This is an unedited version of a paper which appeared in
PSN. Psychiatrie, sciences humaines, neurosciences **3** 12 (2005), pp. 74-87

Les neurosciences cognitives: une nouvelle « nouvelle science de l'esprit »?
Cognitive neuroscience : A new 'new science of the mind' ?

Daniel Andler
UFR de philosophie et sociologie, Université de Paris-Sorbonne (Paris IV)
Département d'études cognitives, Ecole normale supérieure

Résumé (français)

Une question controversée est de savoir si, grâce en particulier aux techniques d'imagerie cérébrale fonctionnelle, une nouvelle science du cerveau occupe désormais, sous le nom de « neurosciences cognitives », le cœur des sciences cognitives. Quels sont les fondements de cette nouvelle spécialité? Comment s'articuleront, demain, les neurosciences cognitives et la psychologie ? On replacera les techniques d'imagerie dans le contexte théorique des neurosciences fondamentales, et on montrera que les neurosciences cognitives en sont tributaires, comme elles le sont de l'ensemble des sciences cognitives, au sein desquelles elles constituent aujourd'hui un programme de recherche parmi d'autres. S'il venait à constituer, dans un avenir lointain, une science achevée des fonctionnalités du cerveau, il ne remplacerait sans doute pas la psychologie cognitive, appuyée sur les autres sciences de l'esprit, mais entrerait avec elle dans un rapport de complémentarité en un sens fort, très éloigné de toute forme de réduction.

Abstract (English)

It is a matter of considerable controversy whether cognitive neuroscience, thanks in large part to functional neuroimaging techniques, is in the process of becoming a new science of the brain and moving into the heart of cognitive science. What are the foundations of this new field ? How will neuroscience and cognitive science coexist in the future ? The paper will attempt to situate neuroimagery in the theoretical framework of fundamental neuroscience, and will show the extent to which cognitive neuroscience depends on it, as it depends on the rest of cognitive science, within which it stands as one research program among several. Should it lead, in a distant future, to a completed science of the brain's functionalities, such a science would likely not replace cognitive psychology and allied disciplines. Instead, I envisage a form of strong complementarity between the two branches, exclusive of any form of reduction.

Mots-clés : neurosciences cognitives, neuroimagerie, sciences cognitives, philosophie des sciences

8 février 2005, rév. 14 février 2005

Les neurosciences cognitives: une nouvelle « nouvelle science de l'esprit »?

Daniel Andler
Université de Paris-Sorbonne (Paris IV) & Ecole normale supérieure

Le rôle des neurosciences dans l'étude scientifique de l'esprit s'est prodigieusement accru au cours des dernières années, tout particulièrement depuis que les méthodes d'imagerie cérébrale se sont généralisées. L'étiquette "neurosciences cognitives", apparue récemment, semble témoigner d'une reconfiguration des sciences cognitives sous l'égide des neurosciences. S'agit-il d'une seconde "révolution cognitive"? Les acquis de la première sont-ils mis en question? Peut-on s'en remettre désormais aux neurosciences pour ériger enfin une science progressive de l'esprit, sous la forme d'une science du cerveau? [1]

1. Position du problème

Ces questions sont déterminantes pour l'avenir des sciences de la cognition, et se posent dans un contexte de fortes tensions qui ne sont pas seulement intellectuelles, mais également institutionnelles. Une discussion sereine n'est pas toujours possible. Une autre source de confusion vient de ce que les choses se présentent différemment selon qu'on se place à l'extérieur ou à l'intérieur du milieu professionnel des sciences cognitives, et il est difficile d'en parler d'une manière qui soit recevable par les deux catégories de lecteurs.

Le lecteur extérieur cherche à identifier une discipline, un programme de recherche, une thèse, une méthode qui soient caractéristiques des sciences cognitives; il n'a ni les ressources ni le désir de mener sa propre enquête, et veut seulement savoir à quel représentant autorisé il doit s'adresser. C'est auprès de ce représentant qu'il va s'enquérir sur les fins et sur les moyens de ce qu'il considère comme un programme de recherche parmi d'autres, et qu'il va juger dans quelle mesure les sciences cognitives, ainsi représentées, apportent ou ont des chances raisonnables d'apporter des réponses à telle ou telle question qu'il juge essentielle concernant l'esprit.

Pour le lecteur de l'intérieur, la seconde interrogation n'est pas d'actualité: revendiquer, ou accepter l'étiquette "sciences cognitives", c'est faire le pari que des questions

[1] Je remercie pour leurs remarques critiques sur une première version de cet article Anne Fagot-Largeault, Bernard Pachoud, Arnaud Plagnol, Jérôme Sackur et Marta Spranzi ; et pour leurs réactions les participants aux séminaires de Jacques Glowinski au Collège de France, et de Dag Westerstahl au département de philosophie de l'Université de Göteborg.

centrales et urgentes sur l'esprit sont à la portée de ces sciences, et qu'il serait vain, au stade présent, de mettre en question leur pertinence[2]. En revanche, il est partie prenante d'une perpétuelle querelle interne sur la bonne stratégie et sur le rôle des disciplines participantes.

Aujourd'hui, un témoin arrivé depuis peu sur la scène, et renseigné par les journaux, les revues savantes, les manifestations scientifiques, n'aurait guère de doute sur la question du représentant: les sciences cognitives ne sont désormais rien d'autre que les sciences du cerveau (les neurosciences), ou peut-être une branche de cette spécialité, celle qui porte le nom de "neurosciences cognitives", et la méthode princeps de cette discipline est l'imagerie cérébrale fonctionnelle.

Pour un acteur des sciences cognitives, cette évidence n'en est pas une. Non seulement il ne va pas de soi que les neurosciences cognitives sont le cœur du domaine, mais le sens et la portée de cette présentation font question. A supposer qu'on voudra lui donner son assentiment, ce ne pourra aucunement l'être en vertu d'une définition ou d'un simple constat, mais seulement à titre d'un bilan, terme d'une analyse complexe de l'évolution du domaine. Et l'affichage du bilan sous cette forme ramassée ne manquera pas d'avoir des conséquences intellectuelles et institutionnelles, tant au sein du domaine que vis-à-vis de l'extérieur.

Après avoir averti le lecteur de leur importance, on n'insistera pas sur les aspects sociologiques du problème. Et on ne traitera (évidemment) pas la totalité des questions épistémologiques que soulèvent les sciences cognitives. Pour pouvoir formuler celles que l'on abordera, et les réponses qui seront proposées, il faut s'entendre sur les termes.

Mise au point terminologique

Les *sciences cognitives* désignent l'ensemble des programmes de recherche consacrés à l'élaboration d'une théorie scientifique de l'esprit humain, et qui souscrivent à une sorte de "cahier des charges" qu'il serait fastidieux de rappeler ici[3], mais dont il est important de retenir qu'il est essentiellement méthodologique, et que non seulement il autorise, mais qu'il favorise la formulation et le déploiement de doctrines et de théories très différentes. C'est ce pluralisme qui donne leur sens aux désaccords dont il va être question,et qui sont tout à fait comparables aux controverses courantes dans toutes les disciplines scientifiques vivantes.

Les sciences cognitives ont une double structure: d'une part elles développent des programmes de recherche pluri- ou interdisciplinaires sur les différentes fonctions mentales;

[2] On peut être plus radical, et considérer que par définition toute question sur l'esprit susceptible d'une formulation scientifique relève des sciences cognitives. Cela laisse ouverte la possibilité de questions importantes, voire essentielles, concernant l'esprit, auxquelles aucune réponse *scientifique* ne semble pouvoir être apportée dans un avenir proche: Chomsky, l'un des pères de la "révolution cognitive", défend cette opinion, d'autres la jugent incohérente, obscurantiste ou irresponsable.

[3] V. par exemple D. Andler, dir., *Introduction aux sciences cognitives*, Paris: Gallimard, coll. Folio, nouvelle éd., 2004; D. Andler *et al.*, *Philosophie des sciences*, Paris: Gallimard, coll. Folio, 2003, vol. 1, chap. 3; O. Houdé *et al.*, *Vocabulaire des sciences cognitives*, Paris : PUF, 1996.

d'autre part elles s'appuient sur certaines branches de disciplines fondamentales. En particulier, les neurosciences y jouent un rôle important. Par ce terme nous entendrons, conformément à l'usage, l'ensemble des branches de la biologie portant sur le système nerveux, quel que soit le niveau d'analyse et la classe de phénomènes qu'elles priviliégient: la neurophysiologie, la neuroanatomie, la neurobiologie cellulaire et moléculaire en sont des parties essentielles, à côté de la neuropsychologie, de la neuropharmacologie et des neurosciences appelées parfois "intégratives"[4]. Il peut alors sembler assez naturel de désigner par "neurosciences cognitives" la partie des neurosciences qui est impliquée dans les programmes de recherche des sciences cognitives. Ce choix est néanmoins très problématique et risquerait d'embrouiller des questions que cet article voudrait contribuer à clarifier. En effet, cette expression est associée aujourd'hui à une certaine pratique scientifique, relevant autant de la psychologie que des neurosciences, et appuyée sur le recours massif à la neuroimagerie fonctionnelle. Il serait maladroit d'introduire dans la terminologie même l'hypothèse que c'est cette pratique particulière qui constitue, pour le dire rapidement, l'intersection des neurosciences et des sciences cognitives, alors que cette hypothèse mérite justement d'être examinée[5]. De plus, pour ce qui est de la place des neurosciences computationnelles, un autre label récent, cette décision créerait une difficulté inutile. Convenons donc de ne pas préjuger de la question de la division des tâches, au sein des neurosciences, entre celles de ses branches qui sont pertinentes pour les sciences cognitives et celles qui ne le sont pas. Et convenons également de considérer les méthodes d'imagerie, d'une part, les techniques de modélisation issues des mathématiques, de la physique, de l'informatique, d'autre part, comme autant d'outils à la libre disposition de toutes les branches des neurosciences et des sciences cognitives.

Questions à prendre, questions à laisser

Ce petit « nettoyage de la situation verbale » (Valéry) étant fait, il devient relativement facile de formuler les questions qui se posent dans le présent contexte, et de dire celles qui seront abordées et celles qui seront laissées de côté:

(Q1) Les sciences cognitives posent-elles des questions importantes relativement à l'esprit (les processus, états, phénomènes mentaux)?

[4] Il n'existe pas aujourd'hui de découpage des neurosciences qui fasse l'unanimité ; l'énumération proposée ne l'est qu'à titre indicatif.

[5] On rencontre un problème analogue en linguistique: le label "linguistique cognitive" désigne un programme de recherche particulier, ou plutôt une famille de programmes liés par des affinités théoriques et institutionnelles. Cette orientation, au sein de la linguistique, revendique une pertinence élevée, voire exclusive, pour les sciences cognitives. Une position parfaitement respectable, mais pour la défendre efficacement, comme pour la contester, il est crucial de ne pas l'introduire par stipulation dans le terminologie même, en décrétant que la linguistique cognitive est l'"intersection" de la linguistique et des sciences cognitives. On peut déplorer ce qui apparaît peut-être, vu de l'extérieur, comme des distinguos byzantins, mais ils sont absolument indispensables.

(Q2) Les sciences cognitives ont-elles les moyens de leurs ambitions, possèdent-elles des méthodes éprouvées, sont-elles progressives?

(Q3) En particulier, la psychologie expérimentale (prise au sens le plus large) s'appuie-t-elle sur une méthodologie éprouvée, antérieure à l'émergence des neurosciences cognitives et de la neuroimagerie?

(Q4) Les sciences cognitives ont-elles vocation, à terme, d'absorber la totalité des enquêtes systématiques impliquant l'esprit, la conscience, la pensée?

Ces quatre questions ne seront pas discutées. Nous postulerons, pour les besoins (urgents) de la discussion, une réponse positive aux trois premières, et écarterons résolument la quatrième. Les questions qui seront examinées, directement ou indirectement, sont les suivantes:

(Q5) Quelle est l'importance de la neuroimagerie fonctionnelle pour les sciences cognitives aujourd'hui?

(Q6) Quelle que puisse être son importance relative, en vertu de quoi modifie-t-elle la situation?

(Q7) Les neurosciences occupent-elles désormais le centre des sciences cognitives, et si c'est le cas, est-ce essentiellement du fait de l'imagerie?

(Q8) Dans l'hypothèse où la neuroimagerie, associée éventuellement à d'autres progrès techniques et conceptuels des neurosciences, conduirait à une science développée des fonctionnalités du cerveau, quelle forme prendraient les sciences cognitives? Se réduiraient-elles alors aux neurosciences?

En posant quatre questions distinctes, quoique manifestement liées, là où certains ne voient qu'une unique interrogation: Oui ou non, les neurosciences cognitives sont-elles le présent, ou l'avenir imminent, des sciences cognitives?, on se ménage la possibilité d'une réponse nuancée, qui ne soit pas pour autant une réponse de Normand. Ce qu'on entend surtout aujourd'hui, ce sont des réponses en tout ou rien: pour les uns, la neuroimagerie est une découverte fondamentale qui bouleverse dès à présent les sciences cognitives et les rabat en réalité sur les sciences du cerveau; pour les autres, la neuroimagerie est une technique coûteuse, sans portée théorique, dont le principal effet est de retarder ou d'entraver le progrès dans le domaine, en monopolisant les ressources et l'attention, et en en donnant une image ultra-réductionniste peu propice au dialogue avec les autres disciplines et courants de pensée, tant au sein des sciences cognitives qu'à l'extérieur.

Résumée en quelques lignes, la position que je défendrai dans le présent article est la suivante. Non, les neurosciences ne bouleversent pas la situation dans les sciences cognitives. Non, elles n'ont pas les moyens à elles seules d'édifier une science de l'esprit. Elles changent bel et bien la donne, mais ce n'est pas principalement du fait qu'elles apporteraient, enfin, des données empiriques suffisamment nombreuses et détaillées pour nous dispenser de recourir aux spéculations incertaines dont nous devions nous contenter jusqu'à l'arrivée de l'imagerie.

Bien entendu, et il ne sera pas nécessaire de le rappeler constamment, les neurosciences apportent des informations importantes, non seulement pour les sciences cognitives théoriques, mais aussi pour la neurologie, peut-être pour la psychiatrie, et la récolte ne fait que commencer. Mais si l'on peut imaginer aujourd'hui que les neurosciences pourraient dans un avenir prévisible modifier la situation en profondeur, c'est davantage en raison des questions nouvelles qu'elles commencent à poser que des réponses qu'elles apportent à certaines questions déjà formulées par les psychologues, les philosophes ou les linguistes. On en verra des exemples, et on s'interrogera pour conclure sur ce que pourrait être la situation d'une science de l'"esprit-cerveau" lorsque l'orientation neurocognitive aura eu le temps de produire son plein effet.

2. Un retour en arrière

Les prémices de ce qui ne s'appelait pas encore les sciences cognitives apparurent il y a un demi-siècle. On put bientôt parler d'une "nouvelle science de l'esprit", fût-ce pour s'interroger sur sa réalité, sur sa nouveauté ou sur son unité. Il s'agissait à première vue d'un ensemble de programmes de recherche rangés sous différentes étiquettes disciplinaires. Leur convergence devait résulter de l'adhésion à certaines thèses, de l'adoption d'un certain vocabulaire théorique, d'une méthodologie particulière, d'une pratique interdisciplinaire. Une histoire bien connue, du moins dans ses grandes lignes, sur laquelle on ne revient que pour attirer l'attention sur la place occupée initialement par deux des composantes de ce mouvement. La première était la nouvellement nommée "intelligence artificielle" (IA), la seconde ce qui ne s'appelait pas encore les neurosciences. L'IA occupait le centre de la scène, rassemblant en elle l'essentiel du projet, sa philosophie et sa méthode; c'est du moins ainsi qu'elle-même concevait son rôle (elle était, selon l'un de ses pères fondateurs, l'"interdiscipline"), et, que ce soit par conviction ou par calcul, on ne songeait généralement pas, dans les milieux de la "nouvelle science", à lui contester cette prééminence. Les arts et sciences du système nerveux occupaient à l'inverse une position marginale. Cela ne tenait pas à leur moindre mérite, mesuré à l'aune de l'ancienneté ou des résultats, théoriques et pratiques (à ce compte, l'IA ne pesait pas bien lourd). Ce qui limitait le rôle de ces précurseurs des neurosciences, c'était leur faible pertinence dans le projet tel que l'IA et les disciplines alliées l'avaient formulé: il s'agissait de caractériser, d'expliquer, de simuler ou de modéliser les processus de la pensée "intelligente". Or cette pensée était conçue comme étant essentiellement constituée de processus conceptuels, délibérés, verbalisables; elle était une capacité des représentants adultes typiques de l'espèce humaine; enfin, et surtout, la caractérisation recherchée se situait à un niveau d'abstraction nettement distingué, et en un

sens indépendant, de celui des processus concrets, de l'enchaînement des causes. Ce que l'on voulait comprendre, c'était la *structure logique* de certains processus de "traitement de l'information". La façon dont ces processus étaient effectués, physiquement, dans le tissu nerveux n'avait qu'une importance relative.

Cette conception de l'esprit (ou de ses productions) alliant un monisme ontologique (tout processus mental est un processus matériel) à un dualisme méthodologique (les catégories explicatives au niveau mental ne sont pas identiques aux catégories explicatives au niveau matériel) porte le nom de "fonctionnalisme" et demeure la doctrine catholique des sciences cognitives, même si elle traverse une crise profonde. Ce qui importe ici, c'est que le fonctionnalisme assignait des limites de principe à toute contribution potentielle des neurosciences, sans toutefois la ramener à zéro, puisqu'ils fournissaient au moins aux architectes de l'esprit des contraintes d'ingéniérie. Il fallait bien que les processus informationnels postulés par le théoricien de l'IA ou de la psychologie cognitive fussent matériellement exécutables par le cerveau tel qu'on le connaissait (composé, par exemple, de neurones lents et peu fiables). En pratique, ce genre de contrainte n'était guère exploitable par les théoriciens de l'IA. A cette restriction de principe imposée par le cadre théorique au ressources explicatives des neurosciences s'ajoutait le fait qu'à l'époque elles avaient surtout prise d'une part sur la perception (essentiellement sur la vision), sur la motricité, sur les émotions ou dispositions caractérielles, d'autre part sur les pathologies massives ou les différences interindividuelles, tous domaines apparemment bien éloignés de la pensée conceptuelle caractéristique de l'adulte humain moyen.

Un spectaculaire renversement de la situation

Aujourd'hui, l'IA est réléguée aux marges (alors même qu'elle s'est enrichie de résultats et de savoir-faire, et guérie des excès de sa jeunesse), et ce sont les neurosciences qui revendiquent un rôle central dans les sciences cognitives, rôle qui lui est souvent accordé – conviction ou calcul?– par les autres branches du domaine. On pourrait penser que ce renversement résulte d'une concurrence théorique directe, l'IA ayant joué "STI" (systèmes de traitement de l'information) et finalement perdu contre les neurosciences qui misaient sur "SNC" (système nerveux central). L'histoire aurait ainsi montré que la voie de l'abstraction informationnelle se perdait dans les sables, tandis que l'approche concrète de l'organe siège de la cognition, une fois munie des instruments nécessaires, mènerait au but. Ou, autre lecture, l'IA aurait misé sur le modèle de l'ordinateur, et perdu tout simplement parce que ce n'est pas un bon modèle pour le cerveau et que seule une modélisation de l'organe, à partir des données empiriques disponibles, est susceptible de conduire à de bons modèles de la fonction. Or ces deux lectures, malgré les bribes de vérité qu'elles contiennent l'une et l'autre,

sont factuellement fausses. Et ce n'est pas là seulement un point d'histoire, car l'échec, avéré, de l'IA est porteur de leçons applicables à la situation présente.

Ce qui a perdu l'IA (celle de la première période, parfois dite "prométhéenne") n'est ni le choix du niveau informationnel ni celui de l'ordinateur comme système de traitement, mais une erreur stratégique consistant à ériger la simulation en méthode universelle de découverte et de justification. L'idée était de prendre un à un les processus de l'intelligence (grosso modo, les tâches intellectuelles ou mentales, au sens le plus large, que l'être humain peut accomplir), de les représenter comme résolutions de problèmes, et d'en proposer des simulations informatiques: pour un processus donné, un programme était proposé, ses performances étaient comparées au processus étudié, les écarts constatés suggérant des modifications, et le programme obtenu au terme d'une suite d'approximations successives constituait à la fois la *théorie* du processus et une *preuve* de la théorie (la théorie étant prouvée si, moyennant idéalisation, les résultats obtenus par le programme sont ceux qu'obtient l'agent humain).

Or cette méthode présupposait une identification suffisamment claire et précise du processus à simuler (ce qu'on appelait en IA le *task domain)* : jouer aux échecs, se remémorer une liste d'objets présentés oralement, empiler des blocs de bois en structures stables, traduire un texte d'italien en russe, démontrer une proposition logique ou géométrique, identifier une scène visuelle, rédiger un chèque bancaire, répondre à une demande de renseignement fiscal, calculer les doses d'insuline nécessaires au traitement d'un diabète... Comment catégoriser et caractériser cette immense variété de tâches? Sans doute n'était-il pas nécessaire que l'identification initiale soit parfaite, car la recherche du programme adéquat était censée fournir les corrections utiles, mais il fallait que le "prédécoupage" respecte certaines frontières naturelles, faute de quoi l'itération divergerait, ou bien conduirait à une "machine" hybride, arbitraire et fragile, manière de robot ménager répondant au téléphone et faisant monter les œufs en neige dans les cas prévus par le fabricant.

Quelle devait donc être la source de ces identifications préalables? L'IA avait là-dessus deux idées complémentaires: le sens commun (des spécialistes de l'IA) et, dans certains cas, le psychologue ou le linguiste, voire le logicien, le philosophe ou l'anthropologue. L'ironie est que cette réponse était bonne, mais que l'IA n'a pas su en conjuguer les deux volets : elle a dramatiquement surestimé les ressources du sens commun[6], et corrélativement sous-estimé la difficulté de la tâche du psychologue, du linguiste etc., commettant du coup la troisième et fatale erreur de se croire capable d'orchestrer les recherches relevant des disciplines fondamentales (en quoi consistait son rôle d'"interdiscipline"). Bref, l'espoir de substituer à l'étude empirique des capacités mentales,

[6] ... tout en les sous-estimant: elle a surestimé le sens commun comme heuristique scientifique (nos intuitions sur nos processus mentaux sont très limitées), et l'a inversement sous-estimé comme composante de la cognition naturelle.

par les méthodes scientifiques traditionnelles, la mise au point de simulations, dans un cadre théorique réduit à l'hypothèse computationnelle, cet espoir était vain (a posteriori, son audace surprend). Et ce sont l'expérience de l'enlisement et la critique par les disciplines directement impliquées (dont les neurosciences n'étaient pas) qui ont mis un terme à l'IA prométhéenne.

L'IA se heurtait à d'autres obstacles, qui l'auraient peut-être arrêtée ultérieurement si, en concluant une alliance plus humble avec la psychologie cognitive et la linguistique théorique, elle avait évité de s'étouffer prématurément faute d'alimentation scientifique. Ces obstacles (intellectualisme, formalisme, fonctionnalisme, problème du cadre, problème du sens commun,...) sont aujourd'hui l'affaire des sciences cognitives dans leur ensemble, qui les négocient ou les contournent dans leur cadre théorique propre. Les neurosciences n'ont pas les ressources conceptuelles nécessaires pour éliminer ces perplexités ou pour développer un point de vue qui leur soit propre.

D'où vient le succès actuel des neurosciences?

Puisque ce n'est pas d'avoir eu raison d'un adversaire naturel (au sens où l'on peut dire par exemple, en première approximation, que le cognitivisme naissant a eu raison du béhaviorisme) que les neurosciences tirent leur dynamisme actuel, quelles en sont les causes? La première est que les progrès des sciences cognitives (y compris, naturellement, de leur composante neurobiologique) ont préparé le terrain: le "prédécoupage" du mental est bien plus avancé qu'il n'était il y a même vingt ans, a fortiori cinquante, et la pratique de l'interdisciplinarité, dans le domaine de la cognition, a atteint une certaine maturité.

La seconde cause est que les neurosciences, en partie grâce à l'imagerie, trouvent énormément de choses à dire sur le cerveau et ses fonctions. Et c'est là que la situation évoque, *mutatis mutandis,* celle de l'IA à l'heure de ses premiers triomphes. Comme l'IA naissante, les neurosciences cognitives semblent parfois revendiquer la responsabilité de refonder l'étude de l'esprit sur des bases réellement scientifiques, comme si la nouvelle science de l'esprit que l'IA entendait incarner était en instance de remplacement. Ouvertes aux autres disciplines, elles sont tentées de se considérer comme l'interdiscipline en charge de rassembler les connaissances nécessaires pour l'enquête scientifique sur l'esprit-cerveau, dont elles sont le maître d'œuvre naturel en vertu du fait qu'elles détiennent la compétence s'agissant du cerveau. Car, dans le complexe esprit-cerveau, le cerveau jouit du privilège d'être un système biologique concret, soumis à une série de contraintes repérables par les méthodes éprouvées de la biologie, et plus récemment rendu accessible à l'observation *in vivo*. Symétriquement, l'IA se réclamait naguère des méthodes éprouvées de la logique et de l'analyse conceptuelle, qu'elle entendait appliquer directement à l'objet de l'enquête, à savoir non le cerveau mais ses facultés. Aujourd'hui les neurosciences, comme hier l'IA, fixent donc

l'objet priviligié, mais aussi la forme des réponses: aux programmes de l'IA répondent les régions ou systèmes impliqués, les codages neuraux. Comme hier l'IA, les neurosciences estiment avoir par définition un rôle à jouer dans toute enquête sur les facultés mentales, et nombreux sont les psychologues, les linguistes, les philosophes, les anthropologues qui ne concevraient plus de publier un article dans leur domaine sans référence à des données neuroscientifiques, acquises ou attendues, si ténue que puisse être leur connexion réelle à la question traitée (hier c'était à l'IA que s'adressait ce genre d'hommage). Comme hier l'IA, les neurosciences sont certaines que les découvertes d'aujourd'hui sont les premiers pas vers une explication asymptotiquement achevée des fonctions mentales.

Enfin, les neurosciences, tout particulièrement lorsqu'elles sont engagées dans le programme de recherche des neurosciences cognitives, sont promptes à s'offusquer de toute réflexion critique sur leur rôle: elles y voient l'expression d'une attitude anti-scientifique ne pouvant avoir d'autre effet, quelles que soient les intentions proclamées de leurs auteurs, qu'un retour à une psychologie coupée de tout fondement naturaliste. Elles l'estiment également inopportune sur le plan politique, alors qu'elles doivent se défendre, au sein de leur discipline-mère, contre des mastodontes tels que la biologie moléculaire ou la génomique. A l'échelle nationale, elles attendent un soutien loyal, pour affronter dans les meilleures conditions possibles la concurrence mondiale. L'IA, autrefois, faisait valoir les mêmes arguments.

Ce parallèle avec l'IA n'a pas pour but de suggérer que les neurosciences sont aujourd'hui dans une situation à tous égards semblable à celle de l'IA hier, et en particulier qu'elles connaîtront demain le sort qui est celui de l'IA aujourd'hui. Il sert seulement de mise en garde contre un certain type d'exagération, et peut nous aider à préciser la nature des apports spécifiques des neurosciences aux sciences cognitives, qui leur donnent un avantage incontestable sur l'IA première manière.

2. L'imagerie fonctionnelle cérébrale

Pendant des décennies (et même, en un certain sens, depuis l'Antiquité), philosophes et psychologues se sont demandé si les images mentales existent réellement, c'est-à-dire si certains processus cognitifs (par exemple, s'imaginer aller d'un endroit à un autre, ou déterminer par la seule réflexion si telle forme géométrique est superposable à telle autre) mettent en jeu des représentations mentales irréductiblement iconiques, ou si au contraire tous les processus cognitifs consistent en manipulations de "phrases" d'une manière de langage interne. Ce débat, écrivait Stanislas Dehaene il y a déjà sept ans, "est aujourd'hui tranché. [...] Quelques expériences de tomographie par émission de positons et d'imagerie

fonctionnelle par résonance magnétique auront suffi" [7] à établir définitivement la réalité de l'imagerie mentale Jugement qui figure aussi dans le titre même du livre de Stephen Kosslyn, *Image and Brain: The resolution of the imagery debate*[8]. En réalité, le débat n'est *pas* tranché, comme le montre en particulier Zenon Pylyshyn[9], même s'il se pourrait qu'a posteriori le jugement de la communauté s'aligne sur celui de Dehaene et Kosslyn. Pour notre propos, plus intéressant que cette double controverse (sur l'existence de l'imagerie mentale[10] et sur la question de savoir si l'affaire est aujourd'hui entendue) est l'argument principal avancé par Dehaene: le cortex visuel primaire V1 est impliqué dans les tâches d'imagerie. Ainsi, les bases neurales du phénomène psychologique de l'imagerie mentale comprennent le cortex visuel primaire, lequel fait partie des bases neurales de la perception visuelle, donc l'imagerie met en jeu de véritables images, c'est-à-dire des représentations iconiques. Nous nous référons plus loin à cet exemple.

Parmi de très nombreux autres exemples disponibles, un second cas, plus subtil et plus récent, est fourni par la découverte que le contôle contextuel de l'action (en gros, la préparation prémotrice d'une action planifiée tenant compte des stimuli perceptifs) dépend des aires de Brodmann 44 et 45, dont la partie incluse dans l'hémisphére gauche n'est autre que l'aire de Broca dont on connaît le rôle dans la production du langage[11]. Contrôle de l'action et langage se trouvent ainsi rapprochés, d'une manière que ne livre nullement l'évidence d'une simple analyse conceptuelle.

La méthode caractéristique des neurosciences cognitives

De manière générale, le domaine des neurosciences cognitives produit une quantité considérable de résultats mettant en relation des événements mentaux (psychologiques) ou comportementaux A, B,... et des activations différentielles dans des zones X, Y... du cerveau. Ces résultats prennent l'une des formes caractéristiques suivantes :

(i) L'événement mental A est corrélé avec une activité accrue des zones X, Y,... du cerveau ("implication de X, Y,... dans A").

---

[7] Stanislas Dehaene, Introduction, *Le cerveau en action*, collectif, Paris: PUF, 1997, p. 2.

[8] Cambridge, MA : MIT Press, 1994.

[9] V. par ex. Le débat sur l'imagerie est-il terminé? Si oui, de quoi s'agissait-il ?, in E. Dupoux, dir., *Les langages du cerveau*, Paris: Odile Jacob, 2002, 65-88; et aussi Nigel Thomas, article "Mental imagery", *Encyclopaedia of Cognitive Science*, Lynn Nadel, ed., Londres: Nature Publ. & Macmillan, 2003.

[10] A ne pas confondre avec l'imagerie cérébrale: celle-ci désigne les techniques de visualisation des aires actives du cortex au cours de l'exécution de tâches ou de processus mentaux particuliers; celle-là est le phénomène psychologique particulier dont il est question dans la controverse.

[11] E. Koechlin, C. Ody, F. Kouneiher, The architecture of cognitive control in the human prefrontal cortex, *Science* **302** (2003), 1181-5 ; E. Koechlin, T. Jubault, From action to human language : Broca's area and the hierarchical organization of behavior, *soumis.*

(ii) La zone X du cerveau est activée aussi bien lorsque l'événement mental A est observé que lorsque l'événement mental B est observé (“implication de X dans A et dans B”).

(iii) L'événement mental A est corrélé avec une activité accrue de la zone X dans le contexte psychologique C et de la zone Y dans le contexte psychologique D (“implication de X dans A quand C et de Y quand D”).

Ces résultats sont souvent présentés comme procédant d'une inspiration unique combinant neurobiologie et psychologie, caractéristique de la nouvelle spécialité dont ils relèvent. Ils ne se prêtent pas moins à deux lectures. Selon la première, les processus ou événements psychologiques sont supposés déterminés et connus, et les activations mises au jour en X, Y,... nous renseignent sur tel ou tel aspect du fonctionnement cérébral.. Selon la seconde, les événements cérébraux sont tenus pour compris, la signification qu'ils prennent dans les meilleures théories neuroscientifiques disponibles étant admise, et c'est notre compréhension de certains processus psychologiques qui pourrait s'en trouver améliorée.

Le soupçon de vacuité

Examinons maintenant certaines critiques adressées à ce type de travaux. Laissons de côté tout ce qui touche à la méthodologie et qui relève d'une compétence strictement neurobiologique: caractère indirect des mesures, artefacts dus aux algorithmes de traitement, faiblesse du signal, résolution insuffisante, variabilité individuelle et valeur douteuse des moyennages, non-standardisation des résultats, précarité des constats de non-activation (faux négatifs), etc. La présente discussion n'a d'intérêt que sous l'hypothèse que ces difficultés sont en voie d'être réduites. La question de savoir si une neuroimagerie infirme ferait progresser la connaissance de l'esprit ne se pose pas vraiment, puisqu'elle ne serait pas même en mesure de contribuer significativement à une science du cerveau. La question intéressante est de savoir si une neuroimagerie mûre et forte, en démultipliant l'efficacité des sciences du cerveau, peut révolutionner les sciences cognitives.

Les sceptiques pensent que non: selon eux la neuroimagerie ne fournit que des réponses sans questions, elle n'est qu'un ensemble de techniques « photographiques », elle accumule des faits sans théorie. Une partie de ces reproches est sans intérêt ici, pour la même raison que précédemment: qu'il y ait dans ce domaine en émergence des expériences mal conduites, des résultats sans grande portée, n'a rien d'exceptionnel. Ce qui nous intéresse est la portée de la *meilleure* neuroimagerie possible. C'est à son propos que le soupçon de vacuité doit être examiné.

Les cas les plus favorables sont ceux dans lesquels l'un des membres de l'“équivalence” X, Y,... ~ A, B... est effectivement identifié au-delà de tout doute raisonnable; ainsi des

expériences portant sur des actions motrices ou perceptives élémentaires, dont la dimension comportementale observable détermine sans réelle ambiguïté la nature mentale ou cognitive (le A, B... de l'équivalence); d'apprendre que X, Y,... sont impliqués dans le décours de A, B,... constitue alors un accroissement sensible d'intelligibilité quant aux propriétés de X, Y..., car A, B,... sont de "bons objets", des catégories naturelles dont les événements singuliers intervenant dans l'expérience sont des représentants. Plus généralement, les processus cognitifs rapides et essentiellement automatiques, associés à des comportements globalement peu modulables, sont un terrain favorable, comme l'attestent du reste les succès de cet ancêtre des neurosciences cognitives qu'est la neuropsychologie traditionnelle. Restent deux familles de cas: (1) ceux dans lesquels A, B,... sont censés être compris mais présentent une complexité ou une variabilité qui font douter qu'ils constituent vraiment des catégories naturelles dont la nature soit conceptuellement claire; (2) ceux dans lesquels ce sont les variables neurales X, Y... qui sont tenues pour fixes et les variables mentales A, B,... que l'on cherche à mieux cerner. Pour les cas de ce genre, qui n'ont rien de rare, l'interprétation des résultats est hasardeuse. Et c'est ici que le bon sens du neurobiologiste, si cultivé et génial qu'il puisse être, peut se trouver en défaut, et que l'interdisciplinarité trouve tout son sens.

Pourquoi, et à quelles conditions, ces résultats peuvent présenter de l'inférêt

Revenons à la situation simple, du type (i) ci-dessus: "L'événement mental A est corrélé avec l'activation de l'aire X". Ce résultat, disent les critiques, est trivial, car personne (dans le monde scientifique) ne doute que tout événement mental ait un corrélat neural, et un corrélat neural doit bien se situer quelque part dans le cerveau. Il n'y a donc pas de question scientifiquement plausible à laquelle un tel résultat pourrait répondre. Arrêtons-nous un instant sur cette objection.

Prenons d'abord le cas dans lequel la seule chose que l'on sache de X, c'est sa position topographique dans le cerveau. Il n'est pas trivial que A soit toujours corrélé avec une activité de X: cela tend à montrer que le type mental A possède un caractère naturel sur le plan neurobiologique. Si les progrès que l'on nous promet en matière de finesse de résolution se réalisent, et si l'aire X était identifiée sans doute possible comme étant homologue chez tous les sujets,on pourrait en arriver à une réfutation[12] du fonctionnalisme. De même, elle pourrait conduire à rejeter l'idée d'une absence de corrélat neurologique des troubles psychiatriques, notamment de pensées délirantes, qui définit encore aujourd'hui, malgré les effets connus, et partiellement compris, des psychotropes, une manière de ligne de démarcation du territoire du psychiatre,

[12] Réfutation partielle: d'une part la même fonction mentale pourrait être réalisée matériellement d'une façon différente chez d'autres espèces, ou chez des sujets anormaux, ou dans des systèmes artificiels; d'autre part, du seul fait qu'en réalité cette fonction n'est réalisée que d'une seule manière chez les sujets normaux on ne pourrait déduire qu'elle n'a pas une nature irréductivement fonctionnelle, c'est-à-dire qu'elle a d'autres propriétés que le rôle qu'elle joue dans l'économie mentale et comportementale.

Cependant, pour en revenir à la pratique actuelle, les résultats du type (i) le plus simple sont rarement très intéressants, et la plupart des expériences portent sur des situations dans lesquelles à X sont attachées des propriétés qui vont au-delà de la seule localisation spatiale. Les neurobiologistes ont accumulé des connaissances considérables sur les fonctionnalités des différentes régions du cerveau, et même si elles sont entâchées d'incertitudes, et sont constamment révisées, parfois de manière radicale, elles n'en ont pas moins suffisamment de consistance pour asseoir des résultats de la forme (ii) ci-dessus : l'aire X est impliquée aussi bien dans des tâches de type A que dans des tâches de type B. La leçon que l'on est tenté de tirer est qu'il y a une parenté entre les tâches de type A et les tâches de type B, et une telle parenté peut être un "scoop". Par exemple, l'hippocampe, avec ses structures associées, semble jouer un rôle double: pour la navigation et la mémoire spatiale (répérage égocentrique sans doute dans le lobe pariétal, et allocentrique dans l'hippocampe droit), d'une part; pour la mémoire épisodique et autobiographique (dans l'hippocampe gauche)[13], d'autre part. Que s'orienter dans l'espace et construire une image ou un concept de soi en tant que centre biographique soient liés (via leur soubassement neural respectif) est loin de résulter d'une simple analyse conceptuelle ou de l'introspection. Cependant, pour passer d'une hypothèse de ce genre, finalement assez floue, à une découverte psychologique précise, certaines conditions doivent être remplies: (1) L'activation de l'aire X doit être un phénomène unitaire, et non une conjonction fortuite de processus distincts. (Imaginons un détective qui aurait remarqué que chaque fois qu'un vol a lieu, les volets du 5^{e} étage d'un immeuble donné restent closes dans les heures précédant le crime, et que chaque fois que *Don Giovanni* est joué à l'Opéra, les volets du 6^{e} étage du même immeuble restent closes dans les heures précedant la représentation. Imaginons alors que le détective, par faiblesse de la vue ou de l'intelligence, classe sous la même rubrique X la fermeture des volets du 5^{e} et du 6^{e} étage de l'immeuble en question. Il serait tenté de conclure que les vols (A) et les représentations de *Don Giovanni* (B) sont "liés", alors qu'on pourrait conjecturer que le voleur habite le 5^{e} étage, le chanteur le 6^{e}, et que cette proximité est fortuite et dépourvue de signification.) (2) L'activation de X doit être, sinon l'unique, du moins le principal corrélat neural de A comme de B, à défaut de quoi le lien entre A et B pourrait être accidentel. (Imaginons que X soit activé uniquement lorsque de petits objets noirs sur fond blanc apparaissent dans le champ visuel, que A soit la lecture, et B l'exécution d'un morceau de violon. A et B ont en commun, dans nos cultures et dans les circonstances habituelles, de mettre en jeu, au stade de l'apprentissage ou de l'exécution, la perception de lettres ou de notes imprimées en noir sur fond blanc. Pourtant, A et B ne sont pas liés (du moins en première analyse). Pour les non-voyants, comme dans une culture où les partitions seraient imprimées en rouge sur fond noir, ou encore dans une culture sans notation musicale, X, on

[13] N. Burgess, E.A. Maguire & J. O'Keefe, The human hippocampus and spatial and episodic memory, *Neuron* **35** (2002), 625-641.

peut le présumer, ne serait pas activé par A et par B. Un neuroimageur qui ne disposerait pas d'un accès indépendant à ces distinctions psychologiques, anthropologiques ou comportementales, et qui s'en tiendrait à ses images, serait incapable de tirer une conclusion valide, et n'aurait de choix qu'entre le silence et l'erreur.

Il ne s'agit pas ici d'affirmer que les conditions qui viennent d'être énoncées ne sont jamais réalisées, ni que les spécialistes ne sont pas conscients de ce genre de réquisits, mais seulement d'insister sur le fait qu'à eux seuls, les résultats d'imagerie de type (ii) ne suffisent pas à conclure quoi que ce soit concernant le lien entre A et B, *sauf* parfois, mais alors ce n'est pas rien, sur le plan clinique: on peut penser que des patients privés, par exemple du fait d'une lésion cérébrale, de la fonctionnalité A, le seront également de la fonctionnalité B, ce qui peut être d'une grande importance pour le diagnostic, le pronostic et la thérapeutique.

Les résultats concernant l'imagerie mentale mentionnés au début de la présente section illustrent bien la situation. Le problème de départ est de savoir si l'imagerie mentale A met en œuvre des processus ou représentations iconiques. L'imagerie fonctionnelle montre que A est corrélé à une activation de X, et l'on sait que X est impliqué dans la vision (B). Si l'on admet (ou que l'on stipule) que la vision fait intervenir à son tour des processus iconiques, on est tenté de conclure que A est de nature iconique et que "le débat est tranché". Mais il faut aussi s'assurer que le cortex visuel primaire est uniquement activé lorsque des représentations iconiques sont manipulées, et que l'imagerie mentale ne déclenche pas V1 pour une raison indirecte (par exemple, si l'imagerie était accompagnée d'un écho, non essentiel pour l'accomplissement de la tâche, en sorte qu'elle serait réalisable malgré une inhibition temporaire de V1 par SMT[14])[15]. A ces difficultés générales s'ajoute dans ce cas la question du bon « découpage » : où faire passer la frontière entre représentations iconiques et représentations symboliques, sachant que l'on se place au niveau des processus eux-mêmes, et non sur le plan émergent des représentations conscientes ou de la reconstruction conceptuelle : de cette difficulté-là il serait téméraire d'affirmer qu'elle est aujourd'hui surmontée.

Descendre au niveau du neurone individuel: le cas des neurones-miroirs

Pour écarter la menace d'une conjonction fortuite, au sein d'une même aire, de deux processus essentiellement indépendants, on peut procéder (mais seulement chez l'animal) à des enregistrements de neurones individuels (*single neuron recordings)*. Il semble en effet

[14] Stimulation magnétique transcrânienne. V. p. ex. *PSN* **I**, 1 (2003), 40-43 ou E.M. Robertson, H. Théoret & A. Pascual-Leone, Studies in cognition : The problems solved and created by transcranial magnetic stimulation, *J. of Cognitive Neuroscience***15** (2003), 948-960.

[15] Il ne s'agit évidemment pas ici d'une discussion scientifiquement réaliste de cette question complexe et controversée, dans laquelle interviennent quantité d'autres considérations et que l'auteur n'a pas compétence à mener. Mais précisément l'un des arguments développés dans cet article est qu'il n'y a en général pas de résolution simple par la seule neuroimagerie de questions psychologiques intéressantes.

plausible que si deux processus cognitifs A et B s'accompagnent de la décharge X d'un même neurone, ils sont liés. Ici encore, il faut supposer que la décharge du neurone joue un rôle essentiel dans les processus en question, sans toutefois qu'elle puisse à elle seule caractériser à la fois A et B, auquel cas ou bien il s'agirait en réalité d'un seul événement cognitif, ou bien A et B ne se distingueraient *que* par une propriété non cérébrale (dans le jargon des philosophes, le mental ne "surviendrait" pas sur le cérébral) –deux éventualités qu'on peut écarter. Un exemple qui fait couler beaucoup d'encre depuis quelques années est celui des "neurones-miroirs"[16]. Certains neurones prémoteurs de l'aire F5 du cortex des macaques ont la propriété de décharger aussi bien lorsque l'animal tend la main pour prendre des cacahuètes disposés devant lui que lorsqu'il observe un expérimentateur (ou un autre macaque) en faire autant. Des systèmes miroirs analogues semblent bien exister chez l'homme (dans l'aire de Broca qui est homologue de l'aire F5 du macaque). C'est une découverte[17] que beaucoup considèrent comme capitale, car elle fournirait selon eux une explication naturaliste directe de la compréhension de l'action intentionnelle d'autrui: en observant le mouvement de l'autre, je comprends ce qu'il veut faire, en vertu du fait que ce geste provoque chez moi une disposition à agir dont la nature intentionnelle m'est révélée par l'anticipation que crée cette disposition. C'est l'idée de base de la *théorie motrice* de la cognition sociale; pour le dire rapidement, je te perçois comme un *agent* semblable à moi, capable d'entretenir des intentions analogues, parce que mon cerveau, d'un côté, identifie ton geste et le mien, et de l'autre me fait connaître le sens de mon propre geste en me permettant d'anticiper ses conséquences. La discussion ne peut être même résumée ici, mais on peut en retenir quelques leçons:

(1) L'existence des neurones-miroirs est significative, et ne pouvait être prédite à partir de considérations seulement théoriques; elle n'est pas empiriquement triviale.

(2) Sa portée dépend crucialement de l'hypothèse qu'il s'agit de neurones prémoteurs, donc remplissant une fonction que l'on a préalablement identifiée avec certitude (à savoir la préparation du mouvement), et cette hypothèse à son tour s'enracine dans des strates épaisses de théorisation tant en psychologie qu'en neurobiologie qui ne doivent rien à l'imagerie.

(3) Sa signification exacte est encore loin d'être claire, et dans les débats en cours les imageurs sont amenés à se placer sur le terrain psychologique, philosophique, voire linguistique, où ils rencontrent des spécialistes de ces disciplines[18]; ils n'ont pas le monopole de la compétence.

---

[16] Rizzolatti, G., Fadiga, L., Gallese, V. and Fogassi, L. Premotor cortex and the recognition of motor actions. *Cog. Brain Res.*, **3** (1996): 131-141; Gallese, V., Fadiga, L., Fogassi, L. and Rizzolatti, G. Action recognition in the premotor cortex. *Brain* **119** (1996): 593-609

[17] V.S. Ramanchandran par exemple y voit la plus importante découverte de la décennie en neurosciences: Edge (2000) – www.edge.org/3rd_culture/ramachandran.

[18] V. par exemple le colloque en ligne sur www.interdisciplines.org.

(4) Dans la mesure où une théorie substantielle stable des neurones-miroirs émergera, elle constituera une contribution non triviale des neurosciences dans leur ensemble aux sciences cognitives, en sorte que...

(5) ...neurosciences et psychologie auront à parts à peu près égales apporté ainsi à la connaissance de l'esprit humain une contribution dont le caractère non trivial est attesté par le fait qu'elle fournirait une explication de la capacité à reconnaître *directement et non intentionnellement* l'intention d'autrui à partir de ses seuls mouvements, invalidant ainsi une conception solidement ancrée dans le sens commun comme dans certaines traditions de psychologie savante (le caractère inférentiel et indirect de cette reconnaissance).

Vers une carte ou un organigramme cérébral?

Plus s'accumulent les résultats de corrélation que nous venons d'examiner, moins on peut se satisfaire d'une simple juxtaposition. La question se pose de savoir comment les articuler. A quel genre de connaissance ou de représentation du cerveau concourent-ils? S'ils avaient tous la forme élémentaire "A ~ X", A étant une fonction cognitive ou comportementale donnée d'entrée de jeu, et X l'activation d'une aire cérébrale, et s'ils étaient simplement juxtaposés, leur assemblage ne constituerait qu'une nouvelle "carte" phrénologique, dont l'intérêt se limiterait à certaines situations cliniques, et dont la portée théorique serait faible.

L'erreur de certains sceptiques, tels Fodor[19], est de croire que c'est l'unique objectif que puissent viser les neurosciences cognitives adossées à l'imagerie. En réalité, elles disposent de résultats des trois types ci-dessus, qui se prêtent à une combinatoire et à un enchevêtrement de fonctions: à la pure mise en correspondance de type (i) simple–une fonction, une aire– s'ajoutent avec le type (i) complexe –une fonction, plusieurs aires, voire plusieurs processus neuraux concomitants– des décompositions en éléments plus simples, avec le type (ii) des partages de ressources (ii), et, avec le type (iii) des différentiations,. On n'a donc plus affaire à une simple carte, mais à un système, à un "organigramme" cérébral. Mais évidemment la constitution de l'organigramme à partir des résultats partiels fait appel à des connaissances et à un art de la conjecture qui ne doivent rien à la neuroimagerie, et tout aux neurosciences fondamentales et à d'autres branches des sciences cognitives.

Reste qu'il semble difficile de contester l'intérêt d'une théorie "organigrammatique" du cerveau, quels qu'en soient les auteurs. Le sceptique ne doit-il pas se rendre? Il devrait au moins être amené à réfléchir. Mais il lui reste des portes de sorties. Il peut d'abord rappeler qu'on est encore bien loin du moment où une telle théorie sera disponible, et soupçonner qu'elle ne le sera peut-être jamais: l'IA n'avait-elle pas été victime, en son temps, de l'illusion

[19] V. p.ex. J. Fodor, Let your brain alone, *London Review of Books*, **21,** 19, 30 sept. 1999. La critique de Fodor s'applique sans doute à certains programmes de recherche, qui semblent en effet se contenter de nourrir une néophrénologie. Nous nous intéressons ici, nous l'avons dit, à la meilleure contribution possible de l'imagerie à la connaissance de l'esprit.

du premier pas (penser que monter à un arbre nous rapproche inexorablement du moment où nous pourrons monter à la lune)? Il peut aussi réaffirmer sa conviction fonctionnaliste et maintenir, pour des raisons théoriques, le principe d'une séparation de deux ordres d'intelligibilité: aucun organigramme du cerveau, qui par définition ne met en jeu que des causes, ne peut nous éclairer sur la nature et sur la dynamique de l'esprit, siège de significations et de raisons. Ou enfin, si le fonctionnalisme lui semble reposer lui-même sur des confusions, il peut contester que la neuroimagerie fournisse des explications, car *ce qui est à expliquer* (le domaine que vise la psychologie) demeure entouré d'un épais brouillard[20].

3. Les neurosciences sont une discipline théorique

Si la discussion précédente a permis, on peut du moins l'espérer, de clarifier les enjeux, elle laisse néanmoins planer le doute. D'un côté, on croit comprendre que la neuroimagerie contribue incontestablement à la connaissance du cerveau, et qu'ainsi elle apporte aux sciences cognitives une moisson de faits qu'elles ne peuvent ignorer. D'un autre côté, on nous invite à relativiser son apport, soit en soulignant sa dépendance vis-à-vis d'autres disciplines, soit en niant qu'elle puisse amener les sciences cognitives à clarifier, réviser ou enrichir notablement leurs conceptions de l'esprit.

Désaccords théoriques sur fond de rivalités institutionnelles? Sans doute, mais il y a également une difficulté épistémologique. Le développement rapide de la neuroimagerie semble encourager l'idée que les *faits,* qui sont concrets et solides, s'opposent aux théories, abstraites et fragiles, et que les neurosciences, grâce à l'imagerie, fournissent les faits permettant à la science de l'esprit/cerveau d'échapper à l'incertitude des théories, produit d'une psychologie encore trop asservie au modèle spéculatif de la philosophie et d'un certain mode idéologique de pratiquer les sciences de l'homme. Ce "neurofactualisme" (ou "neuropositivisme"[21]) doit être clairement récusé: les neurosciences sont une science théorique, ce qui signifie notamment que les connaissances qu'elles produisent sont, au sens technique, des *théories.* Les théories vont au-delà des faits; ceux-ci les sous-déterminent; les faits eux-mêmes sont imprégnés de théorie; les hypothèses d'une théorie confrontent collectivement le tribunal de l'expérience; enfin les théories, ainsi que les explications qu'elles

[20] Un sceptique de ce genre s'inspirerait probablement de Wittgenstein, ou peut-être de la phénoménologie existentielle issue de Heidegger, ou encore de philosophes pragmatiques contemporains.

[21] Un terme qu'il vaut mieux éviter, car il prête à confusion: il peut également servir à une certaine psychologie "humaniste" pour discréditer l'entreprise des neurosciences elles-mêmes, en tant que science de l'homme. En rejetant le "neurofactualisme", je n'entends nullement me ranger dans le camp des "humanistes" (le véritable humanisme se place à un autre niveau): je m'en prends seulement à une conception naïve de la nature et du rôle des faits dans les sciences, abandonnée depuis belle lurette s'agissant des sciences mûres, mais retrouvant du service dans une discipline immature telle que les sciences cognitives.

suggèrent, font intervenir des entités "cachées", c'est-à-dire non directement accessibles à l'expérience.

Affirmer que les neurosciences sont une discipline théorique, c'est donc dire qu'elles possèdent ces traits. Pour être plus précis, elles sont une branche d'une science théorique, à savoir la biologie, et elles héritent de leur caractère théorique. Quant à ce qu'on appelle (depuis peu) "neurosciences cognitives", ce n'est pas, contrairement aux apparences lexicales, une simple spécialité au sein des neurosciences, mais un programme de recherche, ou mieux une toute jeune "matrice disciplinaire", pour emprunter à Thomas Kuhn un terme qu'il a finalement préféré au trop versatile "paradigme"[22]. On peut maintenant préciser les raisons de ce qui a pu apparaître au début de cet article comme une simple précaution verbale, voire une chinoiserie. Les neurosciences cognitives relèvent non pas d'une seule discipline fondamentale, mais de plusieurs, dont l'articulation appelle des décisions qui ne s'imposent pas également à tous. Elles ont de fait une manière déterminée d'aborder leur objet, fondée notamment sur l'option méthodologique de l'imagerie et sur le privilège donné à un certain type d'explication (celle dont on distinguait les types (i), (ii), (iii) au début de la section 2). Ce sont précisément ces décisions théoriques qui leur fournissent les moyens de constituer une représentation de leur objet qui pourrait (une hypothèse que l'on précisera dans un moment) modifier profondément la donne dans les sciences cognitives. Enfin, les neurosciences cognitives ne constituent pas une *théorie* unique, au sens strict, car elles produisent et accueillent une pluralité de théories particulières, qui diffèrent entre elles soit par les phénomènes qu'elles visent, soit par les hypothèses qu'elles défendent dans un même domaine, soit des deux façons.

La juste place de la neuroimagerie et neurosciences cognitives

Sous cet angle, certaines de nos perplexités se dissipent. En replaçant l'imagerie au sein des neurosciences, et les neurosciences au sein de la biologie, on voit que les sciences cognitives ne sont pas davantage fondées à récuser la contribution de la neuroimagerie que les astronomes du début du XVIIe siècle ne le sont à refuser de regarder dans la lunette de Galilée: s'ils veulent élaborer une théorie de la cognition, ils ne peuvent se dispenser de chercher à comprendre en vertu desquelles de ses innombrables propriétés le seul système dont nous sachions à coup sûr qu'il est cognitif l'est effectivement, ni rejeter a priori l'avis globalement favorable des spécialistes du cerveau sur la fiabilité et l'utilité des instruments de neuroimagerie. Mais on voit aussi que *cet avis, scientifiquement motivé, est une garantie essentielle*: les contemporains de Galilée avaient raison de s'enquérir sur la véridicité et sur les conditions d'emploi de la lunette; ils avaient également raison d'exiger des arguments en

[22] T. Kuhn, *The Essential Tension*, University of Chicago Press, 1977 ; trad. fr. *La tension essentielle*, Paris : Gallimard, 1990, chap. 12 (article original paru en 1974).

faveur des thèses copernico-galiléennes qui soient indépendants des images procurées par la lunette, et jamais Galilée n'a pensé fonder la défense de ses théories sur ses seules observations. Enfin, il est clair que les faits que fournit la neuroimagerie sont imprégnés de théorie, d'une part en raison du caractère hautement indirect de la méthodologie, d'autre part pour la raison générale illustrée par l'exemple précédent: jamais la lunette à elle seule n'aurait permis au premier astronome médiéval venu de conclure à la vérité du copernicanisme. Ce sont là des banalités, c'est bien pourquoi il ne faut pas les perdre de vue.

Une question reste en suspens, celle qui figure dans le titre du présent article: les neurosciences cognitives ont-elles un potentiel révolutionnaire au sein des sciences cognitives? Avant d'y répondre, remarquons que seule une matrice disciplinaire peut révolutionner une science: c'est pourquoi il est préférable de conserver l'étiquette "neurosciences cognitives" pour désigner une certaine matrice disciplinaire, et non comme le proposent certains, l'ensemble des recherches en sciences cognitives revendiquant un lien privilégié avec les neurosciences, voire les sciences cognitives dans leur totalité au prétexte qu'elles *devraient* faire allégeance aux neurosciences. (Il est possible que la distinction devienne oiseuse dans un avenir proche ou lointain; ce serait justement le signe d'un ralliement général de la discipline à la matrice en question.)[23]

Un potentiel révolutionnaire?

Se demander si les neurosciences cognitives révolutionnent les sciences cognitives, ou peuvent peut-être le faire, c'est d'abord, pour le redire, leur assigner une place séparée (sans nier leur imbrication, puisque les neurosciences ont toujours été considérées comme un membre actif de la fédération). C'est ensuite exprimer, au risque de fâcher certains éminents collègues, l'opinion que ce n'est pas encore le cas. Mais c'est aussi, à l'inverse, aller au-delà de ce que leurs partisans modérés se contentent de revendiquer pour elles: un rôle dans une construction commune déjà bien engagée – apporter à point nommé des réponses à des questions formulées par les autres disciplines. Révolutionner une science, c'est *aussi* répondre à des questions qu'elle a été incapable de poser.

Parmi ces questions, certaines sont exprimables mais ne se posent pas, car dans le cadre conceptuel considéré on ne peut imaginer une expérience qui imposerait, ou suggérerait seulement une réponse. Un exemple est celui-ci: existe-t-il des représentations mentales? Dans le cadre classique des sciences cognitives, qui définissent la cognition comme une dynamique de représentations internes, on voit mal quelle épreuve empirique conforterait une réponse quelle qu'elle soit. Daniel Amit, il y a dix ans, a remarqué qu'une série d'expériences d'enregistrement de neurones dans le cortex temporal du singe avait permis à

[23] On peut aussi comprendre que les administrations de la recherche préfèrent dès à présent la solution extensionnelle.

Miyashita de mettre en évidence des états neuraux (plus précisément des assemblées de neurones) remplissant les fonctions essentielles attribuées a priori aux représentations internes[24]. Pour le dire autrement, les représentations internes comme nécessité conceptuelle (dans un cadre donné) ou comme métaphore sont remplacées par des phénomènes empiriquement repérables remplissant les mêmes fonctions.

D'autres questions ne sont pas même formulables, car elles nécessitent un vocabulaire dont la science constituée ne dispose pas. Les neurosciences cognitives nous présentent à certains moments une image étrange, qui n'est ni l'image familière du cerveau, ni l'image familière de l'esprit. L'étrangeté ne tient pas à des différences repérables par rapport aux conceptions admises. Elle est due à ce qu'on commence à appeler les "actions à distance": la structure corticale serait radicalement différente de tout ce que la philosophie et la psychologie peuvent imaginer comme "découpage" concevable de l'esprit, en sorte que ce qui semble, dans la conception classique, constituer une unité fonctionnelle, sinon anatomique, est en réalité un simple *pattern* résultant de l'action conjugée de systèmes cérébraux éloignés; inversement, le même système pourrait contribuer à des processus extrêmement différents, conceptuellement très distants. C'est ce que désigne parfois l'expression ambiguë de *complexité fonctionnelle.*

Non seulement donc, comme les connexionnistes l'avaient envisagé, les *représentations* cérébrales seraient distribuées[25], mais il en serait de même des *fonctions* cérébrales. Ainsi, les schèmes caractéristiques de l'imagerie formulés au début de la section 2 ne seraient que des échafaudages provisoires, car les fonctions cognitivo-comportementales A, B,... n'entreraient pas dans des relations systématiques, si complexes soient-elles, avec des processus neuraux X, Y,..., ce dont on s'apercevrait en échouant dans les tentatives de combiner ces résultats en une "neurocarte" légendée dans le vocabulaire psychologique.

Un exercice de prospective scientifique: neurophilosophie ou émergence d'un niveau macroneurobiologique autonome?

Des prémices de cette situation sont déjà apparus, et elles suffisent à montrer que les neurosciences cognitives bousculent dès à présent l'assise des sciences cognitives. Mais rien pour le moment n'indique que les réajustements nécessaires soient hors de portée. La question que je voudrais poser pour conclure, c'est de savoir si les neurosciences cognitives

[24] D.J. Amit, The Hebbian paradigm reintegrated: Local reverberations as internal representations, *Behavioral and Brain Sciences* **18**, 4 (1995), 617-626; Y. Miyashita & H.S. Chang, Neural correlate of pictorial short-term memory in the primate temporal cortex, *Nature* **331** (1988) 68

[25] V. p. ex. G.E. Hinton, J.L. McClelland & D.E. Rumelhart, Distributed representations, *in* D.E. Rumelhart, J.L. McClelland & the PDP Research Group, *Parallel Distributed Processing*, vo. 1, Cambridge, MA : MIT Press, 1986, 77-109. Voir aussi, sur ce point mais aussi pour la perspective générale qu'il propose, P. Smolensky, On the proper treatment of connectionism, *Behavioral and Brain Sciences* **11** (1988), 3-71 ; trad. *in* G. Fisette & P. Poirier, dir., *Philosophie de l' esprit*, vol. 2, Paris : Vrin, 2002.

pourraient déboucher sur une révision si radicale de l'ontologie du mental qu'elle provoquerait une déconnexion de leur domaine avec la psychologie telle que nous la connaissons, déclenchant ainsi une révolution dans les sciences cognitives.

Cette situation n'a-t-elle pas été depuis longtemps envisagée, voire prédite avec assurance par Paul et Patricia Churchland sous le nom de "neurophilosophie"[26]? Il y a deux grandes différences entre leur vision et celle qui vient d'être esquissée. Leur thèse était que les neurosciences fourniraient une image scientifique unifiée de l'"esprit/cerveau", qu'elles constitueraient à la fois une science du cerveau (comme leur nom l'indique) et une science de l'esprit, autrement dit une psychologie scientifique. Au contraire, ce que j'envisagé ici, à titre de pure hyothèse, comme horizon des neurosciences cognitives, est la constitution de *deux* représentations ou niveaux descriptifs au sein d'une branche de la biologie, une neuroscience fondamentale et une "neurocognitologie" ou "macroneurobiologie" . Certes, l'idée de deux niveaux de description neurobiologique n'a rien d'original, elle est présente dans l'organisation même de la recherche et dans les dénominations « neurosciences intégratives » ou « neurobiologie fonctionnelle » opposées à « neurobiologie moléculaire et cellulaire ». Dans ce contexte, le niveau supérieur ou intégratif a vocation à entrer en correspondance ou à s'articuler systématiquement avec la description psychologique.

Ce que proposaient et proposent toujours les Churchland, c'est d'éliminer purement et simplement la description psychologique, plus précisément de considérer que la description macroneurobiologique se substituera graduellement à l'actuelle psychologie fonctionnaliste, qui tendrait vers l'extinction. Le scénario proposé ici présente une première différence avec le leur : d'une part, il donne au niveau supérieur ou macro une certaine autonomie ontologique relativement au niveau micro ; mais d'autre part, il ne postule nullement le dépérissement de la psychologie scientifique comme science du mental. Au contraire, et à la différence également avec la conception « correspondantiste-articulatoire » (le schéma unificateur mais non réductionniste selon lequel les descriptions articulées aux niveaux fonctionnel et cérébral présentent des correspondances systématiques dont la mise au jour étanche la soif d'explication, la possibilité se dessine qu'entre macroneurobiologie et psychologie cognitive existerait une complémentarité, au sens où Bohr l'a proposé pour la mécanique quantique, interdisant l'accès simultané aux descriptions macroneurobiologique et psychologique d'un même processus.[27].

---

[26] Paul M. Churchland, Eliminative materialism and the propositional attitudes, *J. of Philosophy* **78** (1981), 67-90; – *A Neurocomputational Perspective*, Cambridge, MA : MIT Press, 1989. Patricia Smith Churchland, *Neurophilosophy.Towards a unified science of the mind/brain*, Cambridge, MA : MIT Press, 1986; trad.fr. *Neurophilosophie. L'esprit-cerveau*, Paris: PUF, 1999.

[27] Le transfert du concept de complémentarité de Bohr vers d'autres domaines a été récemment proposé par Rom Harré, Resolving the reduction/emergence debate, *in* M. Kistler, dir., *Emergence and Reduction*, numéro spécial de la revue *Synthèse*, à paraître. (V. aussi mon commentaire.)

La seconde différence avec la « neurophilosophie » concerne la « psychologie naïve » ou de sens commun (*folk psychology*), que les Churchland vouent également à l'élimination[28] . Au contraire, dans le scénario proposé, comme chez un philosophe fonctionnaliste partisan du « réalisme intentionnel » comme J. Fodor, la psychologie scientifique pourrait l'abriter, ou du moins en abriter une partie, convenablement amendée ; mais il se pourrait aussi qu'elle constitue une troisième sphère d'intelligibilité. Elle serait ce que le philosophe Wilfrid Sellars appelait l'« image manifeste »[29] de l'esprit, l' « image scientifique » étant alors composée de deux représentations complémentaires (au sens fort), macroneurobiologique et (proprement) psychologique d'autre part[30].

Contrairement aux Churchland, je ne prétends pas que le scénario que je viens d'esquisser soit nécessaire ou seulement probable. Il se pourrait que les neurosciences cognitives ne détiennent aucun potentiel révolutionnaire, et qu'elles se contentent d'un rôle analogue à celui qu'elles remplissent aujourd'hui au sein des sciences cognitives. Pour revenir à la comparaison initiale, comme l'IA première manière, elles auront alors été une matrice disciplinaire qui doit renoncer à s'emparer d'un domaine de la nature pour endosser le statut plus modeste de branche spécialisée. Il subsisterait néanmoins une différence de principe entre les deux programmes. Les neurosciences cognitives sont une science empirique, l'IA mûre est une discipline formelle ou structurale. Les liens de l'IA avec les sciences cognitives sont contingents, mais inversement son domaine d'application peut à tout moment s'étendre à d'autres systèmes que le cerveau, dès lors qu'ils abritent une forme même dérivée ou rudimentaire d'intentionnalité. Les neurosciences cognitives, en tant que spécialité, sont au contraire rivées au système nerveux central, mais elles possèdent une dimension supplémentaire qui leur est propre, la compétence clinique. Si Diderot[31] a raison de penser que seuls les médecins peuvent faire de la bonne métaphysique, les neurosciences cognitives sont peut-être, tous compte faits, la reine des sciences cognitives. Et s'il a tort, il leur restera sans doute l'éminente dignité de servir, à leur rang, à la fois la médecine et les sciences pures de l'esprit.

---

[28] Pour les spécialistes : la position des Churchland est présentée ici à l'envers. C'est d'abord la psychologie naïve qu'ils veulent éliminer au profit des neurosciences, et la psychologie scientifique cognitiviste ou fonctionnaliste ne mérite l'élimination que parce qu'elle est, selon eux, indissociable de la psychologie naïve.

[29] W. Sellars, *Science, Perception and Reality*, Londres : Routledge & Kegan Paul, 1963 ; trad. fr. *in* G. Fisette & P. Poirier, dir., *Philosophie de l' esprit*, vol. 1, Paris : Vrin, 2002.

[30] Une situation qu'on retrouve justement dans la connaissance des phénomènes physiques fondamentaux.

[31] *Encyclopédie*, article « Locke ».