

Modèles du langage et de la pensée

cours de DEA "Sciences de l'Éducation"
1997/98

Alain Lecomte

1- Introduction: une théorie computationnelle de l'esprit?

1.1. Langage et pensée (inférence)

Ce cours est consacré aux modèles. Initialement aux modèles du langage et du raisonnement, plus généralement aux modèles de l'esprit. Qu'est-ce que l'esprit au sens de "mind" en anglais? Face à cette question, on peut ou bien l'attaquer de front et donner une définition de l'esprit, ou bien mettre entre parenthèses une telle définition et se concentrer plutôt sur les propriétés fondamentales. Une façon de voir les choses : quelles sont les propriétés fondamentales de l'esprit? quels sont les outils formels pour les analyser? Ce qui nous intéresse ici principalement ce sont les capacités *cognitives* de l'esprit. On pourrait aussi bien s'intéresser aux sensations, aux affects et à toutes sortes d'"événements" qui ont leur siège dans l'esprit, hors de la sphère intellectuelle proprement dite. Du reste, il est des auteurs qui ont critiqué Piaget au nom de ce qu'il avait trop sous-estimé l'importance de l'affectivité pour les activités intellectuelles du sujet humain. Nous laisserons néanmoins encore cette question ouverte (avec les questions qui iront de pair comme: un ordinateur peut-il souffrir etc.). **Les capacités cognitives de l'esprit incluent deux capacités fondamentales : celle de faire des inférences, et la faculté de langage.**

Souvent, ces deux capacités sont liées, il y a inférence dans la langue, par exemple: de *Marie ne fume plus*, j'infère: *Marie fumait* (effet dit "de présupposition"). Mais il y a aussi inférence sans langage, par exemple la fumée comme indice du feu (distinguer indice et symbole). Peut-il y avoir langage sans inférence? un usage purement mélodique de la langue par exemple....

Une question qu'on peut néanmoins se poser : quand nous parlons de langage, s'agit-il toujours du langage "public" c'est-à-dire de la capacité de communiquer? Peut-être pourrait-il exister une pensée n'utilisant qu'un langage "privé", "interne" (cf. hypothèse du mentalais). On doit ici faire une différence entre *langage* et *langue*. "Langage" désigne une capacité de symboliser, c'est un concept en quelque sorte philosophique. "Langue" désigne les systèmes de signes que nous utilisons pour mettre en oeuvre la capacité langagière, c'est un concept d'une science particulière: la linguistique: voir à ce sujet Saussure, et Chomsky. Nous reviendrons sur Chomsky plus loin, dont la position devient de plus en plus nuancée.

Ainsi, si nous évoquons la possibilité d'un langage privé, cela ne signifie pas nécessairement que celui-ci coïncide avec une langue, telle que le français ou l'anglais (ou le navajo...) mais qu'il y aurait une sorte de disposition interne à symboliser, servant en quelque sorte de fondement ultérieur au développement de la langue. Cette thèse est abondamment développée dans *The Language Instinct*, le livre de Steven Pinker (1994). Ceci dit, cette disposition innée semble être indépendante par rapport aux autres facultés de l'esprit. On a beaucoup glosé à propos de la fameuse "Hypothèse de Whorf-Sapir" selon laquelle notre langue déterminerait notre façon de pensée. L'idée était que chaque langue projetait, pour ses locuteurs, une sorte de grille, de système de catégories à travers lesquels ils voyaient le monde. On proposait comme exemple l'existence, dans la langue inuit, de toute une gamme de mots différents pour désigner la neige dans différents états, ou bien le fait que, semblait-il, différentes langues découpaient le spectre continu des couleurs de manières différentes et arbitraires. De là venait vite l'idée que le langage conditionnait la pensée: George Orwell, dans 1984, décrivait un monde où la langue s'était modifiée de telle manière que les mots correspondant aux concepts de liberté et de justice par exemple ne pouvaient plus être utilisés dans des contextes tels que "liberté individuelle" ou "liberté d'expression". Il en résultait un monde où, par voie de conséquence, les concepts en question avaient disparu. Pinker, dans le livre cité ci-dessus, règle leur compte à ces théories qu'il qualifie de mythes. Les soi-disant observations sur lesquelles elles seraient fondées s'avèrent en général ne pas résister longtemps à une analyse plus poussée. Que les inuits aient tout un vocabulaire pour parler de la neige, quoi de plus normal puisque c'est leur univers... Quant aux couleurs, une étude plus poussée mettrait en doute l'arbitraire supposé du découpage. On sait depuis que, si le spectre des couleurs est continu du point de vue physique de la longueur d'onde, ce fait a relativement peu d'importance pour la perception: celle-ci est guidée par les trois sortes de cônes présents dans l'oeil, qui réagissent chacun à un pigment différent, ces pigments correspondant aux couleurs élémentaires.

En fait, de nombreuses considérations (voir plus loin, §3) conduisent à admettre l'hypothèse qu'il existe une activité de pensée indépendante du langage.

1.2. L'intentionnalité

L'étude de ces capacités n'épuise évidemment pas la question "qu'est-ce que l'esprit?". Joëlle Proust a une manière amusante de faire sentir la profondeur de la question: elle demande: "les animaux ont-ils un esprit?". Cette question nous oblige à nous interroger au moins sur les critères permettant de répondre à une telle question. Il y a une première façon, qui consiste à dire qu'il y a esprit lorsqu'il y a conscience de certains états appelés états mentaux, c'est-à-dire que le porteur de cet esprit ressent les états mentaux qui se déroulent dans l'esprit, et qu'il les ressent à la

première personne. Si là est le critère déterminant, nous ressentons bien le malaise: personne n'est en mesure d'attribuer les propriétés d'un esprit à autre qu'à lui-même. (T. Nagel avait écrit un article, intitulé: "Quel effet cela fait-il d'être une chauve-souris?" afin de mettre en évidence l'impossibilité de répondre à une telle question).

Alors un autre critère repose sur la capacité de contrôler ses comportements au moyen de représentations. Avoir un esprit c'est être capable de former des représentations. Cela signifie que nos états mentaux représentent des états de choses extérieurs, qu'ils se forment afin de désigner ces états de choses. Cette propriété de représentation qu'ont les états mentaux est aussi appelée, depuis Brentano : *intentionnalité*.

Il faut ici s'arrêter un moment : après tout, nous connaissons des systèmes contenant des états qui représentent des états de choses sans que pour autant nous leur attribuions la propriété d'avoir un esprit. Un ordinateur travaille avec des symboles, une syntaxe qui représentent des objets. Mais on voit immédiatement que dans ce cas, la relation de représentation est conférée de l'extérieur par un agent, le programmeur, qui assigne des significations à des symboles. Le propre de la représentation pour l'attribution de l'esprit réside dans ce que la relation à la signification (c'est-à-dire la propriété sémantique du symbole) s'origine dans le sujet lui-même... à moins évidemment de laisser entendre qu'il est lui-même programmé (à la façon du rêveur dans la nouvelle de Borges). C'est cela, à proprement parler, l'intentionnalité: un rapport orienté entre un "sujet" et un "objet", ou plus précisément, un rapport d'un sujet "dirigé vers" un objet. Selon Brentano, l'intentionnalité est le caractère qui différencie les phénomènes psychiques des phénomènes physiques, elle est donc "inéliminable" dès qu'on prétend analyser la conscience.

La question centrale d'une théorie de l'esprit est donc celle de l'origine de l'intentionnalité, c'est-à-dire de la manière dont les propriétés sémantiques émergent. Cette question a été beaucoup discutée ces dernières années et a été traitée dans de nombreux livres et par de nombreux auteurs, citons F. Dretske, E. Pacherie, J. Proust, P. Engel et P. Jacob. Nous y reviendrons dans l'un des cours prochains, en abordant ce qu'on appelle la "sémantique informationnelle".

1.3. Le dualisme

Avant d'aller plus loin, je crois qu'il est important de situer un certain nombre d'autres concepts que nous rencontrerons souvent dans ce cours, ainsi que quelques oppositions fameuses.

On doit d'abord se souvenir que la discussion sur l'esprit et la conscience remonte à très loin. Arrêtons-nous à Descartes. Il y a un changement philosophique important entre Montaigne et Descartes (cf. là-dessus J. Proust) à propos de la question sur l'esprit des animaux. A la doctrine continuiste du premier (pour lequel les animaux communiquent entre eux tout comme les hommes bien que nous ne comprenions pas leur langage), le second oppose l'idée d'une coupure radicale: c'est la fameuse thèse des

"animaux-machines": il manque aux animaux la capacité "de raisonner" qu'ont les hommes, ils n'ont pas d'esprit. Si les animaux "communiquent" c'est comme les végétaux par des propriétés biologiques innées qui font que des signaux sont émis mais sans intention de communiquer. Les signaux en question sont des réactions immédiates à un danger ou à un objet quelconque qui se trouve présent. Enlevez la présence de l'objet, la communication disparaît, alors qu'au contraire le langage humain se caractérise par le fait d'avoir lieu en l'absence des objets ou des événements qui sont évoqués. Les animaux sont donc des machines, des machines physiques, et les hommes en revanche, échappent à ce machinisme, il y a autre chose en eux, qui est l'âme. Et on sait que Descartes fait même l'hypothèse de l'existence d'une glande spéciale dans le cerveau dont le but est de faire la jonction entre l'âme et le cerveau biologique. La notion de machine propre au XVII^{ème} siècle est nécessairement une notion de machine physique, dès lors on pense que s'il y a des comportements qui échappent à un pur déterminisme physique alors ils ne peuvent être que d'un autre ordre que relevant d'une machine. La pensée de Descartes sur ce sujet est le prototype même de ce que les philosophes appellent le *dualisme*. Dualisme du corps et de l'esprit, pour résumer. Ou bien pour reprendre l'expression fameuse d'un philosophe anglais : "the ghost in the machine".

Cette conception se heurte à des objections gênantes, ainsi que le fait remarquer D. Dennett qui fait référence pour les illustrer à la bande dessinée "Casper le fantôme". Dans celle-ci en effet, on voit le sympathique fantôme traverser un mur (car le monde physique n'est en rien une gêne pour lui, puisqu'il est d'une autre essence que celle dont est faite le monde physique).... pour se précipiter afin de ramasser un drap qui tombe du fil d'étendage où il était suspendu. Si le fantôme peut traverser les murs et témoigner ainsi du fait qu'il est d'une autre trempe que les objets physiques, comment peut-il saisir un objet qui tombe?

C'est cette contradiction même qui est à l'oeuvre dans la conception cartésienne. On peut bien en effet développer la théorie de deux mondes parallèles, l'un où se déroulent les événements physiques et l'autre où se déroulent les événements mentaux, mais comment va-t-on faire pour qu'ils puissent se rencontrer?

Une telle question paraît irrésoluble pour la plupart des philosophes de la conscience contemporain. Dennett en fait la raison fondamentale de ne jamais succomber aux sirènes du dualisme et de choisir le monisme et de se tenir à ce choix.

1.4. Le monisme

1.4.1. Définition

Le *monisme* est la conception qui veut que tous les phénomènes, physiques comme mentaux, aient toujours en dernière instance leur origine dans un seul monde: le monde physique. Mais on voit les problèmes qui vont immédiatement se poser. Comment un état mental,

par exemple la sensation de douleur ou bien le fait de penser à telle ou telle chose peut-il s'analyser en partant d'événements purement physiques? Comment donc *l'intentionnalité* peut-elle s'expliquer à partir des phénomènes "de la nature"? Notons que cette entreprise est qualifiée par les philosophes cités ci-dessus de "naturalisation de l'intentionnalité".

1.4.2. Réductionnisme

A priori, la solution aux problèmes posés par l'intentionnalité résiderait dans une réduction des états mentaux à des états physiques. Cette thèse du réductionnisme trouve son origine dans le béhaviorisme, ou psychologie du comportement. Les psychologues de cette tendance ont toujours soutenu en effet que, compte tenu des objectifs d'une psychologie scientifique, on ne pouvait pas faire référence à un point de vue en première personne ni à l'introspection, ni aux états mentaux, dans l'élaboration d'une théorie. Les seules données légitimes étant les comportements (puisque'ils sont observables), tout doit être traduit en termes de comportements. Évidemment, cela ne veut pas dire que les états mentaux "n'existent pas", simplement ils ne font pas partie du mobilier de la connaissance psychologique. Le béhaviorisme philosophique, lui, a voulu tenter de montrer qu'on pouvait éliminer les énoncés mentalistes en les remplaçant systématiquement par des énoncés ne comportant que la mention de dispositions comportementales. Cette tentative a échoué: il est impossible de donner des paraphrases parfaites d'énoncés portant sur la douleur ou la croyance sans d'une manière ou d'une autre réintroduire subrepticement ce qu'on voulait éliminer. Comme le dit Putnam à propos de la douleur, il est difficile "de spécifier la disposition comportementale requise autrement que comme "la disposition de X à se comporter comme si X éprouvait de la douleur"".

D'autres philosophes, tout en admettant l'existence d'états mentaux, ont prétendu qu'il était possible de les réduire à des états et processus neurologiques. Ainsi chaque état mental est-il pour eux identique à un état neurologique (thèse de *l'identité occasionnelle*), et même plus: chaque *propriété* mentale est identique à une propriété neurologique (thèse sur *l'identité de type*). En fait, comme l'ont montré Putnam et Fodor, la portée des deux thèses est très différente: "la première affirme seulement que tous les états mentaux qui se trouvent exister sont neurologiques, la seconde étend cette affirmation à tous les états mentaux qui *pourraient* exister". Alors que la thèse 1 est compatible avec la possibilité logique de porteurs d'états mentaux de nature non biologique, la thèse 2 restreint l'attribution d'états mentaux aux seuls êtres biologiques. Putnam fait aussi remarquer que si jamais on pouvait rencontrer un prédicat psychologique pouvant s'appliquer à la fois au mammifère et à la pieuvre (par exemple "avoir faim"), mais dont le corrélat physico-chimique serait différent dans les deux cas, alors la thèse 2 s'écroulerait. Or, il est très vraisemblable qu'on puisse le faire puisque, même à l'intérieur d'une même espèce, un même processus psychologique peut être mis en

correspondance avec des processus cérébraux différents (labilité du système nerveux central).

La version la plus récente du réductionnisme est celui que l'on nomme "le réductionnisme éliminativiste" (P.M. Churchland et P.S. Churchland). Il consiste à nier tout simplement l'existence d'états mentaux. Pour l'éliminativisme, il ne s'agit donc pas de constater certains phénomènes ou dispositions pour dire qu'on peut en fait les réduire à leurs composantes matérielles, il s'agit de prétendre que ces entités sont purement fictives et n'ont de réalité qu'illusoire. Nous reviendrons sur cette variante du réductionnisme au §1-5-2.

Le réductionnisme, dans le domaine du mental, rencontre donc toujours énormément de difficultés. Ce n'est pas étonnant car après tout, il en est de même dans les autres disciplines du vivant. Ainsi, on aurait pu croire à certaines époques que la biologie se résorberait finalement dans la chimie et la physique. Il n'en a rien été, du moins pour l'instant.

Il reste donc à savoir comment, à partir de quels concepts, on peut être capable de parler de l'intentionnalité et des états mentaux, en évitant le réductionnisme et en demeurant au sein de la conception moniste.

1.4.3. Notion de machine abstraite

Nous pouvons franchir un pas vers une meilleure conception de ces problèmes si nous prenons en compte les bouleversements philosophiques qui se sont produits à la suite de cette sorte de révolution de l'information qu'on a connue depuis les années trente. L'invention fondamentale est ici celle de la machine "abstraite" (ou je dirai plutôt : "à base d'information"). Elle est due à **Alan Turing**. Et on peut voir immédiatement qu'elle nous fait sortir du paradigme de pensée propre à Descartes. Car en effet désormais il existe une notion de machine qui permet de comprendre comment un pur "état informationnel" peut exercer son efficacité sur un processus physique. Si nous admettons que le monde physique inclut la notion d'information alors nous commençons à voir qu'il y a une possibilité d'intégrer des phénomènes comme les états mentaux dans un ensemble de mécanismes "naturels". Par exemple, en robotique, on utilise des langages de commandes, ce sont des langages qui permettent d'écrire des programmes dont l'effet est de déclencher certaines actions physiques: il y a bien eu transformation d'un simple calcul (passage d'un état à un autre dans une machine "informationnelle") en action physique, comme s'il s'agissait du passage d'un état mental à une action. A condition évidemment d'identifier un état mental à un état de calcul...

La théorie des machines abstraites (qui a donné naissance à l'informatique) a donc renouvelé considérablement la manière de poser le problème des rapports de l'esprit et du corps, ou, a l'inverse, celui de savoir si un dispositif physique (mécanique) autre qu'humain peut avoir un esprit, question qu'on peut résumer en: " un ordinateur peut-il avoir un esprit?".

1.4.4. L'"Intelligence Artificielle"

La question de savoir si on peut considérer qu'un dispositif mécanique pense n'est pas nouvelle. Turing avait tenté en son temps de montrer au moyen d'un test qu'elle était loin d'être absurde. Si en effet nous pensons que la conscience est d'abord une propriété qui s'éprouve en première personne, alors nous sommes condamnés à ne l'attribuer que sur la base de tests expérimentaux. On pourrait imaginer la même chose avec la vie: à quoi reconnaît-on que quelque chose vit? On peut imaginer de faire passer des tests à cette chose: si elle répond positivement alors on peut admettre qu'il y a vie. idem pour "l'intelligence".

Le **test de Turing** est censé permettre de dire si une machine pense. Il repose sur le dispositif suivant : on confronte deux candidats: l'ordinateur et un sujet humain. Il y a un interrogateur qui pose des questions aux deux, sans évidemment voir son interlocuteur. Le candidat humain répond aux questions en toute bonne foi et essaie de persuader l'interrogateur qu'il est bien un être humain. De l'autre côté, l'ordinateur est programmé pour mentir: il cherche à convaincre l'interrogateur que c'est lui l'être humain. Si à la suite du test, l'interrogateur ne peut pas identifier le candidat humain de façon cohérente, alors l'ordinateur a gagné la partie. Ceci illustre évidemment un point de vue opérationnaliste, point de vue partagé par les partisans de l'IA forte c'est-à-dire l'idée qu'un programme informatique suffisamment perfectionné peut être déclaré intelligent, simplement parce qu'il répond de manière semblable à ce que ferait un être supposé intelligent : peu importe le moyen d'arriver à la solution. (Dennett définit ainsi l'opérationnalisme : c'est la thèse ou la démarche qui s'exprime par la déclaration : "si vous ne pouvez découvrir une différence, il n'y a pas de différence". Ou, comme on dit souvent : "si cela couaque comme un canard, et marche comme un canard, c'est un canard".) Selon cette perspective en quelque sorte continuiste (penser à Montaigne), il serait possible d'attribuer une qualité mentale au fonctionnement de n'importe quel dispositif capable de calculer (y compris un simple thermostat par exemple): tout ce qui peut s'exprimer sous la forme d'un algorithme contient des propriétés mentales. Remarquons que cela n'entraîne pas que toute activité mentale se ramènerait finalement à exécuter une suite d'opérations, ou algorithme. Il pourrait y avoir certaines activités mentales s'exprimant comme des algorithmes, et d'autres... mystère! Les tenants de ce genre de thèse préfèrent cependant aller à la solution radicale qui consiste à identifier purement et simplement activités mentales et algorithmes. Avec cependant une différence dans le cas d'un auteur comme R. Penrose. Selon ce dernier en effet, il pourrait bien y avoir des activités mentales non basées sur la notion d'algorithmes, mais au sens que nous donnons aujourd'hui à cette dernière notion, c'est-à-dire justement au sens d'un processus analogue à celui qui s'exécute au moyen d'une machine de Turing. Il y aurait, selon Penrose, des moyens d'arriver à des états mentaux autres que ceux actuellement connus. Pour l'illustrer, il prétend qu'on pourrait imaginer dans le futur des espèces de machines travaillant sur des bases

complètement différentes de nos machines actuelles. Ces machines utiliseraient certaines propriétés de la matière au niveau quantique. Selon Penrose, en effet, la clé du mystère réside dans des propriétés de la matière non encore découvertes. C'est une thèse assez isolée dans le monde des sciences cognitives, Penrose n'a pas fait la démonstration convaincante de l'impossibilité d'expliquer la conscience autrement qu'en faisant référence à ces propriétés physiques inconnues.

1.4.5. La "chambre chinoise"

Ce point de vue de l'IA forte a été vivement contesté par **J. Searle**, selon qui même si le test de Turing est passé avec succès, cela ne veut pas dire que l'ordinateur soit "intelligent", dans la mesure où cela n'implique pas qu'il soit doté de la compréhension de ce qu'il fait.

Pour répondre à ce test, Searle invente une autre situation: celle dite "de la **chambre chinoise**".

On imagine un prisonnier enfermé dans une chambre, qui ne connaît absolument pas le chinois et ne peut communiquer avec l'extérieur que par l'échange de petits bouts de papier. On a remis au prisonnier une série d'instructions pour utiliser des mots en chinois qui lui sont communiqués. Ces instructions sont en fait comme celles d'un programme informatique : on ne fait jamais ici que simuler le fonctionnement d'un ordinateur, sauf que c'est un sujet humain qui sert d'unité de calcul. Lorsque tous les mots en chinois ont été communiqué au prisonnier et que celui-ci a réagi selon le mode d'emploi, il se trouve qu'une certaine tâche a été effectuée correspondant à ce que les gens de l'extérieur attendaient. La question est : "le prisonnier connaît-il le chinois?". En adoptant un point de vue opérationnaliste, la réponse est "oui". Mais Searle conteste fortement ce point de vue. Selon lui, il n'y a pas eu compréhension de la part du prisonnier, la preuve en est que si on l'interrogeait, il nous confirmerait qu'il ne connaît pas le chinois, et pourtant, objectivement, vu de loin, il s'est comporté comme s'il le connaissait. Selon J. Proust, cet argument de Searle met clairement en évidence la distinction entre **capacité de calcul** et **capacité "sémantique"** ou **"intentionnelle"**. Et il met en question l'identification de l'esprit ou de l'intelligence avec un ensemble de calculs effectués par un algorithme. La machine de Turing est dans la même position que le prisonnier dans sa chambre chinoise: même si elle est écrite pour traiter les données qui lui sont soumises et si elle parvient à résoudre un problème, elle ne le fait que par sa capacité de calcul et elle ne manifeste aucune capacité sémantique ou intentionnelle. Il faut toutefois noter que le problème se complique dans un raffinement qui est donné à cet exemple de la chambre chinoise: les tâches à effectuer étant tellement complexes, il est déraisonnable de laisser un seul individu en charge de les accomplir. Vient alors l'idée de le remplacer par une équipe dont les membres se coordonnent. Cette équipe peut être très importante numériquement: pourquoi ne pas envisager qu'un continent entier travaille. Alors en ce cas, c'est toujours la même chose: les tâches peuvent être effectuées toujours sans compréhension de

la part des multiples agents. Mais cependant, la comparaison qui vient à l'esprit n'est pas celle de la machine avec un cerveau humain, mais celle d'une multitude de processeurs avec un ensemble de neurones. Le fait que chaque individu n'ait pas la compréhension du chinois a beaucoup moins d'impact du point de vue argumentatif car on ne s'attend pas, de toutes façons, à ce que chaque neurone ait pour lui-même une compréhension de la tâche à effectuer (ou d'une histoire à raconter etc.). Si compréhension il y a, on s'attend alors à ce qu'elle émerge à partir du travail coopératif de l'ensemble des neurones / machines. Searle a beau répondre qu'il n'y a aucun sens à dire qu'un continent "comprenne" une histoire sans que chacun de ses membres ne la comprenne, sa réponse ne colle plus tout à fait à la nouvelle situation si on imagine que la compréhension peut advenir chez un sujet sans qu'aucune région de son cerveau voire aucun neurone ne soit dépositaire d'une telle compréhension. Cette attitude est celle qu'ont les tenants du connexionnisme et de la théorie de l'émergence, dont nous parlerons également plus loin. En attendant, l'argument de Searle est pertinent tant que nous nous limitons au travail d'un seul sujet enfermé dans sa chambre.

1.5. Le fonctionnalisme

1.5.1. Machine Universelle

Cette conception de l'IA forte ne va pas sans une autre conception: le *fonctionnalisme*. Cette conception est elle aussi dérivée des travaux de Turing. En effet, comme nous le verrons plus en détails dans le cours consacré à l'examen du concept de machine de Turing, il est possible de concevoir une machine de Turing universelle TU, c'est-à-dire une machine capable de simuler le travail de n'importe quelle machine (donc aussi en particulier d'elle-même) sur n'importe quelle donnée d'entrée. Cette découverte a été capitale car elle est à la source de notre notion d'ordinateur. En effet un ordinateur est une machine capable de simuler le travail de n'importe quelle machine particulière (imaginons une machine entièrement câblée dévolue à une seule tâche), il obtient cela en stockant dans sa mémoire le programme qui représente symboliquement le fonctionnement de chaque machine particulière. On peut montrer que l'architecture standard des ordinateurs (dite architecture de von Neumann) est équivalente à la machine de Turing universelle au sens suivant : tout programme accompli par une telle architecture peut être converti en un programme accompli par TU et réciproquement. Il y a à cela au moins deux conséquences : d'une part, la programmation d'une machine particulière peut aussi bien s'effectuer sur un substrat physique (câblage) qu'au moyen d'une suite de symboles (inscrits sur le ruban de la machine universelle), et d'autre part le concept abstrait de machine de Turing universelle sert à rendre compte de toutes les espèces d'ordinateurs qu'on a pu construire à ce jour, il exprime "sur le papier" les propriétés communes à toutes les machines capables d'effectuer un algorithme. Autrement dit, l'essence du calcul n'est pas dans quelque

propriété d'un support physique mais peut s'exprimer abstraitement, hors de toute réalisation physique concrète. On dit aussi parfois tout simplement qu'on pourrait bien réaliser concrètement une machine telle que TU au moyen de conduites d'eau et de vannes, de petits bouts de papier qu'on se passe entre amis ou de circuits électroniques, ce serait toujours la même machine de Turing et ce serait toujours les mêmes calculs qui seraient effectués. On sera donc tenté de chercher le "moi" ou la conscience dans ce genre de propriétés "virtuelles" plutôt que dans des propriétés physico-chimiques, comme si le moi pouvait être indépendant de son substrat physique (de son corps par exemple). A partir de cette conception se sont développées beaucoup de chimères et de récits relevant de la science-fiction. Ainsi l'exemple du téléportage: on imagine qu'il est possible de scanner entièrement un individu et de ramasser ainsi toute l'information qu'il contient (concernant sa composition atomique) et d'expédier cette quantité d'information par un rayon laser sur une autre planète. Qu'advient-il de l'original? Y a-t-il désormais deux individus conscients ou un seul? Si l'information est stockée sur un disque avant d'être convertie dans un corps ayant forme humaine, l'information stockée contient-elle encore le moi de l'individu? Bref, peut-on ainsi raisonnablement identifier le moi individuel à une masse d'information digitalisable (numérisable), en sorte que le support de cette information numérique soit complètement indifférent? Aime-t-on toujours son (sa) petit(e) ami(e) quand il (elle) est transformé(e) en gravure d'un CD Rom?

1.5.2. Fonctionnalisme vs réductionnisme

Cette conception de l'IA forte paraît extrémiste. En fait, on peut avoir une position fonctionnaliste sans souscrire à cette conception. C'est le cas de **P. Jacob** (cf. "Le problème des rapports du corps et de l'esprit aujourd'hui" in Andler, 1992), autrement dit sans penser qu'il puisse y avoir un "moi" indépendant de tout substrat. Le fonctionnalisme est alors en tant que philosophie, beaucoup plus élaboré que ce qui peut paraître ici. Il postule un lien permanent entre le "substrat" et la "fonction", celui-ci se traduisant en termes d'état mental et celui-la en termes d'état cérébral. Parler en termes d'états cérébraux et parler en termes d'états mentaux seraient ainsi deux conceptions complémentaires. On peut faire de nombreuses analogies : il est ainsi possible de donner une description d'un organe tel que le coeur à partir de son fonctionnement physico-biologique aussi bien qu'en le caractérisant à partir de sa fonction, qui est de pomper le sang. La propriété de "bien couper" qui peut être attribuée à un couteau n'est pas seulement une manière de parler, elle désigne une propriété fonctionnelle importante, même si elle est liée à un certain état physique de la matière qui compose la lame du couteau. la propriété "bien couper" est donc reliée à un état matériel. *L'idée de base du fonctionnalisme est alors qu'on peut décrire un système cognitif également à partir de ses fonctions et de ses propriétés fonctionnelles.* Mais cela ne confère pas nécessairement à ces dernières une existence autonome. Sans lame de couteau appropriée, impossible de couper

correctement son steak! D'autre part, on peut se demander quel est le statut de ces propriétés fonctionnelles dont on parle: "existent"-elles vraiment ou ne sont-elles que des manières de parler?

La discussion rejoint ici celle qui a lieu entre les tenants de la "psychologie ordinaire" (trad. franç. de "folk psychology") et les tenants de l'"éliminativisme" (ou: réductionnisme éliminativiste). La psychologie ordinaire est l'ensemble des savoirs courants dont nous disposons lorsque nous pensons et discourons à propos de nos attitudes ou de celles d'autrui. Par exemple, nous disons: "il mange parce qu'il a faim", "je prends un aspirine parce que j'ai mal à la tête et que je crois qu'un aspirine va me soulager", "il a pris le train parce qu'il veut rejoindre au plus vite son amie" et ainsi de suite. Autrement dit, nous présentons nos actions et celles des autres comme ayant pour cause des attitudes propositionnelles (croyances, volontés, désirs, peurs etc.). Mais ce faisant ne sommes-nous pas en train de nous rendre sans arrêt coupables de dualisme (au sens évoque plus haut du "fantôme dans la machine"?). En effet, comment un "désir", ou une "croyance", éléments intangibles s'il en est, peuvent-ils avoir un rôle causal sur nos comportements qui, eux, se traduisent bel et bien en actions physiques? Les matérialistes éliminativistes (P.M. et P.S. Churchland en particulier, mais c'est une thèse qui fut également défendue par R. Rorty et par Feyerabend) prétendent donc que tout cela n'est que "manière de parler" ou "illusion", et qu'en fait, il n'y a en dernier ressort que des processus physico-chimiques, qui sont les vrais causes de nos actions. Ils prétendent ainsi que toutes ces propositions sont des manières désuètes de s'exprimer, un peu comme quelqu'un qui continuerait à parler d'"éther" ou de "fluide phlogistique" alors que la science a depuis longtemps condamné ces concepts comme reposant sur des illusions...

Les partisans de la psychologie ordinaire au contraire soutiennent qu'on ne peut se passer de ces manières de s'exprimer parce qu'elles expriment bel et bien une réalité, même si ce n'est pas exactement la même réalité que celle du monde physique. Après tout, nous sommes bien obligés d'utiliser la notion de nombre alors que pourtant elle ne s'identifie à rien de physique. Et nous n'avons aucune idée de la manière dont nous pourrions bien nous passer de cette notion de nombre en la remplaçant par la mention de quelque chose de physique (ou alors ce serait une drôle de régression dans la culture humaine, qui consisterait à évoquer un tas de cailloux chaque fois que nous faisons référence à un nombre entier!). Comme le dit P. Jacob : "de ce que la question suivante est absurde : "Quelle est la composition physico-chimique d'une croyance?", le matérialisme éliminatif conclut à l'inexistence pure et simple des croyances". Il y a pourtant des quantités d'entités abstraites auxquelles nous sommes bien obligés de conférer l'existence, par exemple (l'exemple est de Dennett), nous sommes bien obligés de conférer une existence à la notion de centre de gravité d'un objet, même s'il n'y a aucun sens à demander de quelles atomes est constitué un centre de gravité...

Ce n'est donc pas parce qu'une entité n'est pas identifiable à quelque chose de physico-chimique qu'elle n'existe pas. S'il n'existe ni croyance ni désir, alors il faut trouver d'autres arguments pour le prouver!

1.5.3. L'explication en termes fonctionnalistes

P. Jacob développe l'exemple du comprimé d'aspirine ou de doliprane que l'on prend contre le mal de tête. Nous prenons un tel comprimé parce que c'est un antalgique, c'est-à-dire un produit qui a pour fonction de combattre la douleur, autrement dit nous faisons référence à cette propriété pour expliquer que nous le prenons. Ici, deux observations : premièrement, pouvons-nous totalement éliminer toute référence à une attitude propositionnelle lorsque nous expliquons notre comportement? La tendance dite "du béhaviorisme logique" (Ryle, Wittgenstein) avait le projet de traduire tous les énoncés contenant une telle référence en énoncés ne faisant référence qu'à des observables: ce projet n'a jamais abouti, il y a toujours un résidu d'attitude propositionnelle qui reste quelque part. Deuxièmement, lorsque nous disons que la propriété pour l'aspirine ou le doliprane d'être des antalgiques est la cause qu'ils nous permettent de ne plus souffrir de mal de tête, nous sommes encore dans le cas d'"expliquer" un phénomène physique par une propriété fonctionnelle en quelque sorte immatérielle. Pouvons-nous éliminer ce genre d'"explication"? Si nous pouvons l'éliminer dans une certaine mesure, ne pouvons-nous pas néanmoins continuer de l'utiliser de manière utile, et si oui, dans quel sens? Bien sûr, la propriété d'être un antalgique est liée à une propriété physique matérielle: la présence dans ces produits d'une molécule spécifique, qui est soit l'acide acétylsalicylique (pour l'aspirine), soit le paracétamol (pour le doliprane) en sorte que de fait, et physiquement, c'est ce type de molécule qui est la cause de la disparition de ma migraine. Dans ce cas, nous avons identifié un substrat physique responsable d'une action sur un organisme. N'y a-t-il pas néanmoins un sens à toujours dire que la propriété d'antalgique (ou analgésique) explique l'efficacité du médicament? P. Jacob fait ici référence aux efforts de Carnap pour éliminer du langage de la science les entités non-directement observables (à une certaine époque, c'était le cas des électrons). Il s'en sortait de la manière suivante. Soit une phrase comme : "un électron possède une charge électrique négative", remplacer "électron" et "charge électrique" chacun par une variable, puis quantifier existentiellement. On obtient quelque chose comme : $\exists x, \exists C, \exists \alpha, C(x, \alpha) \wedge (\alpha < 0)$. Cet énoncé est une formule de logique du second ordre, en effet C désigne un prédicat du premier ordre (qui s'applique aux "individus" x et α , x étant un individu de cette théorie et α un nombre), c'est donc une variable du second ordre et l'énoncé comporte une quantification sur C. On peut paraphraser cet énoncé en : "il existe une entité individuelle non nommée, et une propriété physique non nommée telles que l'entité individuelle possède la propriété en question". Et on parle en ce cas d'énoncé du second ordre. Mais il apparaît maintenant que la propriété d'être un antalgique est une propriété du second ordre, elle peut

s'exprimer en effet comme une disjonction de propriétés physiques du premier ordre ("être composé de molécules d'acide acétylsalicylique ou de molécules de paracétamol") (or, comme nous le savons, une disjonction est un cas particulier d'énoncé existentiel). Au terme de cette analyse, on voit que la propriété fonctionnelle a un sens: elle est analysable si on souhaite l'analyser, en fonction des propriétés du premier ordre qui lui sont sous-jacentes. Faire référence à une propriété fonctionnelle est donc certes une manière de parler, mais c'est une manière de parler qui est fondée logiquement. Mais ce n'est pas tout, car cela ne justifie pas qu'on continue de s'en servir (si on a à notre disposition une "meilleure" explication en termes de propriétés physiques du premier ordre). P. Jacob identifie deux raisons d'attacher encore de l'intérêt aux propriétés fonctionnelles. D'une part, il est faux de prétendre que les propriétés fonctionnelles n'exercent en elles-mêmes aucun rôle causal: il suffit de mentionner l'effet "placebo". En ce cas, le contenu de la croyance qu'un médicament est un antalgique a de l'effet sur la douleur alors même qu'il n'y a pas la présence des molécules citées ci-dessus. D'autre part, dire que "la douleur a été supprimée par l'ingestion d'un antalgique" a la fonction de donner une explication sans nommer explicitement la molécule responsable, et c'est une "explication" en ce sens qu'elle en écarte d'autres (la douleur aurait pu disparaître parce que le malade est guéri ou bien parce qu'il est mort...). On peut ainsi distinguer le fait de *nommer* une cause (utilisation d'une propriété de premier ordre) et le fait de fournir une explication causale en affirmant qu'il existe une cause.

Cette analyse s'applique bien aux problèmes que nous nous posons concernant le cerveau et l'esprit. La psychologie cognitive, la linguistique, l'ergonomie cognitive etc. étudient les fonctions cognitives, les neurosciences étudient la manière dont elles sont implémentées. Le rapport est évidemment le même que celui qu'on établit depuis longtemps entre software et hardware d'une machine. Les livres d'informatique usuels (Aho et Uhlmann...) tout autant que les philosophes cognitivistes comme Marr font une distinction entre au moins trois niveaux lorsque nous essayons de résoudre un problème informatiquement. Il y a un premier niveau où on modélise le problème grâce à des outils mathématiques et logiques, un second niveau où à partir de cette spécification, on écrit un algorithme (en utilisant une notation, certaines structures de contrôle etc.) et enfin un troisième niveau qui est celui du mécanisme physique qui implémente l'algorithme. De nos jours, le passage du deuxième au troisième s'effectue automatiquement: le programme découlant de l'algorithme écrit dans un langage quelconque (Pascal, LISP, Prolog ou C++) étant immédiatement compilé ou interprété c'est-à-dire traduit en instructions machines (exprimées dans le langage propre de la machine). Et c'est une perspective de l'informatique future (encore qu'on y soit presque) que de faire en sorte que le passage du premier au deuxième niveau soit tout aussi automatique (c'est le sens par exemple des recherches en informatique théorique aujourd'hui qui

utilisent la théorie de la preuve en logique pour produire automatiquement des programmes qui ont de plus l'avantage d'être corrects car ils sont déjà prouvés au moment même où on les obtient!).

Le fonctionnalisme dans les sciences cognitives établit un parallèle avec cette démarche en identifiant le mécanisme physique au fonctionnement cérébral et les deux premiers niveaux aux états mentaux.

On peut donc résumer le fonctionnalisme en disant qu'états mentaux et états cérébraux sont distincts mais néanmoins toujours liés (tout état mental s'accompagne d'un état cérébral de la même façon que tout état de calcul dans un algorithme s'accompagne, lorsqu'il est implémenté, d'un état physique de la machine).

Évidemment demeure un point mystérieux dans le fonctionnalisme: si la métaphore de l'ordinateur est maintenue, qu'est-ce qui, dans le cerveau humain, tient le rôle du langage-machine? Certains philosophes, dont Pierre Jacob, font l'hypothèse d'un langage-machine propre à l'homme, qu'ils appellent le "mentalais". Cette hypothèse est hautement spéculative et personne n'a pu encore étudier le mentalais...!

1.5.4. Au-delà du fonctionnalisme turingien

Jusqu'à quel point doit-on prendre au sérieux l'analogie entre esprits humains et machines de Turing? Putnam, au départ, (dans les années soixante) défendait la thèse selon laquelle les machines de Turing fournissaient un bon modèle de l'esprit humain. Cette thèse a été plus tard critiquée par Block et Fodor et rejetée par Putnam lui-même. L'identification des états psychologiques et des états d'une machine de Turing pose de nombreux problèmes. On peut en particulier signaler le caractère fixe, non évolutif d'une machine de Turing: l'ensemble de ses états est donné une fois pour toutes. Les seules nouveautés peuvent provenir d'inscriptions nouvelles sur son ruban. On ne peut donc pas imaginer un processus d'enrichissement de ses états. D'autre part, on peut être amené à simplifier une machine de Turing si on constate que deux états sont tels que leurs entrées et leurs sorties sont exactement identiques. Cela veut dire qu'un état est totalement identifié à l'ensemble de ses entrées et sorties, or il n'est pas du tout évident que nos états psychologiques puissent être identifiés de la sorte. Deux états psychologiques peuvent se ramener à un seul tout en ayant des sorties différentes (la douleur par exemple chez un individu donné, qui peut réagir à celle-ci de différentes manières), et au contraire on peut imaginer des états psychologiques distincts conduisant aux mêmes comportements.

Il ne faut donc peut-être pas prendre le modèle des machines de Turing "à la lettre". Certes, elles font la preuve de l'existence de dispositifs très puissants pouvant simuler des opérations intellectuelles, dispositifs qui peuvent être définis indépendamment d'une réalisation physique particulière, mais la manière dont une solution à ce problème est choisie dans les machines de Turing doit demeurer secondaire: il y a peut-être des modèles plus pertinents des tâches accomplies par l'esprit humain.

D'autre part, nous voyons qu'une limite grave au fonctionnalisme est atteinte dès que nous réfléchissons au parallèle entre états mentaux et états computationnels. En effet, ces derniers ne sont que les éléments d'un ensemble fini: ils n'ont d'identité, comme dit plus haut, que celle que leur procurent leurs entrées et leurs sorties dans le diagramme sagittal que l'on peut faire pour représenter la machine. Leur nom n'importe pas. Ils n'ont pas de "contenu" intrinsèque. Or, notre psychologie ordinaire nous demande de distinguer les désirs des croyances, les douleurs des plaisirs, les envies des frayeurs... Le fonctionnalisme "turingien" est capable de nous dire comment un état peut avoir un rôle causal, il n'est pas capable de nous dire que c'est en vertu d'un certain contenu qu'il le peut.

Résoudre cette question du contenu, c'est franchir une étape supplémentaire, c'est expliquer comment les propriétés sémantiques des symboles émergent, c'est comme le dit Pierre Jacob, expliquer "pourquoi les choses ont un sens". On doit alors faire référence à la théorie de **F. Dretske**, selon laquelle les propriétés sémantiques peuvent s'expliquer à partir de la notion d'*information*. Pour ce philosophe en effet, l'"information" est une catégorie ontologique, c'est-à-dire: les éléments d'informations (que certains appellent aussi *infons*) font partie de l'existant, au même titre que les particules atomiques, et c'est à partir d'eux que se forment les représentations sémantiques (le "sens" des choses), relais pour passer ensuite à l'intentionnalité. Il existe aussi une certaine parenté à ce sujet avec les thèses de D. Dennett qui seront évoquées au § suivant.

1.6. Une thèse évolutionniste: celle de D. Dennett

D. Dennett a écrit en 1991 un gros livre traduit en français en 1993 sous le titre "La conscience expliquée". Titre très ambitieux et qui laisse entendre que tous les mystères seraient enfin levés. De fait, il n'en est rien, ou au prix d'autres mystères, plus profonds encore! Il part pour exposer sa thèse d'un point de vue évolutif, et même évolutionniste au sens darwinien. "Au début, il n'y avait pas de raisons; il n'y avait que des causes. Rien n'avait un but, et rien n'avait quoi que ce soit qui puisse ressembler à une fonction; il n'y avait pas du tout de téléologie dans le monde. L'explication en est simple : il n'y avait pas d'êtres doués d'intérêts. Mais après des millénaires apparurent des *réplicateurs* simples." Les réplicateurs simples en question sont des êtres qui n'ont pas d'autre intérêt que celui de se reproduire. Ces êtres ont donc dû tout de suite "apprendre" à éviter ce qui les mettait en danger et au contraire à rechercher ce qui était bon pour eux. En opérant ce genre de division, est apparue une certaine diversification des intérêts de ces êtres à l'égard du monde, d'où un certain nombre de raisons à leurs comportements. Mais de plus, lorsque les créatures en question ont voulu rechercher ce qu'il y avait de meilleur pour elles-mêmes et rejeter ce qu'il y avait de moins bon, il a fallu qu'elles s'identifient elles-mêmes comme étant un "moi" (il faut savoir ce que l'on doit préserver avant de trouver les moyens de le

préserver!). Un exemple fameux de problème de reconnaissance d'un soi est fourni par le système immunitaire (qui est souvent présenté comme un "autre" système cognitif). Ceci dit, l'exemple de ce système est instructif à plus d'un titre, notamment en ce qu'il nous prouve qu'il n'est nul besoin d'avoir un "chef" pour que le système agisse et réagisse (dans le cas de la conscience: nul besoin donc d'un point "ultime" qui serait en dernière instance celui qui commanderait les fonctions cognitives). De plus, il peut évidemment se passer une foule d'accidents dans la reproduction, de même que la recherche à tout prix de solutions pour une meilleure survie peut très bien à un moment donné se traduire par l'apparition d'effets secondaires inattendus et même indésirables. On peut aussi avoir des divergences et des systèmes fonctionnels émergents imprévus mais qui, en interagissant, peuvent créer de nouvelles possibilités et ainsi de suite.

Dans une première phase de cette évolution, on a des organismes qui sont "câblés": les réactions au milieu viennent de circuits non modifiables, et puis dans une deuxième phase apparaît la plasticité, c'est-à-dire "l'émergence de phénotypes individuels dont l'intérieur n'est pas entièrement câblé de façon innée mais plutôt variable ou plastique et qui peuvent ainsi apprendre durant leur vie". Cette plasticité joue un grand rôle car elle permet de mieux répondre à des processus chaotiques, une façon de faire étant, lorsqu'apparaissent certaines circonstances, de se réorganiser c'est-à-dire finalement apprendre. De plus, non seulement elle donne cet avantage mais en plus elle peut accélérer le processus évolutif. Il y a ainsi une grande variété de câblages possibles à la naissance des individus. On peut imaginer qu'il y ait un individu ayant un câblage meilleur que d'autres, lui conférant un avantage déterminant (Dennett appelle ça un "Bon Truc"). Évidemment s'il n'y a pas de plasticité du cerveau, on risque d'en rester là: les rares individus ainsi câblés ne vont pas créer une descendance nécessairement nombreuses. Mais s'il y a plasticité, alors les voisins de cet individu vont chercher à l'imiter, puis finalement obtenir une organisation semblable. Ces individus, devenus plus nombreux et ayant plus d'avantages (donc une meilleure survie) que ceux qui n'atteignent pas un tel résultat vont eux-mêmes se reproduire en plus grand nombre et leurs descendants vont à leur tour, même s'ils n'ont pas le bon câblage de façon innée, tenter d'obtenir le "Bon Truc" et ainsi de suite. De cette manière, l'évolution va progresser plus vite encore.

Parmi ces "Bons Trucs", Dennett en suggère certains. Par exemple "on" aurait pu remarquer qu'en s'adressant à autrui alors qu'il n'y a personne, la voix revient au locuteur et lui procure ce genre de stimulation ou d'encouragement que l'interpellation par autrui aurait pu lui causer. Il apparaît alors que ce comportement engendre une communication de soi avec soi qui supplée les manques de certaines connexions internes, d'où l'habitude de se parler à soi-même à haute voix, puis de manière silencieuse, ce qui serait l'origine de la conscience. ("Le processus silencieux maintiendrait la boucle de l'autostimulation "). Du reste, il est facile de mettre en évidence la facilité avec laquelle les êtres humains sont

capables de créer de nouvelles voies de communication internes lorsque certaines sont insuffisantes ou bien viennent à être détruites par des lésions cérébrales.

Reste à expliquer comment se transmettent les "Bons Trucs". L'autre processus évolutif est, pour Dennett, l'évolution culturelle. Dans le domaine biologique, le gène est l'entité de réplication. Une hypothèse est alors faite (au départ par le zoologue R. Dawkins) selon laquelle nous aurions atteint un nouveau stade de l'évolution au cours duquel un autre type de réplicateur serait en train de prendre le dessus. Dawkins, dans *Le gène égoïste*, appelle ces entités (sur le modèle des gènes): les mèmes. En gros, un *mème* est une idée élémentaire, c'est (Dawkins): " une unité de transmission culturelle, ou une unité d'imitation... Des exemples sont : des airs musicaux, des idées, des expressions toutes faites, des modes vestimentaires, des façons de faire des poteries ou de construire des voûtes. Tout comme les gènes se propagent dans le pool génétique en sautant d'un corps à un autre par l'intermédiaire du sperme et des oeufs, les mèmes se propagent dans le pool mémétique en sautant de cerveau en cerveau à travers un processus qui, au sens large, peut être appelé imitation."

Finalement, trois véhicules essentiels ont contribué à l'organisation de la conscience humaine : l'évolution génétique, la plasticité phénotypique, et l'évolution mémétique. Le cerveau devient alors une sorte de machine pour gérer le flux de mèmes et si nous voulons tenter de décrire la manière dont il fonctionne alors, le seul genre de dispositif dont nous disposons est, là encore, emprunté à l'informatique. Et nous retombons ainsi sur le rôle (envahissant!) du modèle proposé par Turing.

Dennett résume sa thèse de la façon suivante :

"La conscience humaine est *elle-même* un énorme complexe de mèmes (ou plus exactement d'effets de mèmes dans le cerveau); elle fonctionne en quelque sorte comme une machine virtuelle à la von Neumann *implémentée* dans l'*architecture parallèle* d'un cerveau, lequel n'était pas conçu pour de telles activités. Les pouvoirs de cette *machine virtuelle* augmentent énormément les pouvoirs sous-jacents du *matériel* organique sur lequel elle tourne."

Finalement, Dennett nous donne une utilisation spectaculaire du fonctionnalisme en fondant le fait qu'on puisse décrire les fonctions cognitives au moyen de propriétés fonctionnelles sur l'hypothèse que celles-ci existent réellement en tant que portées par certaines entités non physiques appelées "mèmes". Ce faisant, il ne s'éloigne pas tellement de la position exprimée par Jacob dans la mesure où il admet toujours qu'en dernier ressort, les états mentaux se traduisent par des inscriptions physiques, mais sa justification du niveau fonctionnel de la description est beaucoup plus forte.

L'hypothèse des *mèmes* nous paraît pourtant tout aussi hasardeuse que celle du *mentalais*.... De plus, toute explication en termes d'évolution pose inévitablement la question de "pourquoi l'évolution" et pourquoi des mutations dans l'évolution qui aboutiraient notamment à la conscience.

L'explication par "le hasard" (favorisée en son temps par Jacques Monod) est assez faible. Si le "hasard" fait aussi bien les choses en les forçant à évoluer vers un point de convergence comme la conscience, pourquoi alors ne pas supposer que c'est le point de convergence qui est au départ de tout, renversant ainsi la perspective pour mettre la conscience à l'origine du monde. Cela paraît à première vue farfelue, mais c'est pourtant la voie choisie par certains physiciens qui défendent ce qu'on appelle la *théorie anthropique* ¹!

1.7. Le connexionnisme

Il y a aussi une autre position, qui est non-fonctionnaliste et se base sur les propriétés d'*émergence* que peuvent avoir des réseaux extrêmement complexes, comme les *réseaux de neurones*, c'est la voie suivie notamment par **Francisco Varéla**. Cette voie n'est pas sans évoquer les difficultés inhérentes au dualisme: on fait reposer sur un concept mystérieux ("émergence") le soin d'établir le lien entre la matière neuronale et la conscience. Nous verrons toutefois qu'il y a des arguments en faveur de cette autre voie et qu'il y a peut-être une façon de sortir du cercle monisme-dualisme consistant à mettre en doute l'importance du concept de "moi" et, partant, de conscience au sens classique (occidental) du terme.

1.7.1. Critique du cognitivisme

Ce que Varéla nomme *cognitivisme*, et que d'autres nomment *computationnalisme*, ou bien simplement "approche symbolique" de la cognition, est la conception dont nous nous sommes inspirés dans les paragraphes précédents, qui fait de l'esprit humain un système de traitement de l'information. Il se base sur quelques idées de fond:

- a) que le cerveau humain traite des informations venant de l'extérieur par le canal des sens et qu'il transforme en représentations internes,
- b) que ces représentations internes sont exprimées dans une sorte de langage (le mentalais),
- c) que le fonctionnement de l'esprit humain sans être exactement celui d'une machine de Turing y ressemble dans ses aspects majeurs, que sont: l'aspect symbolique (syntaxique) des traitements et leur caractère séquentiel.

On peut ajouter que, selon cette approche, toute information est "digitalisée": si en effet, le cerveau humain traite les informations quand elles sont traduites en symboles et si ceux-ci sont traités d'après leur forme physique, alors cela semble impliquer que leur place et leur rôle sont les mêmes que dans un ordinateur, autrement dit qu'ils sont analysés comme des suites d'unités d'information.

¹ On retrouve aussi cette idée dans plusieurs courants de la philosophie indienne (çankara) et du bouddhisme, ainsi que dans les écrits philosophiques de Shrödinger.

Or, de nombreux arguments existent qui tendent à nous montrer qu'il n'en est pas ainsi dans la réalité. Les biologistes ont par exemple acquis beaucoup d'éléments sur la vision et sur l'oeil. Le "dogme" computationnaliste voudrait que l'oeil soit comme le caméscope qui filme des images et les convertit en information numérique. Le nerf optique relie les yeux à une région du thalamus appelée le corps genouillé latéral (GNL), puis au cortex visuel. Varéla (1989) écrit : "L'explication typique en termes de traitement de l'information est que l'information pénètre par les yeux pour être séquentiellement transmise, via le thalamus, au cortex, où un "traitement plus approfondi" est effectué. Mais cette idée de fonctionnement séquentiel résiste mal à une analyse un peu fine. [En fait], il apparaît de façon évidente que 80% de tout ce qu'une cellule du CGL voit passer vient du dense réseau qui la relie aux autres régions du cerveau plutôt que de la rétine. De plus, on voit bien qu'il y a plus de fibres allant du cortex vers le CGL que dans le sens inverse. La description séquentielle du réseau de la vision semble complètement arbitraire: la séquence pourrait tout aussi bien être orientée dans l'autre sens". Ainsi finalement, dans une activité cérébrale paraissant aussi évidente que la vision, les choses se passent un peu comme si les activités internes au cerveau prenaient plus de place que le rapport avec le monde extérieur. Que sont ces activités internes? Là encore, l'étude neurophysiologique a beaucoup progressé et a su mettre en évidence des activités neuronales sous forme de mises en résonance de zones cellulaires. Les activités conscientes ressemblent alors à des obtentions d'états globaux dans des systèmes complexes mettant en jeu des milliards de neurones. De l'ensemble de ces observations, naît l'idée que ce que nous appelons "état mental" se caractérise par des états d'équilibre atteints dans l'activation de réseaux de neurones. Et loin de pouvoir identifier précisément tel ou tel neurone, voire même tel ou tel ensemble de neurones comme responsable de telle ou telle activité de pensée, nous sommes conduits à n'observer que des états globaux d'excitation se réalisant dans différentes zones du cerveau.

Ainsi la nature symbolique des activités mentales est fortement contestée. D'abord, son aspect séquentiel est critiqué: une machine séquentielle a de nombreuses limites. On sait ainsi que certains problèmes (par exemple l'analyse d'une phrase, la reconnaissance d'une image...) ont sur une machine ayant une architecture classique une complexité qu'on peut déterminer *a priori*, et qui se traduit en un temps d'exécution fonction de la longueur de l'input, cette fonction étant de l'ordre d'un polynôme de degré n (n pouvant être 3 ou 4 dans le cas de l'analyse d'une phrase par exemple), ce qui ne coïncide pas avec les temps extrêmement rapides que met l'esprit humain pour effectuer de telles tâches. De plus en plus, apparaît l'idée que si l'esprit humain va si vite dans l'accomplissement de ces tâches, c'est parce qu'il opère d'une manière *massivement parallèle* (comme une machine qui aurait un très grand nombre de processeurs, ces processeurs effectuant chacun leur part de calcul sans nécessairement attendre que les calculs partiels effectués par les autres soient terminés).

On commence à connaître de telles machines mais leur technologie et surtout leur utilisation sont encore balbutiantes.

L'aspect symbolique est également contesté parce que la comparaison de l'esprit humain avec un système formel, souvent mis en avant, conduit là encore à des conséquences irréalistes. En effet, un système formel est fait pour démontrer des énoncés vrais. Il fonctionne quasiment sans erreur, contrairement à notre intelligence humaine qui connaît le tâtonnement, l'erreur et l'approximation. En contrepartie de cette capacité d'erreur et d'approximation, le fonctionnement de notre cerveau gagne en souplesse. Une erreur quelque part ne met pas "en panne" tout le système (comme c'est le cas d'un système logique formel: la moindre contradiction au sein du système le rend complètement inconsistant et inapte à toute utilisation!).

1.7.2. Le connexionnisme

Face au courant cognitiviste, Varéla oppose le connexionnisme, qui est la conception selon laquelle les activités intellectuelles doivent être vues directement comme des produits de l'activité neuronale, sans passer par les symboles. Pour illustrer son propos, il donne l'exemple de réseaux de neurones qui ont été conçus de manière artificielle, même s'ils le sont à partir d'une extrême simplification. Imaginons par exemple un "anneau" de cellules, pouvant chacune prendre un état 1 ou 0 (activé, non activé) et tel que chaque cellule calcule son état à partir de ses deux voisines au moyen d'une fonction simple (un connecteur "ou" par exemple etc.). On peut étudier mathématiquement de manière très précise ce qui se passe alors. On assiste presque toujours à des oscillations des états cellulaires (balançant entre le 1 et le 0) qui finissent soit par se stabiliser, soit par être le siège d'une certaine cyclicité: on parle dans ces cas d'"attracteurs". On a défini des systèmes dynamiques dotés d'attracteurs. C'est ce genre de phénomène, la naissance de régularités au sein d'un système dynamique, que Varéla appelle *l'émergence*. Selon lui, l'émergence expliquerait l'apparition d'états qu'on peut qualifier d'états mentaux. Dans cette perspective, on n'a plus, comme dans le cas des machines de Turing, un ensemble *fini fixe* d'états, donné *a priori* une bonne fois pour toutes, avec des calculs se faisant par transitions d'état à état. On a au contraire des états d'équilibre qui se constituent au gré d'une dynamique et n'ont d'existence que temporaire. Un tel système a l'avantage de la souplesse que n'ont pas les systèmes symboliques, en effet le mode de traitement est plus "statistique" que "symbolique". Si une donnée est défaillante à un endroit (par exemple une activation d'un neurone), la défaillance sera compensée par des cellules proches. Ainsi par exemple peut-on fabriquer des réseaux de neurones artificiels qui sont capables de "reconnaître" des formes mêmes si ces dernières ne correspondent pas exactement à celles pour lesquelles ils ont été entraînés.

1.7.3. *L'énaction*

Il y a plus profond encore dans l'approche de Varéla. En effet, on peut se demander si l'activité cognitive humaine est correctement représentée par les systèmes d'IA connus, qui fonctionnent sur des *micro-mondes*. Ainsi, l'un des grands succès de l'IA est le robot SHRDLU de Terry Winograd, qui était capable de manipuler correctement des blocs géométriques, même si on s'adressait à lui dans une langue naturelle (simplifiée...). Un autre grand succès est aussi évidemment *Deep Blue*, le fameux ordinateur d'IBM, capable de battre le champion du monde d'échecs. Aussi impressionnantes que soient ces machines, il n'en reste pas moins qu'elles ont été programmées pour accomplir des tâches précises, dans un univers où toutes les possibilités de mouvements ont été étroitement définies. Elles ignorent les perturbations de l'extérieur (le bruit de la salle, la température plus ou moins élevée, le problème étranger qui soudain surgit dans la tête du sujet, le perturbe et peut aller jusqu'à prendre le devant de la scène). Inversement, elles n'ont pas eu à décider qu'elles allaient s'attaquer à ce type de problème ou à un autre: elles ont été programmées pour cela et ne se sont jamais posées la question. Évidemment, l'esprit humain fonctionne d'une manière complètement différente. Les activités effectuées par le sujet humain, même si en apparence elles sont moins difficiles que gagner aux échecs contre Kasparov, sont en réalité extrêmement complexes parce que non entièrement formalisables a priori (la conduite d'une voiture, la lecture du journal, l'éducation d'un enfant...). Et qui plus est, elles surgissent a priori dans un arrière-fonds de relative indécision, et elles nécessitent de la part du sujet des talents créatifs. Dans la conception connexionniste de Varéla, les connaissances viennent non pas d'un simple *enregistrement* de données de l'extérieur, comme le veut la tradition, mais d'une action qui s'origine dans le sujet lui-même, d'une sorte d'activité interprétative permanente jamais en sommeil (qui explique en partie la densité des relations internes au cerveau par rapport à celles qui le lient aux organes des sens), que Varéla appelle *énaction*. C'est cette activité qui ferait émerger les structures mentales, et qu'il convient d'encourager².

Ce faisant, il se rattache à des courants de la philosophie occidentale qui ont eux-mêmes fait la critique de la conception "idéaliste" de la connaissance, et qui sont représentés principalement par Heidegger, Merleau-Ponty, et d'autres phénoménologues.

1.7.4. *L'inscription corporelle de l'esprit*

La perspective connexionniste développée par F. Varéla conduit à une critique radicale de la notion de **représentation**, notion qui, comme nous l'avons vu, est au coeur même de la tradition fonctionnaliste. Comment imaginer une connaissance sans représentation? Le concept d'*énaction* est une tentative de réponse. La position varélienne est schématiquement la suivante : notre connaissance du cerveau et du

² Ce thème a parfois été repris par des spécialistes en Sciences de l'Education....

système nerveux nous révèle qu'il y a beaucoup plus de connexions internes que de connexions "dirigées vers l'extérieur", nous pouvons donc en déduire que l'activité principale des cerveaux n'est pas de *représenter le monde extérieur*, mais d'effectuer en permanence *des modifications de soi*. Une métaphore parfois utilisée par Varéla est celle du système immunitaire. Celui-ci est en effet tel qu'il "réagit à tout ce qui n'est pas soi", assurant ainsi la défense de l'organisme. Le fonctionnement du réseau immunitaire est resté longtemps mystérieux: comment un anticorps agit-il? comment un lymphocyte s'y prend-il pour distinguer ce qui fait partie du soi de ce qui n'en fait pas partie? C'est le biologiste danois Niels Jerne qui a eu le premier l'idée qu'il fallait concevoir le système immunitaire comme un réseau d'événements interconnectés, qu'autrement dit il faut étudier dans sa "clôture opérationnelle". Il ne vient à personne l'idée de dire que ce système fabrique des "représentations" des corps étrangers, ni même qu'il fabrique une représentation du moi. Son action est simplement permanente, c'est une suite de modifications internes qui s'effectuent en chaîne pour rétablir des équilibres chaque fois que des perturbations proviennent de l'extérieur, et on qualifie de reconnaissance d'un corps étranger cette suite de modifications elle-même. Il est pertinent de rapprocher le fonctionnement de l'esprit de celui du système immunitaire. Au lieu de dire que notre pensée "représente" le monde extérieur, on dit alors qu'elle est un système autonome qui *énacte* un monde (qui *l'agit* en quelque sorte, si le verbe pouvait être utilisé transitivement), en tant que domaine de distinctions qui s'incarne dans notre cerveau. Ainsi, le monde ne nous est-il pas "prédonné". Et à la question du "siège de la conscience" qu'un philosophe cartésien pourrait poser et qui n'aurait plus guère de sens en face d'un univers tellement distribué (car une telle suite d'événements s'enchaînant les uns à la suite des autres n'aurait pas de lieu pour installer une permanence de la conscience), Varéla répond tout simplement en remettant en doute la permanence du soi, se référant en cela à la tradition du bouddhisme tibétain (principalement *la Madhyamika* ou *voie du milieu*, inventée par le sage *Nagarjuna*).

1.7.5. Critique du connexionnisme

Comme on vient de le voir, la théorie de l'émergence et de l'énaction possède de nombreux arguments séduisants en sa faveur. Elle est une voie complètement différente du fonctionnalisme. Il est dit parfois que les deux voies pourraient se rejoindre. Mais cela paraît très hypothétique, même si Varéla lui-même conçoit une perspective de complémentarité entre les deux approches :“ dans la perspective inclusive, la nécessité d'un niveau symbolique est reconnue, mais on laisse ouverte la possibilité que ce niveau soit seulement approximatif. Autrement dit, les symboles ne sont pas pris pour argent comptant, ils sont considérés comme des descriptions approchées à un niveau macroscopique de certaines opérations dont la détermination effective se situe à un niveau subsymbolique”.

Il faut en effet parvenir à l'établissement d'un lien entre ce que nous reconnaissons comme des symboles dans notre activité cognitive de tous les jours et nos activités neuronales inconscientes existant au niveau "subsymbolique". Ce lien n'existe pas aujourd'hui. La notion d'attracteur dans un système cognitif est séduisante (et observable empiriquement) mais on n'est pas capable d'assigner à un attracteur observé un contenu et on fait encore comme si les contenus n'intervenaient pas dans le fonctionnement de la cognition, ce qui est manière, en un sens, de renouveler la démarche béhavioriste.

1.8. Conclusion :

Nous avons abordé beaucoup de conceptions concernant la théorie de l'esprit. On doit remarquer que chacune possède une part de conviction mais qu'aucune n'est entièrement convaincante parce que chacune recèle une part de mystère qu'elle n'arrive pas à éliminer. Pour le dualisme, c'est bien sûr le paradoxe du fantôme ("the ghost in the machine"), mais pour le monisme, nous avons aussi bien, dans le cas du matérialisme éliminativiste, le paradoxe qu'il y aurait à éliminer les entités abstraites (alors qu'on est bien obligé de faire appel à elles dans l'énoncé même de l'hypothèse!), que dans le cas du fonctionnalisme, le problème du mentalais, ou que dans celui des théories connexionnistes, le mystère de l'"émergence".

Dans le cours qui suit, nous nous concentrerons principalement sur des modèles qui proviennent du courant fonctionnaliste et que l'on qualifie aussi souvent de *symboliques*, de *cognitivistes* ou de *computationnalistes*. La raison principale en est qu'ils sont les plus connus et les plus étudiés et qu'il nous apportent, en dépit des critiques venant du connexionnisme, si ce n'est toute la vérité sur la cognition, du moins *une approximation* intéressante de celle-ci. On peut par exemple citer le cas du modèle chomskyen du langage: il peut être critiqué (voir à ce propos les critiques de G. Lakoff, reprises par Edelman), mais néanmoins il conduit à formuler de grandes hypothèses (concernant par exemple l'innéité du langage) qui peuvent être vérifiées empiriquement (voir à ce propos le livre de S. Pinker cité ci-dessus). Mais nous ne nous limiterons pas à l'exposé de ces modèles: nous envisagerons également leurs critiques.