plus féconde. Elle prend chez les théoriciens contemporains des formes diverses, non nécessairement compatibles, voire défendues par des écoles qui peuvent s'opposer durement. Par-delà ces différences, on peut discerner un noyau commun, possédant

1. « *Outward expression, however close and consecutive, is still hop, skip and jump. It does not supply the full sequence of mental movements.* » Je dois les références à Hamilton (1859) et à Bain, ainsi que la citation de ce dernier, à un chapitre de Martin Davies homonyme de celui-ci (Davies, 2005). Sur Hamilton, on dispose en français de Dupont (2007).

deux composantes. La première thèse est que le niveau auquel se produisent les enchaînements causaux réels responsables de la cognition est disjoint de la conscience. La seconde thèse est que les entités et processus, à ce niveau, sont doués d'un contenu sémantique qui est d'un grain plus fin que celui des significations ordinaires, présentes à la conscience et dans la langue. Pour mettre un nom sur la première thèse, on peut emprunter la notion proposée par Chomsky de « connaissance tacite » de la grammaire. Pour la seconde thèse, on peut penser au niveau « sub-personnel » de Dennett (1978), aux états et processus « sub-doxastiques » de Stich (1978 et 1983) , ou encore à la « microstructure » de la cognition que les théoriciens du connexionnisme (dont il sera question dans un moment) veulent mettre au jour (Rumelhart & McClelland, 1986 ; Smolensky, 1987).

Ces approches, on l'a dit, sont différentes, mais on retrouve dans chacune d'elles trois hypothèses : celle d'un niveau sous-jacent qui explique la formation des pensées et démarches conscientes ; celle d'une différence radicale avec les attitudes propositionnelles ordinaires ; celle, enfin, d'une nature informationnelle ou représentationnelle des entités du niveau en question : les états et processus de ce niveau ne sont pas directement physiques (ce ne sont pas, directement, des états et des processus neurophysiologiques[1]).

2.3 Le rôle fondamental mais limité des modèles dans la recherche de fondements

Pour un lecteur déjà familier des recherches en cours dans les sciences cognitives, ou pour celui qui entrerait par hasard dans un laboratoire actif dans le domaine, ce qui vient d'être exposé peut sembler très éloigné des questions actuellement étudiées. Ce sentiment, justifié, a plusieurs causes. La première, très générale, est que la quête philosophique de fondements n'est pas directement pertinente pour la recherche scientifique. La deuxième est que les choses évoluent très vite, et que beaucoup de recherches échappent au cadre que les philosophes se sont efforcés de donner à l'ensemble de l'entreprise. Ces nouveaux courants sont accompagnés par des groupes de philosophes qui veulent ébaucher d'autres cadres, mais comme on le verra ces efforts restent très dispersés, annonçant d'ailleurs peut-être la fin du projet unitaire. Enfin, comme on l'a indiqué d'entrée de jeu, les sciences cognitives restent très incertaines sur la nature et l'extension de leur objet, et cette incertitude persistante donne à la philosophie un rôle plus important que d'ordinaire dans le débroussaillage de la situation proprement scientifique ; elle jouit d'une autonomie inhabituelle, comparable à celle des disciplines positives, et développe ses propres idées sans toujours se référer aux programmes de recherche en cours, tandis que

1. Nous retrouvons le problème terminologique mentionné à la note de la page 545. Tout état ou processus particulier est physique (neurophysiologique) sur le plan de sa nature ; mais ses propriétés pertinentes sont celles d'une classe d'entités fonctionnellement semblables, et s'énoncent dans un autre vocabulaire.

ceux-ci poursuivent leur trajectoire sans se soucier du cadre dans lequel ils sont censés trouver place. C'est pourquoi une articulation entre philosophie et sciences positives de la cognition est indispensable. Cette articulation est assurée par des *modèles*. Il ne s'agit pas ici de discuter du rôle des modèles dans les sciences en général, et la question de savoir si le terme recouvre ou non des choses très différentes sera laissée de côté. Dans les sciences cognitives, il y a comme ailleurs différentes sortes de modèles, et le terme est doté d'une élasticité considérable. Mais il a aussi un emploi bien particulier, et le dispositif théorique dans lequel il s'insère est d'une importance décisive.

2.3.1 *Modèles classiques, connexionnistes, dynamiques*

Si l'ordinateur n'avait pas existé, il est très difficile d'imaginer dans quel horizon théorique les sciences cognitives auraient pris leur essor, et ce qu'elles seraient aujourd'hui. Le rôle de l'ordinateur en la circonstance est souvent mal compris, donnant lieu à des critiques aussi faciles qu'injustifiées. L'ordinateur a été d'abord conçu, par Turing, comme un modèle de l'homme calculant (le « computor » déjà mentionné à la note de la page 552) : Turing identifie des aspects déterminants du processus réel et crée une structure formelle constituée d'éléments et de relations représentant ces aspects et leurs interactions. Il s'agit à ce stade d'un modèle abstrait, comme le sont les systèmes différentiels en physique. Puis les premiers ordinateurs matériels voient le jour ; ils incorporent le schéma de Turing, démontrant ainsi sa cohérence et soutenant ses hypothèses de modélisation. Mais ils reflètent également d'autres choix théoriques, d'inspiration technologique ou logico-mathématique et non psychologique, qui en retour suggèrent des hypothèses supplémentaires importantes sur le « computor ». Ce sont ces choix qui conduisent à l'architecture dite de von Neumann, qui constitue encore aujourd'hui le patron des ordinateurs tout-venant. Bientôt, Turing et d'autres proposent de voir dans l'ordinateur un modèle de la pensée humaine en général, cette fois-ci dans le sens pratique et non théorique de « modèle », quelque chose de comparable à une maquette ou à un modèle réduit. Enfin, une expérimentation sur ce modèle, et un réexamen de ses principes de construction, conduisent à modifier et à enrichir considérablement le modèle théorique de départ.

C'est donc autour de ce processus complexe de modélisation (dans lequel les modèles sont alternativement abstraits et concrets) que se sont élaborés conjointement (« co-construits ») un cadre général pour les sciences cognitives et une famille de systèmes matériels incorporant ce cadre et, le cas échéant, permettant de mettre à l'épreuve des hypothèses formulées dans ce cadre. Nous donnerons dans un instant (voir *b. infra*) un sens précis à ce double mouvement. Je parle maintenant d'une *famille* de systèmes plutôt que de *l'*ordinateur au singulier, pour deux raisons : d'abord, comme chacun sait, il n'existe pas un seul type d'ordinateur, mais une grande variété, qui ne diffèrent pas seulement par les paramètres connus du

grand public (vitesse du processeur central, mémoire vive, mémoire morte…) mais par leur architecture au sens informatique du terme ; en second lieu, un ordinateur est nécessairement doté d'un langage de base, ou système d'exploitation, qui en fait une machine particulière, différente du même ordinateur doté d'un autre système d'exploitation (et en réalité chaque spécification additionnelle, sous la forme d'un langage d'ordre supérieur, introduit une nouvelle différence). Il est vrai que toutes ces machines ont tant en commun qu'il est souvent légitime de les regrouper sous un seul chapeau ; on peut même arguer qu'elles ne sont que différentes façons de réaliser un système matériel de calcul, au sens logico-mathématique du terme, ce qui leur confère une identité unique. Mais la simple considération de la finitude des ordinateurs réels montre qu'ils diffèrent du modèle idéal de la machine de Turing, et suggère que la manière dont ils diffèrent d'elle peut introduire entre eux des différences ayant une signification théorique. De manière plus générale, les *conditions aux limites* de fonctionnement d'un ordinateur particulier, résultant des nombreuses décisions architecturales prises par ses concepteurs, mais aussi ses conditions d'utilisation et la manière dont on interprète ses résultats, constituent des caractéristiques qui peuvent compter autant que sa fonction calculatoire originelle[1].

Venons-en à un deuxième cadre pour les sciences cognitives, résultat d'un processus de co-construction très semblable à celui qui a conduit au cadre lié à la machine de Turing. Quoique son élaboration soit à peu près contemporaine, il est parvenu à maturité plus tard. Cela explique que le cadre turingien soit souvent appelé « classique » ; on l'appelle aussi parfois « symbolique », par référence aux symboles postulés par l'HLP. Le deuxième cadre est généralement appelé « connexionniste », nous allons comprendre pourquoi. Le connexionnisme trouve son origine dans une tentative, faite au début des années 1940, de modélisation de l'unité fonctionnelle de base du cerveau, telle qu'on pouvait la concevoir à l'époque (et que le psychologue canadien Donald Hebb appellera des « assemblées de neurones » : Hebb, 1949). L'hypothèse était qu'une telle unité est constituée d'un réseau de neurones qui se transmettent, par le canal des *connexions* synaptiques, des impulsions électriques. Les auteurs du modèle, Warren McCulloch et Walter Pitts (membres du groupe qui créa la cybernétique[2]), partaient d'une conception schématique du neurone (le « neurone formel ») et des réseaux que forment les neurones pour montrer que ces réseaux sont capables d'effectuer les calculs logiques de base, et partant toute espèce de calcul (McCulloch & Pitts, 1943 ; Anderson & Rosenfeld, 1988). Ce mouvement est en un sens symétrique de celui de Turing, qui part d'une conception schématique du calcul et conçoit une machine capable d'exécuter ce schéma.

Aujourd'hui, les réseaux de neurones formels constituent une famille de systèmes matériels qui jouent vis-à-vis du connexionnisme le même rôle que les ordinateurs vis-à-vis du classicisme. Ils incorporent des hypothèses fondamentales

1. Une illustration amusante (mais superficielle) en est fournie par l'épisode du « 2KY bug ».
2. Groupe qui fit de Turing un « membre d'honneur » (Heims, 1991).

quant à la nature de la cognition, hypothèses qui forment un cadre au sein duquel des hypothèses plus spécifiques peuvent être formulées et, en un sens, testées sur les réseaux connexionnistes. En retour, ceux-ci suggèrent des modifications ou bien des hypothèses entièrement nouvelles. Inversement, les théories issues des sciences cognitives suggèrent des principes architecturaux pour la conception des réseaux : la variété des suggestions concevables est ici plus grande que dans le cadre classique, d'une part, en raison de la grande diversité d'architectures possibles pour les réseaux, d'autre part, parce que les hypothèses neuroscientifiques peuvent s'appliquer, au même titre que les hypothèses psychologiques, dans le processus de coévolution de la théorie psychologique et des modèles computationnels. Selon que l'on accorde une place plus grande aux premières qu'aux secondes, ou l'inverse, on se place dans un courant neuroscientifique, ou au contraire psychologique, au sein du connexionnisme. Ce dont les réseaux connexionnistes sont des modèles s'accorde de conceptions très diverses de ce qui constitue l'objet des sciences cognitives (on y reviendra bientôt).

Le cadre connexionniste ne peut être décrit ici même de manière sommaire (Hinton & Anderson, 1981 ; Rumelhart & McClelland, 1986 ; Smolensky, 1987 ; Amit, 1989 ; Anderson, Pellionisz & Rosenfeld, 1990 ; Clark, 1989 ; Andler, 1990 ; Dayan & Abott, 2001). On peut cependant commencer à le situer par rapport au cadre classique à l'aide d'une série d'oppositions. Les processus de traitement de l'information sont, dans le cadre classique, essentiellement séquentiels ; dans le cadre connexionniste, massivement parallèles. L'opération fondamentale est, dans le premier cas, l'inférence, ou encore des processus gouvernés par une règle explicite ; dans le second, l'association, guidée par des mesures continues de distance. Les représentations internes classiques sont symboliques et locales (c'est-à-dire que chaque symbole représente à lui seul un concept et un seul) ; les représentations connexionnistes sont souvent sub-symboliques et distribuées (chaque support représentationnel ne représente rien à lui seul, les concepts étant représentés par des ensembles de supports, ce qui ne laisse à chacun qu'une valeur « micro-représentationnelle » susceptible d'entrer dans une pluralité de représentations). Le classicisme repose sur une distinction nette entre connaissances (les valeurs des variables, dans un programme) et opérations (la suite des instructions du programme), le connexionnisme mêle les deux. Enfin, l'apprentissage, dans le cadre classique, se réduit à l'acquisition discrète de nouvelles connaissances, alors qu'il se présente très naturellement comme une forme d'adaptation graduelle dans le cadre connexionniste.

Mais ce ne sont là que des contrastes très généraux qui ne présentent qu'une image simpliste de la situation. La question des rapports entre les deux cadres est complexe. Aucun d'entre eux n'étant très contraignant, et chacun admettant une grande variété d'interprétations, plusieurs façons d'envisager ces rapports ont été élaborées, allant de l'incompatibilité totale à la compatibilité complète, en passant par différentes

positions intermédiaires, et incluant notamment le partage de compétences, le cas limite (selon le principe de correspondance de Bohr[1]) et l'émergence.

L'apparition tardive du connexionnisme dans le présent chapitre, et la place très modeste qui lui est dévolue peuvent conduire le lecteur à deux supputations : que le connexionnisme joue un rôle secondaire aujourd'hui dans les sciences cognitives, ou que l'auteur du chapitre n'en voit pas l'intérêt. C'est le contraire qui est vrai, dans les deux cas. L'explication est d'ordre éditorial. D'une part, il fallait faire des choix : on ne peut parler de tout en détail dans un simple chapitre. D'autre part, il est difficile d'exposer le connexionnisme sans parler du classicisme, alors que l'inverse est vrai.

Mais il faut évoquer un troisième couple candidat, beaucoup plus récent, nommé parfois le « dynamicisme » (Thelen & Smith, 1994 ; Port & Van Gelder, 1995 ; Ward, 2001). La famille de systèmes physiques de référence est ici constituée par les systèmes dynamiques, compris au sens de la théorie mathématique du même nom, c'est-à-dire des systèmes matériels évoluant dans le temps, dont l'état est caractérisé à un instant donné par les valeurs, en général réelles, d'un ensemble de variables et dont les trajectoires sont déterminées par un système d'équations, en général différentielles. C'est là une classe immense dont font partie toutes sortes de systèmes, depuis le système solaire, le système météorologique terrestre ou l'économie mondiale jusqu'aux gyroscopes, aux ordinateurs et aux réseaux connexionnistes, vus sous une description adéquate. Le dynamicisme a en vue certains systèmes particuliers, sur lesquels la cybernétique avait mis l'accent, possédant notamment des propriétés d'autonomie ou d'auto-régulation assurées par des boucles de rétroaction. Ces systèmes sont typiquement des systèmes de contrôle : le thermostat est un exemple particulièrement rudimentaire, le régulateur de Watt un exemple plus riche. Certains robots, construits selon les principes du cadre dynamiciste, sont des illustrations plus explicites de systèmes cognitifs considérés sous l'angle dynamique ; ils peuvent être vus comme des systèmes de contrôle lorsqu'ils sont placés dans un environnement sur lequel ils agissent.

Quant au cadre dynamiciste, il est de loin le moins développé des trois, et il n'est pas sûr, dans son état présent d'élaboration et de ses choix théoriques, qu'il soit appelé à jouer un rôle durable. Ses principaux points d'opposition avec le cadre classique sont les suivants. (i) Il rejette tout recours aux représentations internes. De manière concomitante, il conçoit les rapports entre système cognitif et environnement sur le modèle du couplage et du contrôle, et non de la représentation et de l'action. (ii) Il accorde à la temporalité des processus une importance cruciale, alors que le cadre classique n'y voit que l'effet de la succession des opérations, entraînant des contraintes qui peuvent être importantes, mais ne constituant pas une détermination fondamentale. Une caractéristique centrale de cette temporalité

1. Principe selon lequel une nouvelle théorie (telle que la relativité restreinte) doit subsumer une approximation de l'ancienne (telle que la dynamique newtonienne), qui apparaît à son tour *a posteriori* comme une approximation d'un cas particulier de la nouvelle.

est qu'elle est continue : le système interagit continûment avec l'environnement, alors qu'un système classique reçoit des informations à des moments discrets, évolue selon un protocole discret, et exécute une suite discrète d'actions. (iii) Il épouse un holisme radical, inspiré notamment par la *Gestalttheorie* (Koffka, 1935 ; Köhler, 1945 ; Kanizsa, 1997 ; Smith & Ehrenfels, 1989). Selon ce point de vue, seules sont significatives les configurations du système et du couplage système-environnement, et non tel ou tel élément ou aspect distingué : pris isolément, aucun élément simple n'a de signification, la notion même d'élément simple, ou de base, constituant le germe d'une erreur fondamentale.

Dans ce contraste, le connexionnisme occupe une position intermédiaire : il rejoint, dans certaines de ses versions les plus intéressantes, une partie, la plus solide à mes yeux, du programme dynamiciste, sans l'épouser complètement (ce qui l'amènerait à renoncer à une bonne partie de ce qui fait sa fécondité) : il met en question la conception classique de la représentation, sans rejeter l'idée que la représentation est essentielle pour la cognition ; il adopte la perspective des systèmes dynamiques, faisant du temps une dimension essentielle ; il favorise un certain holisme.

Certains portent sur le dynamicisme un jugement tout différent de celui que je formule ici sommairement : ils estiment que le connexionnisme ne va pas assez loin dans son rejet des hypothèses classiques, et que seul le dynamicisme offre une réelle possibilité d'échapper à ce qu'ils voient comme les limitations rédhibitoires, voire les incohérences du classicisme.

2.3.2 *Préciser et diversifier les options théoriques*

Mais de quelle manière les « grands modèles » (ceux dont on vient de parler) contribuent-ils effectivement aux recherches en sciences cognitives ? La question peut paraître déplacée : ne vient-on pas d'y répondre longuement ? Elle nous invite pourtant à y aller voir de plus près.

Commençons par le cadre classique. On le présente souvent comme découlant de la « métaphore de l'ordinateur », métaphore qui serait aussi peu pertinente que possible, vu que le système nerveux central n'est d'aucune manière raisonnable comparable à un ordinateur. C'est là commettre un contresens précisément sur le rôle joué dans les sciences cognitives par le grand modèle qu'est l'ordinateur.

En réalité, ce rôle est triple. *Primo*, il fournit une détermination concrète précise des concepts théoriques employés dans la psychologie cognitive naissante ; pour le dire en un mot, l'ordinateur sert de preuve d'existence (ou, ce qui revient au même, de cohérence) et permet de *fixer les idées*. Prenons l'idée très générale de système formel, en partant de la notion d'origine leibnizienne de « pensée aveugle ». Peut-on concevoir une « machine syntaxique » qui rende les services d'une « machine sémantique » idéale, c'est-à-dire capable d'éviter les multiples pièges du langage, de la pensée et de la perception ordinaires ? Certainement, fût-ce au prix d'un long cheminement, d'Aristote à Turing en passant par Frege, Russell, Gödel… Mais est-on

sûr que la proposition théorique à laquelle on aboutit est libre de contradictions (des contradictions ne sont-elles pas apparues dans des théories dont la rigueur abstraite et l'apparente simplicité semblaient garantir la cohérence ?) ? Est-on sûr que cette proposition peut se réaliser dans le monde matériel que nous connaissons sous la forme d'un système physique ? Un système physique n'est-il pas voué à ne produire que des réactions réflexes, pouvant aller, peut-être, jusqu'aux opérations de l'arithmétique élémentaire, mais pas au-delà ? Il est tout à fait remarquable que Turing parvienne à mettre un terme final à ces doutes, et que c'est dans sa tentative, couronnée de succès, pour déterminer les *limites* de la pensée formelle ou mécanisable qu'il montre son étendue *illimitée*. Pour prendre un autre exemple, l'idée générale que notre réaction à une situation donnée dépend de notre propre état au moment considéré reste flottante ; en la rapportant à la notion technique précise d'état interne d'une machine de Turing (notion que son inventeur éclaire lui-même par une comparaison avec les modes « majuscule » ou « minuscule » d'une machine à écrire[1]), on se donne une prise ferme qui permet de progresser dans la réflexion conceptuelle, sans être rivé au modèle.

Secundo, le grand modèle de l'ordinateur est la source de concepts, de distinctions et d'hypothèses que la psychologie et, plus largement, les sciences cognitives peuvent chercher à exploiter. Les exemples ne manquent pas. La notion de « valeur par défaut » est d'origine informatique ; elle fait partie du vocabulaire de base des sciences cognitives. De même pour la « mémoire vive » (qui donne naissance aux notions de mémoire à court terme et de mémoire de travail), l'idée de mémoire « adressable par le contenu », ou la notion de « contrôle central ». Ou encore la notion d'« heuristique » introduite par Herbert Simon dans le contexte de la prise de décision et transférée par lui au domaine de l'IA, où elle prend un sens précis et peut de là migrer vers les sciences cognitives. Remarquons que beaucoup de ces notions ont également envahi le langage commun : l'omniprésence de l'outil informatique produit des effets dans la « théorie naïve » des processus mentaux (la notion de théorie naïve est la généralisation de la « physique naïve », un autre concept forgé par l'IA). D'autres transferts vers les sciences cognitives sont plus locaux et plus techniques, par exemple en théorie de la vision, et la place manque pour les présenter. L'apport du modèle informatique, sur ce plan, est cependant disputé : peu probant pour certains, il est selon d'autres d'une importance essentielle.

C'est dans son troisième rôle que l'utilité du modèle est le moins contestable. L'ordinateur est pris comme un terrain d'expérimentation : expériences au sens littéral comme y insistaient les fondateurs de l'IA, quoique d'un genre particulier, expériences par la pensée, également d'un genre particulier, comme les pratiquent davantage les sciences cognitives d'aujourd'hui.

---

1. Ancêtre mécanique puis électromécanique des traitements de texte contemporains ; a laissé pour trace le clavier des ordinateurs, et une grande nostalgie à la génération déclinante.

Voyons d'abord en quel sens l'ordinateur permet aux sciences cognitives de faire des expériences réelles. Pour l'IA des débuts, un programme d'ordinateur lui permettant d'accomplir une tâche cognitive qui, chez l'homme, résulte de l'exercice de la capacité cognitive C constituait, littéralement, une théorie de C relevant de plein droit de la psychologie scientifique (pour fixer les idées, prenons pour C la capacité de lire un texte à haute voix, ou bien la capacité de résoudre une certaine famille de problèmes géométriques, ou encore la capacité d'empiler des blocs de taille différente en sorte que la pile soit stable). Donc si le psychologue formule une conjecture T relative à la capacité C, il peut (et, selon certains, il doit) traduire T en un programme P et mesurer le degré de succès que P remporte dans l'accomplissement de C ; un échec peut amener le psychologue à rejeter T, ou, si l'échec n'est que partiel, à modifier T en T', puis à traduire T' en un programme C' qui sera testé à son tour. Voilà pour l'ordinateur comme « laboratoire » de sciences cognitives. Pour diverses raisons, cette démarche a été pratiquement abandonnée, sauf dans certains domaines particuliers, mais elle conserve, au moins, une valeur heuristique et constitue un schéma qui sera repris dans d'autres cadres.

C'est finalement comme terrain d'expériences de pensée que l'ordinateur est aujourd'hui le plus utile pour les sciences cognitives. Lorsqu'un chercheur cherche à expliquer une capacité cognitive, s'il épouse le cadre classique il proposera de décomposer cette capacité, comprise comme une transformation d'informations, en capacités plus élémentaires, et celles-ci à leur tour en capacités plus simples encore, jusqu'au point où il aura réduit la capacité d'origine à une combinaison de capacités dont il est moralement certain qu'elles sont réalisables mécaniquement. Il est, en général, impraticable de traduire cette décomposition en un modèle mécanique explicite complet. L'expérience de pensée consiste à se demander si un ordinateur pourrait être programmé en conformité avec la décomposition proposée, et si ainsi programmé il obtiendrait le résultat requis. Comme toute expérience de pensée, une démarche de ce genre n'a de valeur probante qu'entre les mains d'un chercheur expérimenté : l'ordinateur sert de « discipline », décourageant les fausses solutions.

Mais un tout autre type d'expérience de pensée est également envisageable. Soit à nouveau une capacité cognitive C dont on cherche à rendre compte. Supposons que l'on soit parvenu, par un ensemble convergent d'arguments, à la conviction que toute décomposition possible réalisable sur un ordinateur d'une certaine architecture présenterait des caractéristiques qui ne sont pas observées sur C. Alors on dispose d'un argument en faveur du rejet de cette architecture comme modèle de l'esprit (ou peut-être seulement comme modèle de ce genre de capacité). Si l'on parvient à une conclusion plus forte, à savoir qu'aucune décomposition réalisable sur un ordinateur, quelle que puisse être son architecture, ne répond aux principales caractéristiques observables de C, alors on dispose d'un argument contre le cadre classique ou symbolique lui-même.

Et c'est à ce point que l'utilité théorique des grands modèles est peut-être la plus claire. Si la capacité C n'est pas réalisable dans une architecture classique, et s'il en est

d'autres concevables, on peut chercher à réaliser C dans ces autres architectures (et à reconceptualiser C en conséquence). Le connexionnisme ainsi, malgré sa fragilité relative, que le dynamicisme se présentent comme des solutions de rechange au classicisme. C'est ainsi que beaucoup de théories particulières postulent une réalisation connexionniste, sans aller jusqu'à une modélisation effective, ni nécessairement la présenter comme un schéma de fonctionnement neural. De manière générale, l'existence de grands modèles concurrents permet de formuler avec une précision inédite en psychologie toute une série de questions allant du niveau le plus local au niveau le plus général. Parmi les questions locales, les cadres classique et connexionniste conduisent à des conceptions radicalement opposées (du moins en première analyse) de la mémoire, de la reconnaissance des formes, de l'acquisition des règles morphologiques dans les langues naturelles (un exemple qui a donné lieu à une controverse célèbre est celui de l'apprentissage, par l'enfant, du passé des verbes de l'anglais), de la formation des concepts, etc. Au niveau intermédiaire, c'est le format de représentation des connaissances, le rôle des règles dans la cognition, la nature de l'apprentissage qui sont en jeu. Au niveau supérieur, s'affrontent des conceptions différentes de la cognition. Le classicisme place la logique au centre de la cognition, le connexionnisme place la perception, et le dynamicisme le mouvement. La cognition est, dans le cadre classique, essentiellement informationnelle ; dans le cadre connexionniste, elle est comprise comme une fonction informationnelle de systèmes ayant la forme très particulière des structures corticales ; dans le cadre dynamique, comme un couplage évolutif avec l'environnement.

Comment choisit-on un cadre plutôt qu'un autre ? C'est l'une des principales questions de la philosophie des sciences cognitives, et elle est liée aux autres grandes questions de multiples manières. Sa difficulté vient de deux sources principales. L'une est que les différences intrinsèques entre les grands modèles, on l'a vu, ne sont pas une donnée de fait, mais constituent une question ouverte, dont la résolution ne peut venir que d'un effort à la fois philosophique et scientifique qui n'a pas encore abouti. L'autre est qu'on ne peut s'appuyer, comme on pourrait être tenté de le faire, sur l'arbitrage des sciences cognitives telles qu'elles se font. On pourrait penser que ces cadres ont pour pierre de touche l'adéquation au domaine dont ils prétendent révéler la structure fondamentale : en proposant des hypothèses de très grande généralité sur ce qu'est la *cognition*, ils s'offrent comme reconstructions rationnelles des *sciences cognitives*, conçues comme l'ensemble des travaux empiriques locaux portant sur différents aspects, à différents niveaux de description, de différentes fonctions cognitives particulières. Le cadre qui subsume ces travaux de la manière la plus satisfaisante pourrait alors être déclaré vainqueur, de manière révisable comme toujours dans les sciences. Or ce qui compte comme un résultat ou comme un programme de recherche admissible dans les sciences cognitives n'est pas une donnée, mais une hypothèse qui se place dans un ensemble d'hypothèses dont celle du cadre général. En d'autres termes, le cadre détermine (en partie) ce qui compte comme un résultat ou une théorie, on ne peut donc partir des résultats et des théories pour trancher la

question du choix du meilleur cadre. C'est donc, dans le meilleur des cas, au terme d'un long cycle d'allers-retours entre hypothèses de haut niveau, théories de niveau plus local, résultats empiriques que se stabiliseront, simultanément et solidairement, le cadre, la conception de l'objet des sciences cognitives et de la structure de ses théories, et le corpus de ses concepts et résultats fondamentaux.

Fort heureusement pour les sciences cognitives, le choix du cadre n'est pas un préalable, pour une raison que nous allons maintenant examiner.

2.3.3 *Tout ce qui reste à déterminer : l'incomplétude des grands modèles*

Imaginons une psychologue du développement qui cherche à rendre compte de la manière dont un très jeune enfant se rend maître d'une capacité, d'un concept, d'un savoir-faire particulier. Imaginons un neurolinguiste qui veut comprendre pourquoi certains déficits linguistiques massifs, consécutifs à un infarctus cérébral, disparaissent spontanément ; pourquoi d'autres s'atténuent sous l'effet d'une thérapie, pourquoi enfin certains sont irréversibles. Imaginons un psychologue qui s'interroge sur la dépendance, suggérée par certaines pathologies, entre capacité de navigation et conscience autobiographique. Imaginons un neurophysiologiste qui se demande comment le système visuel peut suivre la trajectoire de plusieurs objets simultanément. Imaginons un psychophysicien qui veut améliorer l'audition des sourds profonds à l'aide de meilleurs implants cochléaires. Imaginons un linguiste qui veut comprendre quels indices permettent d'attribuer les bonnes valeurs référentielles à certains pronoms dans des phrases d'un certain type (« il » dans « Le chat a mangé le bifteck parce qu'il était affamé » / « Le chat a mangé le bifteck parce qu'il était appétissant » ; ou dans « Pierre demande à Jean s'il croit vraiment qu'il aime Julie » / « Pierre explique à Jean qu'il croit vraiment qu'il aime Julie »). Imaginons un informaticien chargé de concevoir un logiciel d'aide à la décision pour les agents de sécurité des centrales nucléaires. Imaginons un anthropologue qui étudie les croyances surnaturelles et leur coexistence avec les croyances communes. Imaginons un économiste qui cherche à compenser les biais cognitifs du sujet moyen pour l'orienter vers une conduite propice à ses intérêts à long terme, par exemple en matière de retraite ou de sécurité routière. Imaginons un philosophe qui se demande si une image perçue, une image imaginée et une image remémorée sont de même nature.

Comment ces chercheurs vont-ils procéder ? Ils n'ont rien à espérer, du moins au début de leur enquête, des grands modèles et des cadres qui leur sont associés, pour la raison simple que ceux-ci sont absolument muets sur les questions qui les occupent. Ils ne peuvent que se pencher, en psychologue, linguiste, informaticien, neurobiologiste, anthropologue, économiste, philosophe, sur le phénomène lui-même, en poursuivant toutes les pistes suggérées par leur propre tradition disciplinaire, mais en tirant aussi parti (selon le principe organisationnel de base des sciences cognitives) des indices fournis par les autres disciplines. Les grands modèles parlent surtout des processus mentaux. Si la question des processus est importante (comme l'a longtemps souligné Fodor, alors que, selon lui, la tradition philosophique et psychologique

les avait négligés), les sciences cognitives naissantes ont, à l'inverse, eu tendance à sous-estimer la difficulté de la question des états mentaux et de leurs contenus spécifiques. En mûrissant, elles se sont intéressées à des capacités de plus en plus spécifiques ou « domaniales », concernant les nombres ou autrui, la notion d'objet ou les anaphores, la dyslexie ou la perception du mouvement, et les états mentaux sont revenus sur le devant de la scène, réléguant au second plan les processus et, partant, les grands modèles.

Toujours est-il que la plupart des chercheurs en sciences cognitives sont en général indifférents à la question du cadre, qu'ils traitent un peu comme Newton faisait pour la gravité : *Hypothesis non fingo*. Les questions qui les occupent ne sont pas sans rapport avec les hypothèses générales constitutives des grands modèles. Mais les connexions sont généralement lâches, et quand elles se resserrent, ce n'est que graduellement et de manière réversible. Pour le dire rapidement, les sciences cognitives opèrent le plus souvent selon un régime libre, affiché ou tacite, ne s'engageant pas réellement sur la question des fondements.

Pour illustrer ce régime, prenons un ou deux exemples, choisis parmi ceux qui suscitent chez le philosophe des sciences un intérêt en eux-mêmes. Il y a plus de trente ans, deux psychologues spécialistes des grands singes posèrent la question suivante : les chimpanzés possèdent-ils une « théorie de l'esprit » (TdE) (Premack & Woodruff, 1978) ? En d'autres termes, sont-ils, comme nous, capables d'attribuer à un congénère des croyances, des désirs, des intentions qui lui soient propres et diffèrent, le cas échéant, des leurs ? Cette question a donné naissance à un programme de recherche sur l'humain : quelle est cette capacité ? Quels en sont les mécanismes psychologiques ? Dépend-elle, en particulier, de notre capacité à comprendre nos propres attitudes propositionnelles, ou bien celles-ci ne nous sont-elle accessibles que par le même canal que celles d'autrui ? Quelles sont les bases neurales de la TdE ? Est-elle immergée dans une capacité plus générale, telle qu'une « psychologie naïve[1] » conçue comme une théorie plus ou moins tacite du fonctionnement de l'esprit, ou bien se limite-t-elle à l'identification des attitudes propositionnelles d'un congénère ? Possède-t-elle les caractéristiques d'un module au sens de la modularité massive ? À quel âge et comment est-elle acquise par l'enfant ? L'autisme est-il marqué par une TdE absente ou déficiente, et ce déficit est-il une cause ou une conséquence d'autres aspects du syndrome ? En particulier, la « cécité mentale » (l'incapacité présumée de l'autiste à voir en l'autre une entité munie d'un esprit, contrairement aux pierres et aux camions) est-elle la cause de son incapacité à établir des relations sociales ? Chez l'enfant normal, inversement, la théorie de l'esprit est-elle nécessaire, est-elle suffisante pour lui permettre de développer son « intelligence sociale » ? Et quelles

1. En anglais, *folk psychology*. Il n'y a pas d'acception consensuelle de cette locution, que certains emploient comme un synonyme de *theory of mind* au sens techique (TdE), mais que d'autres comprennent de manière plus large.

sont les bases de la cognition sociale chez l'adulte ? (Deux bilans récents, *pro* : Malle, 2005 ; *contra* : Radcliffe, 2007.)

Ces questions concernent les philosophes autant que les psychologues, et si l'on avait le loisir d'examiner leur contribution, on aborderait bon nombre de domaines relevant de la philosophie des sciences cognitives au sens le plus large et dans toute sa diversité. On ne peut d'ailleurs manquer d'être frappé par le caractère ou, du moins, par la formulation de la question initiale, qu'un philosophe aurait pu poser si des psychologues ne l'avaient fait ; le penseur qui en a le premier pressenti l'importance, Piaget, n'est-il pas à égalité philosophe et psychologue (Piaget, 1926) ? Mais pour en venir à la question du rôle des grands modèles, on voit qu'ils n'offrent aucune prise sur le sujet : ils n'ont de ressource ni pour formuler les questions, ni pour suggérer des réponses, ni même pour recommander une méthode de recherche. Et si l'on venait dire à tel chercheur que son explication de la TdE, ou la réponse à l'une des multiples questions qu'elle soulève, est incompatible avec tel grand modèle, il est peu probable qu'il s'en soucierait, ne serait-ce que parce que la preuve avancée de la l'incompatibilité lui semblerait davantage sujette à caution que sa propre théorie.

Un second exemple, lié au premier, illustre plus nettement encore cette pauvreté des grands modèles. Chez le singe macaque ont été découverts un peu par hasard[1], il y a une quinzaine d'années, des neurones qui déchargent indifféremment dans deux conditions : soit lorsque l'animal exécute un mouvement intentionnel (tel que tendre la main vers des cacahuètes qu'on lui offre), soit lorsque l'animal observe un congénère (ou un humain) faire le même geste (Rizzolatti, 1996). Ces « neurones miroirs », selon certains chercheurs, permettent à l'animal d'identifier l'intention d'autrui, telle qu'elle s'exprime par un geste ; Bobby « comprend » mon intention d'attraper une cacahuète parce que, lorsque je tends la main à cette fin, un neurone miroir de Bobby décharge qui déchargerait également si lui, Bobby, avait tendu la main avec la même intention. Bobby peut donc rapporter son observation à sa propre intention, et identifier ainsi la mienne. Ces observations et cette interprétation ont donné naissance à une théorie « motrice » de la cognition humaine (Rizzolatti, 2003 ; Gallese, 2004), tout particulièrement de la cognition sociale humaine, qui fait l'objet de vifs débats impliquant ici encore des philosophes et des psychologues, mais aussi des neurobiologistes (Jacob & Jeannerod, 2005 ; Jacob, 2008). Ces débats ne croisent à aucun moment la question des grands modèles ; plus encore, la découverte qui les a déclenchés se situe hors du contexte le plus englobant dans lequel les grands modèles peuvent être comparés : une interprétation béhavioriste, donc non mentaliste ou informationnelle, semble possible. Seul peut-être le dynamicisme (qu'on peut d'ailleurs voir comme une forme de béhaviorisme) a l'élasticité nécessaire pour pouvoir prétendre intégrer la théorie motrice au sens fort où l'on peut arguer que

1. Les hasards de ce genre ne sont jamais purs. Voir notamment la véritable histoire de la découverte « fortuite » de la pénicilline en 1928 par Fleming.

chacun apporte un soutien à l'autre ; le classicisme comme le connexionnisme peuvent également l'intégrer, mais en un sens faible : leur sort et celui de la théorie motrice ne sont pas liés. Mais de ce genre de question, ceux qui s'intéressent aux neurones miroirs n'ont généralement cure.

2.3.4 *Existence et unité des sciences cognitives*

Il existe une tension entre les deux dernières sous-sections. L'une souligne le caractère central et les vertus heuristiques des grands modèles, l'autre leur absence de pertinence pour des secteurs entiers de la recherche contemporaine. Que faut-il comprendre ?

Les grands modèles – tout particulièrement le modèle classique, mais aussi les autres modèles dont certains précurseurs ont joué un rôle important – ont d'abord eu une fonction historique. Cette fonction, on l'a vu, a été de fournir aux sciences cognitives naissantes une perspective dans laquelle elles ont pu prendre forme, forger leurs premiers concepts, obtenir leurs premiers résultats, rapatrier les acquis assimilables de programmes de recherche qui les précédaient, en psychologie et dans d'autres domaines, regrouper un nombre suffisant de chercheurs, et atteindre une masse critique. Cette fonction à la fois sociologique et méthodologique n'a pu être assurée qu'en vertu d'une conceptualisation relativement précise, quoique d'applicabilité limitée, prenant la forme de thèses sur la nature de l'objet d'étude et sur la méthodologie complexe qui lui est applicable. Compris de manière stricte, les grands modèles défendent, en des termes et pour des raisons qui diffèrent de l'un à l'autre, une unité ontologique et une unité méthodologique des sciences cognitives. Dans le cadre qu'ils proposent, les sciences cognitives ont un objet, constituant un domaine aux contours naturels et stables ; et ce domaine doit être étudié à plusieurs niveaux, sachant qu'il existe entre ces niveaux une articulation qui permet de les subsumer comme aspects d'un même phénomène.

Ainsi, les grands modèles procurent des conditions de viabilité pragmatique aux sciences cognitives, fondées sur une perspective théorique. Ce qui est en cause aujourd'hui, c'est cette perspective théorique, mais les conditions de viabilité ne sont pas nécessairement affectées.

Essayons d'expliquer cet apparent paradoxe. Les sciences cognitives n'ont pas besoin, pour se développer, d'une garantie de l'unité ontologique de leur domaine. Elle n'ont besoin, au fond, que de la présomption que cette unité est pensable, qu'aucun argument décisif ne conclut à son incohérence. Comme dans le cas de la physique ou de la biologie, l'unité peut ne se dégager qu'à un stade ultérieur de développement. Les sciences cognitives n'ont pas non plus besoin d'interpréter littéralement les prescriptions méthodologiques de tel ou tel grand modèle. Un *modus vivendi* méthodologique leur suffit, fondé sur l'absence de frontières fixes, sur des références communes, sur une pratique de dialogue, sur un objectif de convergence conçu comme idéal régulateur. Ces conditions intellectuelles étant réunies, une communauté se constitue et prouve le mouvement en marchant. La réflexion sur les

grands modèles se replie alors sur le terrain des fondements, comme c'est le cas dans les disciplines mûres. On n'en est peut-être pas encore là, mais on peut interpréter l'évolution en cours comme une transition vers ce stade.

Mais si les grands modèles sont ainsi remis à une plus juste place, ce n'est pas seulement parce que les sciences cognitives ont commencé à mûrir et poursuivent leur trajectoire en se passant largement de leur aide. C'est aussi parce qu'ils ont leurs propres soucis.

Ces soucis sont de deux ordres. D'une part, les grands modèles sont en quête de réponses à tout un ensemble de questions d'ordre ontologique, en l'absence desquelles ils continuent de flotter dans le vague. D'autre part, ils sont en butte à des critiques franchement destructrices, visant, à travers eux, le projet même des sciences cognitives tel qu'il se déploie aujourd'hui. Cette dichotomie est simpliste, car il existe une continuité entre les deux sortes de préoccupations qu'on vient de distinguer, et qui s'étagent selon un gradient de radicalité. Mais elle reflète une certaine réalité institutionnelle : il y a deux groupes assez différents d'auteurs, qui se parlent beaucoup entre eux et peu d'un groupe à l'autre, s'inscrivant dans des perspectives distinctes.

Le premier groupe d'auteurs est d'orientation naturaliste, et recherche activement des solutions naturalistes aux problèmes de fondement des sciences cognitives. Ils peuvent être pessimistes (au sens où Borges fait dire à l'un de ses personnages qu'un gentleman ne s'intéresse qu'aux causes perdues), mais ils travaillent aux côtés des optimistes, acceptant les termes dans lesquels les questions sont posées. Ce n'est pas le cas du second groupe d'auteurs, qui sans récuser nécessairement toute forme de naturalisme, rejettent la conception qu'en proposent les premiers.

Les deux groupes (d'inégale importance numérique) travaillent en pratique sur des thèmes distincts. Le premier groupe met au cœur de son enquête trois grandes questions : celle de l'intentionnalité, celle de la causalité mentale, celle de la conscience.

La première a longtemps été considérée comme la plus centrale, ou du moins celle qui devait être attaquée en premier. Comment comprendre qu'un processus naturel se traduise, dans le vocabulaire psychologique, par le fait qu'une entité matérielle soit porteuse d'un sens, qu'elle signifie quelque chose (objet, classe, relation, état de fait) qui se situe en dehors d'elle ? Dans le cadre de l'HLP, par exemple, la question, comme on l'a vu, est de savoir en quel sens et comment les symboles du mentalais possèdent ou acquièrent leur référence ou dénotation, c'est-à-dire l'entité qu'ils désignent. La question se divise en deux : la première est celle de la référence en général, la seconde celle de l'assignation d'une référence particulière à un symbole donné. Une chose est donc de comprendre ce que signifie qu'un symbole ait une référence, une autre ce qui fait que ce symbole-ci désigne les camions plutôt que Jules César ou le triangle équilatéral que je suis en train de tracer au tableau noir. L'intentionnalité ainsi circonscrite ouvre une perspective vertigineuse : elle semble introduire le monde dans l'esprit, mettant en péril l'image de la forteresse du for

intérieur, de la tour de contrôle. L'« externalisme » est l'étiquette générale posée sur cette perspective. Il en existe des formes plus ou moins radicales, chacune offrant une conception différente de la manière dont le monde fait irruption dans l'esprit (Clark & Chalmers, 1998 ; Hutchins, 1995 ; Rowlands, 2003 ; Wilson, 2004 ; Kelly, 2000).

La seconde grande question pour les philosophes naturalistes est une version moderne du problème que Descartes pensait résoudre par l'artifice de la glande pinéale. Elle porte le nom de problème de la causalité mentale, et se formule de la manière suivante[1]. Le monde matériel évolue selon les lois de la physique. Ces lois sont complètes par principe : la physique, même si elle est encore inachevée, rassemble la totalité des lois de la nature. Elle détient donc par principe, sinon de fait, tous les moyens nécessaires pour rendre compte de tout processus ou enchaînement causal. Il n'y a pas place, dans ce tableau, pour une cause dont la physique ne pourrait rendre compte. Mais d'un autre côté, nous sommes tentés de penser que nos pensées ont un effet causal : n'est-ce pas mon intention d'ouvrir la porte qui cause l'ouverture de la porte ? Faut-il alors rejeter cette intuition, au risque de voir disparaître la psychologie de sens commun et une bonne partie de la psychologie scientifique d'aujourd'hui ?

Un troisième grand questionnement porte sur la conscience. Possède-t-elle une réalité propre, ou bien est-elle un épiphénomène ? A-t-elle plusieurs formes ou modalités, ou bien est-elle d'un seul tenant ? Joue-t-elle un rôle propre dans la cognition, et lequel ? Si elle est réelle, comment trouve-t-elle, et comment a-t-elle trouvé initialement sa place dans la nature ? À ce faisceau d'interrogations se rattachent plusieurs autres problématiques : la question des propriétés phénoménales, c'est-à-dire celles qui n'interviennent pas dans le traitement de l'information, mais accompagnent certains processus cognitifs (le goût de la poire : ce que « ça me fait » de la sentir dans ma bouche) ; la question de la nature et du rôle des émotions ; la question du moi.

Il est plus difficile de dresser une liste des thèmes autour desquels s'organise la réflexion des philosophes qui critiquent l'orientation naturaliste du premier groupe. Je me risquerai pourtant à en mentionner trois. Les deux premiers sont étroitement liés : peut-on penser l'esprit, même dans une étape préliminaire, indépendamment de la *société* ? L'esprit n'est-il pas à ce point façonné par la *culture* que sa structure naturelle, biologique, disparaît pratiquement de la description et de l'explication ? Si, comme le pensent les philosophes (et certains scientifiques) qui posent ces questions, la réponse est négative, alors il devient concevable que l'esprit, tel qu'il est conçu par les sciences cognitives actuelles (sciences de la *cognition*), ne constitue pas un authentique objet de science (Erneling & Johnson, 2005). (Rappelons qu'il ne suffit pas d'exister dans le monde matériel pour constituer un objet de science : il n'existe pas de science des objets pesant moins de 350 grammes, ni de science des textes dans

1. On en trouve un exposé beaucoup plus complet dans le chapitre « Réduction et émergence ».

lesquels la lettre *x* n'apparaît pas ; il n'existe pas une science de la prestidigitation, ni une science des malheurs, ni une science des visages.) Le troisième thème est celui du *corps* (Bermudez *et al*., 1995 ; Kelly, 2000) : est-il légitime de considérer que l'esprit est *logé* dans le corps, et qu'il est *relié* au corps, alors qu'il *est* corps, qu'il est une partie constitutive du corps ?

J'ai pu donner l'impression que ces débats, qu'ils se développent dans l'un ou l'autre camp, ou dans un entre-deux, sont sans effet sur les sciences cognitives. C'est évidemment faux. Les critiques radicales du second camp suscitent des programmes de recherche « hétérodoxes » dans les sciences cognitives, programmes qui nourrissent en retour les remises en question philosophiques. Les travaux des philosophes naturalistes, quant à eux, entrent en résonance avec des problématiques scientifiques (conformément à l'une des principales thèses du naturalisme, affirmant la continuité de la science et de la philosophie). Il s'agit autant de problèmes du premier ordre – comme lorsqu'une solution connexionniste est proposée au problème de l'origine du langage, ou que les neurosciences proposent un modèle de la conscience – que de questions du second ordre, non moins importantes, telles que celle de savoir dans quelle mesure la psychologie, la linguistique ou l'anthropologie peuvent poursuivre des enquêtes indépendamment des données et des recherches en cours dans les neurosciences (Ravenscroft, 1998 ; Gold & Stolja, 1999 ; Bennett & Hacker, 2003 ; Andler, 2005).

Les questions ontologiques des philosophes, on le voit, ont donc une pertinence pour la question de l'existence et de l'unité des sciences cognitives, envisagées dans leur état présent ou dans leur devenir. Le lecteur actif aura suivi cette piste tout au long du chapitre. Mais il lui faudra chercher ailleurs une présentation moins allusive des questions ontologiques, et des conséquences à en tirer pour les sciences cognitives elles-mêmes. Car il est grand temps que s'arrête ce chapitre.

Il se termine donc là où d'autres auteurs l'auraient fait commencer. J'ai posé une série de questions de nature ontologique qui non seulement relèvent, selon eux, de la philosophie des sciences cognitives, mais en constituent le cœur, et je les ai laissées sans réponse après les avoir tout juste formulées. Je voudrais donc, en conclusion, dire quelques mots de la division technique du travail chez les philosophes s'intéressant à la cognition.

Plusieurs termes existent pour désigner leurs aires d'activité : philosophie des sciences cognitives, philosophie de la psychologie, psychologie philosophique, philosophie cognitive, philosophie de l'esprit, philosophie de la cognition. Glissons rapidement sur deux évidences : *primo*, la terminologie varie d'un philosophe ou d'un ouvrage à l'autre, et on ne peut donc en tirer, du moins directement, d'information fiable ; *secundo*, aucune classification ne doit viser à éliminer les chevauchements[1],

1. Ils sont, de fait, si importants que certains philosophes se refusent à établir les distinctions que je propose, n'y voyant que des effets terminologiques ou des nuances sans portée théorique.

qui sont non seulement inévitables, mais qui jouent un rôle essentiel à la fois pour faire circuler les concepts et les idées, et pour prévenir des cristallisations doctrinales et la formation de chapelles.

Portons plutôt notre attention sur les objectifs que peuvent se proposer les philosophes, et sur leur position par rapport aux sciences. Le philosophe A s'interroge sur les sciences cognitives sur un mode à la fois descriptif et normatif ou critique : il est proche de cette discipline, mais il ne se donne pas pour objectif unique de l'assister dans sa tâche, et ne prétend pas y contribuer directement. Son attitude est semblable à celle qu'adoptent la plupart des philosophes de la physique, des mathématiques ou de la biologie. Le philosophe B, au contraire, veut contribuer aux sciences cognitives par tous les moyens dont il dispose : analyse conceptuelle, mais aussi participation à des recherches interdisciplinaires, impliquant de sa part l'acquisition de compétences scientifiques, fussent-elles ponctuelles. Le philosophe C, quant à lui, s'interroge directement sur l'objet des sciences cognitives, mais d'une manière qui ne dépend pas entièrement d'elles et de leurs choix méthodologiques et qui s'inscrit dans une tradition philosophique. Le philosophe D s'intéresse, pour sa part, à la psychologie dans toute son étendue et sa diversité. Les objectifs de D sont à la fois plus étroits et plus larges que ceux de A : il tend à laisser de côté certaines questions du domaine de A (par exemple, des questions relatives au langage, à l'évolution des cultures, à l'intelligence artificielle, à la méthodologie des neurosciences), mais peut inversement se concentrer sur des écoles ou des branches de la psychologie qui ne sont pas (du moins pour l'instant) de la compétence des sciences cognitives (la psychologie du travail, la psychanalyse, la psychologie de l'éducation, la psychologie du caractère et de l'intelligence…). D'autre part, il prête attention à la méthodologie propre à la psychologie scientifique, de la chronométrie ou de l'amorçage à la mesure du temps de regard chez les tout-petits ou à la succion non nutritive chez les nourrissons, de l'héritabilité des traits de caractère ou de l'intelligence[1]. De même, le domaine de C est à la fois plus restreint et plus limité que celui de B : C peut, par exemple, contrairement à B, défendre le dualisme, ou se placer dans une perspective phénoménologique, ou encore wittgensteinienne, sans chercher, comme le fait B, à rejoindre d'une manière ou d'une autre les sciences cognitives[2]. Ces idéaux-types (au sens de Weber) sont représentatifs de ce que j'appellerai, respectivement, philosophie des sciences cognitives (pour A), philosophie cognitive ou psychologie

1. Dans la mesure où la linguistique, les neurosciences, l'anthropologie sont également partiellement immergées dans les sciences cognitives, elles donnent lieu à une distribution des tâches un peu comparable : la philosophie des sciences cognitives met l'accent sur les rapports entre les disciplines composantes, sur leurs convergences, etc., tandis que la philosophie de la linguistique, des neurosciences, etc., d'une part embrasse par définition tous les courants, y compris non « cognitifs », de la linguistique, etc., d'autre part se concentre sur les problèmes spécifiques à la discipline.

2. Il est cependant apparu récemment un courant d'inspiration phénoménologique qui veut contribuer très directement aux sciences cognitives (voir Dreyfus, 1982 ; McClamrock, 1995 ; Petitot *et al.*, 2002 ; Smith & Thomasson, 2005 ; Andler, 2006b).

philosophique (d'orientation cognitive) (pour B), philosophie de l'esprit (pour C) et philosophie de la psychologie (pour D). La philosophie cognitive et la psychologie philosophique sont proches des sciences cognitives au sens où elles en partagent les objectifs directs ; la philosophie des sciences cognitives et la philosophie de la psychologie en sont plus éloignées : leurs objectifs ne coïncident pas nécessairement, entièrement et à tout moment, avec ceux des sciences cognitives. La philosophie de la psychologie et la psychologie philosophique sont évidemment proches de la psychologie comme discipline distincte et autonome, la philosophie des sciences cognitives et la philosophie cognitive en sont plus éloignées puisqu'elles s'intéressent précisément à une approche qui se propose de plonger (voire parfois de dissoudre) la psychologie dans un cadre théorique beaucoup plus large. Enfin, la philosophie de l'esprit recoupe largement les autres branches, tout en disposant d'une autonomie propre par rapport à la perspective scientifique.

La division des tâches n'est pas la seule explication de cette géographie des spécialités. Il y a aussi des désaccords de doctrine, qu'ils soient du premier ordre (par exemple, sur la question du naturalisme) ou du second (portant sur une conception normative du rôle du philosophe). C'est là encore un sujet qui ne sera pas abordé ici.

Ce chapitre a choisi le point de vue du philosophe A. Il n'a pas cherché à éviter la compagnie de B, C ou D. Mais il n'a pas suivi l'un ou l'autre des chemins qu'ils auraient pris à sa place. Il a aussi dû laisser de côté bon nombre de questions qui relèvent incontestablement de sa compétence. L'objectif, pour le dire une dernière fois, était de tenir au sujet des sciences cognitives le genre de propos que le philosophe de la biologie tient au sujet de la biologie, le philosophe de l'économie à propos de la science économique, et ainsi de suite. Si, comme je le crois, cet objectif n'a pas été pleinement atteint, la faute en revient pour partie, comme j'en ai prévenu le lecteur, au domaine, et pour partie, naturellement, à l'auteur.

Daniel Andler
Université Paris-Sorbonne (Paris IV)
et institut universitaire de France