

Peut-on modéliser la conscience à l'aide d'un système informatique ?

La réponse à la question de savoir si l'on peut modéliser la conscience implique d'abord, quels que soient ses *a priori* théoriques, de préciser non seulement ce que l'on entend par conscience, mais aussi le sens précis que l'on donne à modéliser. On a souvent insisté sur l'importance de distinguer les diverses acceptions du terme de conscience (cf. en particulier Block, 1993, 1995) et, dans cet ouvrage, Bernard Lechevalier décrit bien la polysémie de ce mot, et Etienne Balibar montre combien l'histoire même de ce mot est indissociable de l'histoire des théories philosophiques sur le sujet. Mais la polysémie du mot « modélisation » mérite aussi qu'on s'y arrête pour dégager, là encore, les différentes acceptions de ce terme. En effet, depuis plusieurs années, avec le formidable développement de l'outil informatique, cette activité multiforme a pris une ampleur considérable, sans que soit toujours bien défini le statut épistémologique des « modèles » ainsi réalisés.

Parfois ces modèles sont conçus avec une finalité propre, qui ne doit rien au système modélisé, sinon une inspiration qui a conduit à en utiliser certaines caractéristiques. Dans ces cas, la confection du modèle n'apprend pratiquement rien sur le système auquel il fait référence. On est dans le domaine de l'analogie et de la métaphore, ce qui ne signifie pas d'ailleurs que cela ne présente aucun intérêt : au contraire, le « nomadisme » des concepts scientifiques est une source de richesse et de productivité, à condition que l'on ne se méprenne pas sur sa nature heuristique. A l'autre extrême, certains modèles sont de véritables démonstrations constructives : en exhibant un système qui fonctionne comme le système étudié, en respectant aussi bien ce que l'on connaît de son organisation interne que ce que l'on peut observer de ses comportements, on prouve la cohérence et la complétude de la théorie qui a permis de construire ce modèle. De tels modèles peuvent jouer alors un rôle prédictif : en le faisant fonctionner dans telle ou telle condition, on obtient des valeurs d'observables que l'on peut ensuite comparer expérimentalement avec des mesures sur le système réel, les résultats amenant généralement à modifier tel ou tel aspect du modèle : en bref, un modèle de ce type est une expression opérationnelle d'une théorie du phénomène étudié. Bien entendu, la grande majorité des modèles se situe entre ces deux extrêmes. Ils ne cherchent à représenter qu'un aspect partiel du phénomène, ou encore ils ne modélisent qu'une structure et une organisation simplifiées du système étudié. Du coup, la vérification expérimentale de la validité du modèle n'est plus réellement possible. Néanmoins, ces modèles gardent une valeur explicative dans la mesure où l'analyse de leur comportement permet de mieux comprendre des mécanismes généraux qui aident à construire une théorie plus complète du phénomène étudié. Quoi qu'il en soit, on reste avec ces modèles dans le cadre de la théorie explicative : de la même façon qu'un modèle informatique de dynamique des fluides permet de prévoir assez précisément les conditions dans lesquelles une aile d'avion atteindra sa charge de rupture, sans que cette simulation ne reproduise réellement le phénomène (il n'y a ni vent, ni avion, ni « crash » dans l'ordinateur), on ne demande pas à ces modèles de la conscience de « produire » un phénomène de conscience dans et pour la machine. Or, contrairement à ce qui se passe pour les catastrophes aéronautiques, le problème peut se poser de savoir si une machine « intelligente » est susceptible d'être le siège de véritables phénomènes conscients : la science-fiction nous a habitué à considérer cette éventualité en alimentant notre imaginaire d'ordinateurs n'en faisant qu'à leur tête et de robots fomentant des révoltes spartakistes... C'est alors un tout autre sens de modèle qui est utilisé, plus proche de celui de modèle réduit : le problème que pose ce type de modélisation, c'est de savoir s'il est possible de faire émerger, dans un système informatique, un nouveau type de phénomène que l'on pourrait appeler « conscience artificielle ».

On a donc affaire à toute une « gamme » de modélisations possibles, quel que soit le

phénomène auquel on s'intéresse. La question qui se pose est donc de savoir jusqu'où l'on peut aller sur cette échelle, quand il s'agit de modéliser la conscience, en tenant compte bien sûr des diverses acceptions de ce mot, et donc des divers phénomènes que l'on regroupe sous ce terme. Nous n'avons pas la prétention de remplir ici complètement ce programme : cela réclamerait un travail colossal, et cela dépasse de toute façon nos compétences très limitées dans ce domaine. Nous allons donc plutôt décrire quelques types de modèles informatiques, situés à différents endroits de l'échelle, qui vont nous servir de jalons pour discuter des limites et de l'intérêt que peut représenter cette activité de modélisation pour mieux comprendre et cerner les phénomènes de conscience.

1. Les métaphores de l'Intelligence Artificielle

L'Intelligence Artificielle, de par la nature même de ses objectifs, n'a jamais été avare de métaphores concernant le fonctionnement de l'esprit humain. Il est important de noter d'ailleurs que les analogies se sont développées, à un rythme parfois extravagant, dans les deux sens. D'un côté, les informaticiens ont souvent pris leur inspiration dans ce que l'on connaissait de la structure et de l'organisation du cerveau humain pour concevoir les outils qui leur ont permis de faire réaliser à des systèmes informatiques des tâches dites « intelligentes ». Réciproquement, en sens inverse, les productions de l'Intelligence Artificielle ont été presque systématiquement prises, avec plus ou moins de bonheur, comme source de métaphores, par des chercheurs en quête de schèmes explicatifs du comportement humain, que ce soit en psychologie cognitive, en neuropsychologie, ou en philosophie de l'esprit. Il est donc intéressant de relever les usages du terme de « conscience » dans les systèmes produits par l'Intelligence Artificielle, et de discuter de la validité de l'extension nouvelle qu'a pris ce terme dans ces milieux.

1.1. Les systèmes à base de connaissances

Les systèmes à base de connaissances sont les systèmes les plus classiques de l'Intelligence Artificielle. Héritiers de la génération des systèmes experts, ils sont structurés en deux composantes essentielles :

- d'une part, une ou plusieurs bases de connaissances, qui contiennent des « faits », c'est-à-dire des connaissances sur « l'état des choses » représenté par le système informatique, des « règles », c'est-à-dire des connaissances sur les lois d'interaction et d'évolution de l'univers représenté, et enfin des « méta-règles », c'est-à-dire des connaissances sur les processus de création et de modification de ces règles, ainsi que la manière dont elles doivent s'appliquer ;

- d'autre part, un ou plusieurs « moteurs d'inférence », qui sont généralement des mécanismes de déduction capables d'utiliser de manière systématique les connaissances contenues dans une base pour en tirer de nouvelles connaissances (très grossièrement, en appliquant les règles aux faits déjà présents pour en tirer de nouveaux faits, et en appliquant les méta-règles aux règles déjà présentes pour en tirer de nouvelles règles et, parfois même, de nouvelles méta-règles).

Il n'est pas question de présenter ici en détail les différentes architectures qui respectent peu ou prou cette définition très approximative et incomplète de tels systèmes, mais cela va nous suffire pour repérer quelques uns des usages de la métaphore de la conscience dans ces systèmes.

On a souvent associé à des connaissances « conscientes » les connaissances contenues dans les bases de connaissances, qui sont déclaratives et explicites, au sens où elles sont exprimées dans des langages formels de haut niveau, en les opposant aux connaissances procédurales, du type de celles qui sont intégrées dans les mécanismes des moteurs d'inférence qui sont, elles, implicites, « enfouies » dans la programmation des procédures de traitement, et qui sont donc associées à des connaissances « inconscientes ». L'analogie porte sur plusieurs aspects :

- d'abord, ces connaissances sont facilement accessibles et verbalisables, au sens où il est très facile de faire correspondre automatiquement à chaque déclaration dans le langage formel une proposition énoncée dans une langue naturelle ;

- la manière dont une nouvelle connaissance a pu être déduite peut aussi, avec la même facilité, être explicitée et verbalisée : autrement dit, le système peut produire des explications sur son comportement en « racontant » le déroulement de son raisonnement ;

- enfin l'existence de méta-règles (qui, dans le meilleur des cas, peuvent aussi s'appliquer à elles-mêmes) permet d'implémenter des mécanismes « auto-réflexifs », au sens où le système peut lui-même analyser ses propres déductions et modifier en conséquence ses connaissances et même sa façon de les utiliser.

Cette analogie a été développée dans de nombreuses directions. Pour ne prendre qu'un exemple, on a souvent comparé le processus de « compilation de règles », qui consiste à rendre procédurale l'utilisation d'un ensemble de connaissances pour obtenir un système plus rapide et efficace, à « l'acquisition de réflexes » : de la même façon que l'on peut grâce à l'entraînement rendre automatiques et stéréotypés certains comportements (conduite automobile, mouvements sportifs, etc.) qui ne pouvaient être exécutés qu'avec difficulté et qui réclamaient beaucoup d'attention consciente en début d'apprentissage, la compilation de règles remplace un processus plus lent d'interprétation de règles explicites et modifiables par un processus plus rapide, mais figé et beaucoup plus difficilement analysable. Le processus inverse, à savoir l'acquisition de règles explicites par apprentissage à partir de données non explicites existe aussi dans des systèmes de ce genre, mais il a été plus amplement développé dans un autre paradigme de l'Intelligence Artificielle, le connexionnisme, sur lequel nous allons maintenant nous pencher.

1.2 Les systèmes connexionnistes

Le connexionnisme a en effet permis d'étendre considérablement les métaphores de l'Intelligence Artificielle, en proposant des outils appelés « réseaux neuronaux » en raison même de l'analogie de leur fonctionnement avec celui des structures neuronales que l'on peut observer dans le système nerveux animal. Bien entendu, les « neurones formels » du connexionnisme sont des modèles très simplifiés de leur correspondants naturels, et les réseaux connexionnistes ne comportent généralement qu'un nombre ridiculement petit de neurones quand on les compare avec le cerveau du moindre batracien, mais ces réseaux permettent néanmoins de reproduire un certain nombre de caractéristiques générales du fonctionnement d'une population de neurones et donc de mieux comprendre les potentialités de calcul et de représentation que l'on est en droit d'associer à l'activité de telles populations. Dès le début, ces perspectives de modélisation de systèmes cognitifs ont joué un rôle important dans le développement du connexionnisme, au moins autant que des objectifs d'ingénierie, en particulier dans le groupe qui a le plus contribué à populariser ce domaine de recherche, le groupe dit « PDP » (*Parallel Distributed Processing* : cf. McClelland *et al.*, 1986). Aujourd'hui, ce domaine de recherche s'est approfondi et diversifié, aussi bien en direction de réalisations d'ingénierie (en particulier en reconnaissance de formes) qu'en direction de la modélisation du système nerveux (on trouvera un exposé assez complet des recherches dans cette voie dans Churchland et Sejnowski, 1992).

L'un des intérêts majeurs du connexionnisme, c'est de donner une large place à des mécanismes d'apprentissage : les réseaux sont en général adaptatifs, au sens où leur comportement évolue au cours du temps en fonction de lois d'apprentissage, qui leur permettent de mieux remplir des tâches auxquelles ils ont été « exercés ». Ainsi peut-on faire « émerger » des régularités dans le comportement du réseau, et un problème que l'on s'est posé, c'est celui d'explicitier les règles ainsi apprises par le système : c'est ce que l'on a appelé « l'extraction de règles » dans un réseau. D'une manière générale, les chercheurs en Intelligence Artificielle ont beaucoup discuté de la relation que l'on devait postuler entre ces

systèmes « connexionnistes » et les traitements plus classiques, appelés par opposition « symboliques » (cf. par exemple Smolensky, 1988). On devine comment la métaphore de la conscience a pu se glisser dans ces débats, en particulier à propos des systèmes dits « hybrides », c'est-à-dire possédant à la fois une composante connexionniste et une composante symbolique (cf. Grumbach, 1994) : le niveau connexionniste représenterait des processus non-conscients, dont émergerait un niveau conscient, la composante symbolique, qui possède les propriétés que nous avons passées en revue plus haut : capacités d'explicitation et de verbalisation, auto-réflexivité, etc.

1.3 Des métaphores « unidirectionnelles »

Quel statut épistémologique doit-on accorder à ces analogies ? Peuvent-elles nous apprendre quelque chose sur les phénomènes regroupés sous le terme de conscience ? Pour répondre correctement à cette question, il faut analyser de plus près les éléments qui sont à la base de l'analogie entre certains processus d'un système informatique et des phénomènes liés à la conscience. L'un des points essentiels, comme on l'aura remarqué, c'est le caractère explicite et verbalisable de certaines connaissances et de certains traitements. Mais, le problème qui se pose, c'est de savoir *pour qui* ces connaissances sont explicitables : pour le système lui-même, ou pour l'humain qui conçoit et/ou manipule le système ?

A vrai dire, pour la machine, que ces connaissances soient écrites en langage machine, en langage formel de haut niveau, ou en langue naturelle, cela ne lui donne pas plus accès à leur sens : les chaînes de caractères qui les représentent ne forment un système de signes que pour les humains, soit directement s'il s'agit de langue naturelle, soit de toute façon par l'intermédiaire de la langue naturelle dans laquelle ont été formulées les conventions qui assignent une syntaxe et une sémantique au langage formel utilisé. De tout cela, rien n'est accessible au système, que l'on désigne par là la machine, le logiciel, ou les processus de traitement. À cet égard, le célèbre argument de Searle (1980), dit « de la chambre chinoise », s'applique parfaitement. Rappelons que Searle évoque un homme enfermé dans une pièce, qui ne connaîtrait pas le chinois, et à qui on demanderait de manipuler des idéogrammes suivant des règles précises. Celui-ci pourrait très bien répondre correctement par écrit à des questions en chinois, donnant ainsi de l'extérieur l'impression qu'il sait le chinois, alors qu'il ne comprend en fait pas un traître mot à ce dont il est question. De la même manière, les systèmes informatiques qui manipulent des symboles n'ont pas pour autant accès à leur signification.

Ainsi les connaissances « explicites », « symboliques », « de haut niveau », ne le sont que pour les humains qui observent le système, et du coup, il ne saurait être question de parler de connaissances conscientes pour le système lui-même : tout au plus, peut-on parler de connaissances explicitables et verbalisables pour l'ensemble constitué par le système informatique et l'humain qui le manipule. Il s'agit donc en quelque sorte d'une métaphore « unidirectionnelle ». En effet, dans un sens, l'analogie permet effectivement de construire des systèmes informatiques dont l'architecture reflète en partie des propriétés associées à la conscience, du moins dans certaines de ses acceptions : en fait il s'agit grosso modo de propriétés de ce que Block appelle *access-consciousness* et *monitoring-consciousness* (Block, 1993, 1995). Mais en revanche, dans le sens opposé, l'analogie n'a rien à nous apprendre sur le phénomène de conscience lui-même, puisque les systèmes ainsi construits, quelles que soient leurs qualités par ailleurs (qu'il n'est nullement question de dénigrer ici, en particulier pour l'étude de mécanismes cognitifs autres que celui de la conscience), ne simulent pas véritablement un équivalent de ce que serait un phénomène de conscience. Au fond, le seul enseignement que l'on peut en tirer, c'est que ces systèmes fournissent autant de contre-exemples qui, comme l'expérience de pensée « de la chambre chinoise » de Searle (1980), tendent à démontrer qu'il n'est pas possible de statuer sur l'existence de phénomènes conscients dans un système à partir d'une observation objective extérieure du comportement

du système en question.

2 Les modélisations neurophysiologiques

Tournons-nous maintenant vers d'autres types de modèles, qui proviennent d'une toute autre source, et dont les finalités sont aussi très différentes : il s'agit des modèles issus de la neurophysiologie. Généralement, ces modèles se donnent comme objectif de rendre compte de l'activité cérébrale telle qu'on peut l'observer et la mesurer, en l'intégrant dans un cadre explicatif qui relierait les états et les processus physiologiques à des états et des processus mentaux. La plupart de ces travaux se situent dans le cadre de l'hypothèse d'une correspondance, plus ou moins étroite selon les cas, entre états physiologiques et états mentaux. De fait, il s'agit pour l'instant plutôt de modèles théoriques, dont l'implémentation n'est pas encore à l'ordre du jour, ou qui se limite, quand elle est tentée, à des systèmes très rudimentaires, bien en dessous du niveau de complexité qui serait nécessaire pour qu'elle serve à autre chose qu'à une illustration pédagogique des idées théoriques dont elle est issue. Nous allons décrire rapidement, à titre d'exemples, deux de ces modèles assez « typiques » de ce que l'on peut rencontrer dans la littérature avant de discuter plus avant de ce qu'on est en droit d'en attendre d'une manière générale.

2.1 Les boucles réentrantes

Une des hypothèses souvent émises fait correspondre aux états de conscience une certaine localisation dans le cerveau, au moins sous la forme de circuits spécifiques, sinon sous la forme de modules spécialisés. Le modèle d'Edelman fait partie de cette catégorie (Edelman, 1989, 1992 ; voir aussi Clancey, 1993, pour une analyse de ce modèle du point de vue de l'Intelligence Artificielle). Edelman définit d'abord ce qu'il appelle la « conscience primaire », que nous partagerions avec un certain nombre d'espèces animales, et qui serait une sorte de conscience dans le présent de nos perceptions et de nos intentions, mais qui ne saurait s'étendre dans une temporalité plus large, et qui exclurait donc la mémoire de nos états conscients passés, la capacité d'imaginer d'autres états conscients, la conscience complète de soi, des autres et du monde : toutes ces propriétés étant caractéristiques de ce qu'il appelle la conscience de haut niveau, réservée aux humains, et intimement liée à la faculté de langage. Pour nous en tenir ici à la conscience primaire, Edelman la décrit comme résultant de l'activité d'une boucle réentrante qui implique deux grands ensembles corticaux : d'une part les aires corticales sensorielles, qui catégorisent les entrées extéroceptives (qui sont donc le siège des activités proprement perceptives), et d'autre part, le cortex frontal, temporal et pariétal, qui mémoriserait une catégorisation dite « conceptuelle », résultat de l'interaction entre perceptions extérieures et états internes de l'organisme (intéroception).

Dans la théorie générale d'Edelman, un mécanisme, qu'il appelle *neuronal group selection*, joue un rôle fondamental : il s'agit d'un mécanisme de sélection, analogue de la sélection darwinienne pour le niveau neuronal, qui permet de structurer les apprentissages en organisant des groupes de neurones réagissant de manière cohérente, et reliés entre eux par des boucles réentrantes. Ainsi, c'est ce même mécanisme, qui est déjà mis à contribution pour expliquer d'autres activités cognitives (les boucles sensori-motrices, les catégorisations perceptives, la mémorisation, etc.), qui serait à l'origine de cette conscience primaire, en s'appliquant à des circuits spécifiques entre des modules précis. Si le modèle de la conscience primaire n'a jamais été implémenté, le mécanisme en question a été testé dans un système informatique, appelé Darwin III (Reeke et al., 1990). Ce système est une sorte d'énorme réseau connexionniste, ne comportant pas moins de cinquante mille « neurones » et de six cent mille « synapses », structurés en une cinquantaine de répertoires différents, et dont la tâche est de simuler la coordination d'un système visuel avec le mouvement d'un bras articulé muni d'un sens tactile. La taille du réseau nécessaire pour implémenter une tâche somme toute assez

simple et de bas niveau donne une idée de la difficulté d'aller beaucoup plus loin dans l'implémentation de ce modèle.

2.2 Les synchronisations d'oscillations

Un autre grand ensemble de modèles, plutôt que de supposer une localisation de l'activité neuronale liée à la conscience, propose de faire correspondre un certain mode de fonctionnement du système neuronal aux phénomènes associés à l'activité consciente. En particulier, à la suite de Crick et Koch (1990) qui semblent avoir été les premiers à développer une telle approche, un certain nombre de modèles font reposer le phénomène de conscience sur l'établissement d'oscillations synchrones dans le cortex. Par exemple, Varela (1995) a récemment proposé un modèle de ce type, dans lequel il reprend l'idée tout à fait classique (on la trouve par exemple déjà dans Changeux, 1983) d'associer à chaque acte cognitif l'activité d'une « assemblée neuronale », ensemble distribué de neurones fortement interconnectés, recrutés en une configuration spécifique de l'acte cognitif en question. L'hypothèse de Varela (1995) est que ce qu'il appelle, lui aussi, « conscience primaire » émerge de la stabilisation, pendant plusieurs centaines de millisecondes, d'une activité synchronisée (dans la bande de fréquences gamma) d'une assemblée neuronale. Ainsi la conscience ne serait pas l'apanage de tel ou tel circuit neuronal, mais serait la propriété partagée de tous les neurones du cortex, pour peu qu'ils puissent participer à de telles oscillations synchrones stabilisées. C'est la cohérence de cette activité dans tout le cortex pendant une période de temps bien déterminée qui produirait à la fois l'unicité du phénomène de conscience et son caractère à la fois stable et transitoire.

Sous l'influence de ces modèles, on a vu fleurir ces dernières années un certain nombre d'implémentations informatiques de mécanismes de synchronie dans des réseaux neuronaux. Celles-ci peuvent être très simplistes, au sens où elles se contentent de doter les unités d'un réseau connexionniste classique d'une « dimension » temporelle supplémentaire, représentant leur phase (en les considérant comme des oscillateurs de même fréquence ; cf. par exemple Shastri et Ajjanagadde, 1993). D'autres systèmes, plus intéressants mais aussi du même coup plus complexes, s'attaquent aux problèmes ardues que pose le couplage de chaînes d'oscillateurs (cf. par exemple Von der Malsburg et Bienenstock, 1986 ; ou les travaux d'Abeles sur les *synfire chains* : Abeles, 1991, Abeles et al., 1994). Là aussi, on peut dire que l'informatique est quelque peu « en retard » par rapport aux modèles théoriques : les systèmes restent très frustes par rapport à ce dont on aurait besoin pour commencer à simuler de manière un tant soit peu réaliste les phénomènes décrits.

2.3 Les présupposés théoriques de ces modèles

Comme nous le disions, dans la plupart de ces travaux, est posée l'hypothèse d'une correspondance, plus ou moins étroite selon les cas, entre états physiologiques et états mentaux. Cette attitude n'est pas spécifique des modélisateurs : c'est la position défendue plus ou moins explicitement par les neurophysiologistes en général, à de notables exceptions près (Popper et Eccles, 1977). Mais pour les modélisateurs, cette hypothèse retentit immédiatement sur le statut épistémologique de leurs modèles. En effet, puisqu'il s'agit de modéliser uniquement des états et des processus physiques, le modèle ne peut rien nous apprendre sur les mécanismes qui produisent les états mentaux, c'est-à-dire sur la nature de cette correspondance entre états mentaux et états physiques qui est supposée a priori. Autrement dit, ces modèles ne peuvent pas aider à répondre aux questions suivantes : les états mentaux sont-ils une conséquence obligatoire des états physiques auxquels ils sont associés ? Cela dépend-il de la nature biologique du substrat matériel dans lequel les processus se produisent ? etc. Leur intérêt et leur ambition sont ailleurs : c'est de pouvoir tester la plausibilité d'avoir choisi de mettre en correspondance tel état ou tel processus physiologique

avec tel phénomène mental plutôt qu'avec tel autre. Et bien entendu, il s'agit là de problèmes tout aussi importants, dont la plupart des chercheurs de ces disciplines pensent sans doute qu'ils doivent d'abord être résolus, avant que l'on puisse se poser le problème de la nature de la correspondance avec quelque chance de succès.

Ainsi, ces modèles ont un statut épistémologique assez clair, ce qui ne les rend pas plus faciles à construire pour autant : comme on l'a vu, on est encore loin de la confection de systèmes de taille réaliste, qui permettent de pouvoir réellement tester leur validité en les comparant avec les faits (qui, rappelons-le, concernent avant tout le cerveau humain, le seul pour lequel l'expérience subjective de phénomènes liés à la conscience peut être relatée). Cela ne signifie pas, une fois de plus, que l'implémentation de modèles « réduits » n'apporte rien pour l'instant. Même s'ils ne sont pas réalistes, ces systèmes peuvent être étudiés pour eux-mêmes, et l'analyse de leurs comportements permet souvent de dégager de nouvelles hypothèses de travail, de conforter certaines intuitions ou au contraire d'en invalider d'autres, par exemple sur des conditions nécessaires à l'existence de telle ou telle propriété du système (en montrant qu'un système restreint peut posséder ces propriétés sans que les conditions en question soient vérifiées, par exemple). Il est prévisible d'ailleurs que l'éventail de ces systèmes va s'étendre considérablement, au fur et à mesure que seront émises de nouvelles hypothèses sur le substrat neuronal des phénomènes de conscience. Pour ne prendre qu'un exemple, les nouvelles hypothèses sur le rêve avancées par Tassin (1995), dans lesquelles le phénomène de conscience associé au rêve ne se produirait pas pendant le sommeil paradoxal, mais serait une « reconstruction » au moment du réveil, liée à des relâchements massifs de neuromodulateurs spécifiques sur des intervalles de temps très courts, pourraient bien engendrer une série de nouveaux modèles, qui n'aient rien à voir avec les deux types de modèles présentés ci-dessus.

3 Vers des systèmes « conscients » artificiels ?

Une dernière grande question se pose. Si ces modèles aboutissaient un jour à des systèmes informatiques capables de « copier » fidèlement le fonctionnement du système nerveux, seraient-ils pour autant dotés de conscience phénoménale, ou n'auraient-ils que le statut de « zombies », se conduisant comme s'ils étaient conscients, mais sans rien éprouver qui ressemble à nos *qualia* ? Autrement dit, la simulation du système physiologique sous-jacent aux phénomènes de conscience suffit-elle pour faire apparaître une conscience dans le système simulateur lui-même ? Comme on l'a vu, l'état actuel des réalisations ne laisse pas présager d'expérimentations de ce type avant longtemps, si tant est qu'elles soient possibles un jour. Cette question reste donc pour l'instant une pure spéculation que les modélisations informatiques présentes ne peuvent en rien contribuer à éclaircir. Elle reste donc essentiellement l'apanage des philosophes de l'esprit qui, il est vrai, ne se font généralement pas prier pour occuper le terrain, palliant avec une facilité déconcertante l'impossibilité d'expérimentation par leurs célèbres expériences de pensée, avec la conséquence que l'on imagine aisément : toutes les réponses à la question posée ont été proposées, analysées, débattues, combattues. Nous nous limiterons ici à présenter rapidement quelques arguments en faveur d'une réponse positive et d'une réponse négative, sans aucune prétention à l'exhaustivité, avant de conclure en nous posant une autre question : pourquoi devrions-nous nous efforcer de résoudre ce problème ?

3.1 Oui

Les tenants du oui, appelés « fonctionnalistes », s'appuient sur un raisonnement simple : si l'on reproduit à l'identique les conditions d'apparition de la conscience, quel que soit le support matériel que l'on utilise, il n'y a pas de raison de supposer que cela ne produira pas le même effet, à moins de penser que les cellules vivantes que sont les neurones possèdent des

propriétés mystérieuses non élucidables, hypothèse qu'une saine attitude scientifique doit rejeter à moins d'y être contraint par les faits. Un de leurs arguments repose sur des expériences de pensée telles que celle-ci : on remplace progressivement les neurones d'un cerveau humain par des composants électroniques. Il est difficile d'imaginer que la modification d'un seul élément suffise à elle seule à supprimer l'expérience phénoménale de la conscience. Or quand le processus est terminé, le tour est joué : le système est entièrement électronique, et la conscience est restée intacte.

Cette prise de position s'accompagne le plus souvent d'une description des phénomènes conscients en termes de processus informationnels, ce qui bien sûr conforte l'idée que ces phénomènes seraient indépendants du support sur lequel « circule » cette information. C'est ainsi que Dennett (1991), pour « expliquer » la conscience (le titre de son livre, *La conscience expliquée*, est sans ambiguïté sur cette ambition), utilise pleinement la métaphore informatique, dans la version très à la mode des systèmes multi-agents (cf. Ferber, 1995) de l'Intelligence Artificielle Distribuée (avec des processus parallèles traitant chacun l'information à sa façon, et communiquant entre eux, à la manière de *La société de l'esprit* de Minsky, 1985). Dans ce cadre, la conscience est décrite comme un logiciel acquis par Homo Sapiens : « *Il [Homo Sapiens] a créé une sorte de logiciel [souligné par Dennett] pour augmenter ses pouvoirs sous-jacents* » (Dennett, 1991, trad. fr., p. 240). Ce type « d'explications », qui illustre parfaitement l'utilisation abusive de la métaphore informatique que nous dénonçons (cf. section 1.3 de ce chapitre), est entièrement revendiqué par Dennett : « *Il se trouve que la façon dont on peut imaginer cela [comment le cerveau pourrait être le siège de la conscience] consiste à penser le cerveau comme une certaine sorte d'ordinateur. [...] En considérant nos cerveaux comme des systèmes de traitement d'information, nous pouvons progressivement dissiper le brouillard et frayer notre chemin à travers le grand fossé, en découvrant comment nos cerveaux peuvent produire tous ces phénomènes.* » (Dennett, 1991, trad. fr., p. 537). Sans revenir sur les raisons qui nous ont fait critiquer plus haut ce type de discours, remarquons simplement qu'il n'est pas étonnant que ce type de conception de la conscience conduise à répondre oui à la question que nous nous posons : si par avance la conscience est décrite comme un processus de traitement de l'information, il serait surprenant qu'un ordinateur ne puisse lui servir de support.

3.2 Non

À l'opposé, les partisans du non réfutent bien entendu que la conscience soit un tel processus informationnel. On a déjà parlé de la position de Searle (1980, voir aussi Searle, 1996) qui montre donc que ce n'est pas parce que l'on peut décrire les phénomènes de conscience en termes de traitement d'information que toute entité qui reproduirait ce traitement serait automatiquement dotée de cette « conscience » : la manipulation de symboles ne fait sens que pour les entités pour lesquelles ces symboles font sens, et jusqu'à présent, tous les symboles manipulés par les systèmes informatiques ne forment des systèmes de signes que par l'intermédiaire du langage humain.

Cet argument négatif, même s'il semble irréfutable, ne nous dit pas pour autant ce qu'est la conscience, et n'exclut pas a priori que des systèmes artefactuels puissent être, un jour, le siège de phénomènes de conscience. Après tout, on peut imaginer une population de « machines », que l'on aurait dotées de tout ce qu'il faut pour interagir avec le monde à la manière des espèces vivantes, sauf qu'elles seraient faites de composants électroniques qui auraient donc des capteurs, des effecteurs, des circuits adaptatifs permettant des apprentissages, qui de plus se reproduiraient et obéiraient même à une loi d'évolution de type néo-darwinien si cela peut aider : pourquoi serait-il impossible qu'émerge de cette évolution une génération de machines douées de l'équivalent de notre conscience ? Bien sûr, on peut arguer d'une spécificité de la matière vivante qui expliquerait l'irréductibilité du phénomène, mais alors la charge de la preuve est du côté de ceux qui avancent cet argument, puisqu'ils

doivent exhiber quelle est cette spécificité.

D'une certaine manière, c'est ce défi qu'a voulu relever Penrose (1994) qui développe une hypothèse tout à fait cohérente à ce sujet : la conscience serait due à un phénomène de cohérence quantique à l'échelle macroscopique, qui serait basé sur une mise en cohérence de phénomènes quantiques intervenant dans des structures microtubulaires du cyto-squelette des cellules nerveuses. Cette hypothèse a l'avantage d'être suffisante pour entraîner (si elle est vérifiée) la non-calculabilité de ce phénomène, et donc sa non-reproductibilité par une machine de type machine de Turing-von Neumann (Penrose s'appuie par ailleurs sur le théorème de Gödel comme point de départ pour démontrer que, de toute façon, il y a quelque chose qui relève du non-calculable dans notre capacité à comprendre et à construire des concepts mathématiques, à commencer par l'ensemble des entiers naturels). On le voit : il existe au moins une théorie qui s'appuie sur une hypothèse scientifique (c'est-à-dire falsifiable par l'expérience, du moins en principe : comme le dit Penrose lui-même, c'est une nouvelle théorie physique, celle des effets quantiques macroscopiques, qui doit d'abord être construite), qui expliquerait que les phénomènes de conscience ne pourraient jamais émerger d'un système informatique.

3.3 Bof

Mais on peut douter, au fond, de l'intérêt de la question elle-même. L'état actuel de l'implémentation informatique des modèles neurophysiologiques montrent que l'on est encore loin de pouvoir apporter une réponse qui soit autre chose qu'une profession de foi, en tout cas, à partir de l'informatique elle-même. Nous avons donc de toute façon bien des progrès à faire dans notre discipline avant de pouvoir apporter une contribution positive à ce débat. Comme nous l'avons dit, l'intérêt actuel des modèles informatiques n'est pas dans des simulations « réalistes » des processus neuronaux, mais plutôt dans la confection de systèmes qui testent l'intérêt théorique de tel ou tel mécanisme que les disciplines expérimentales mettent en évidence : en bref, l'informatique est utile parce qu'elle permet d'expérimenter des idées.

Or, l'objectif de réaliser une « conscience artificielle » ne s'intègre absolument pas (en tout cas, dans l'état actuel de nos techniques : il faudrait être très imprudent pour chercher à prédire les rythmes d'évolution dans ce domaine encore très jeune et encore en pleine révolution technologique) à ce travail d'expérimentation d'idées. Il faut d'ailleurs remarquer que l'on n'est même pas capable aujourd'hui de donner un critère de réussite de cette entreprise. Supposons en effet que l'on mène à bien un programme du type de celui que nous évoquions plus haut : une population de machines interagissant dans le monde à la manière des êtres vivants et se transformant progressivement par apprentissage et évolution pour atteindre une complexité et des performances qui méritent que l'on se pose la question de leur conscience. En fait, on n'aura pas alors plus d'éléments pour décider que ces machines possèdent une conscience que l'on n'en a aujourd'hui pour répondre à cette question à propos d'animaux comme les chimpanzés, les dauphins ou les fourmis ! Il serait pour le moins paradoxal que les informaticiens, qui connaissent bien les difficultés que l'on a pour démontrer si un programme informatique se termine ou non, se lancent dans un programme de recherche sans s'assurer au moins d'avoir un test de fin de ce programme !

Références

- ABELES M. (1991), *Corticonics : Neural circuits of the cerebral cortex*. Cambridge, Cambridge University Press.
- ABELES M., PRUT Y., BERGMAN H., VAADIA E. (1994), Synchronization in neuronal transmission and its importance for information processing, in G. Buzsaki, Llinas, W. Singer, A. Berthoz Y. Christen (eds), *Temporal coding in the brain*, Berlin, Springer-Verlag, 39-50.

- BLOCK N. (1993), Review of Dennett : Consciousness explained, *The Journal of Philosophy*, 60, 181-219.
- BLOCK N. (1995), On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, 18, 227-287.
- CHANGEUX J.P. (1983), *L'homme neuronal*, Paris, Fayard.
- CHURCHLAND P.S., SEJNOWSKI T.J. (1992), *The Computational Brain*, Cambridge, Mass., MIT Press.
- CLANCEY W.J. (1993), The biology of consciousness : Comparative review of Israel Rosenfield, The strange, familiar, and forgotten : An anatomy of consciousness, and Gerald M. Edelman, Bright air, brilliant fire : On the matter of the mind, *Artificial Intelligence*, 60, 313-356.
- CRICK F., KOCH C. (1990), Towards a Neurobiological Theory of Consciousness, *Seminars in the Neurosciences*, 2, 263-275.
- DENNETT D. (1991), *Consciousness explained*, Boston. Little, Brown and Cy, Trad. fr. par P. Engel, *La conscience expliquée*, Paris, Odile Jacob, 1993.
- EDELMAN G.M. (1989), *The remembered present. A biological theory of consciousness*, New York, Basic books.
- EDELMAN G.M. (1992), *Bright air, brilliant fire : On the matter of the mind*, New York, Basic books.
- FERBER J. (1995), *Les systèmes multi-agents*, Paris, InterEditions.
- GRUMBACH A. (1994), *Cognition artificielle*, Paris, Addison-Wesley France.
- McCLELLAND J.D., RUMELHART D. and the PDP Research Group (1986), *Parallel Distributed Processing : Explorations in the microstructure of cognition*, Cambridge, Mass., MIT Press, 2 volumes.
- MINSKY M. (1985), *The society of mind*, New York, Simon et Schuster, Trad. fr. par J. Henry, *La société de l'esprit*, Paris, InterEditions, 1988.
- PENROSE R. (1994), *Shadows of the mind. A search for the missing science of consciousness*, Oxford, Oxford University Press.
- POPPER K.R., ECCLES J.C. (1977), *The self and its brain*, Berlin, SpringerVerlag.
- REEKE G.N., FINKEL L.H., SPORNS O., EDELMAN G.M. (1990), Synthetic neural modeling : A multi level approach to the analysis of brain complexity, in G.M. Edelman, W.E. Gall & W.M. Cowan (Eds), *The Neurosciences Institute Publications, signal and sense, local and global order in perceptual maps*, New York, Wiley, 607-707.
- SEARLE J. (1980), Minds, brains and programs, *Behavioral and Brain Sciences*, 3, 417-457.
- SEARLE J. (1996), Deux biologistes et un physicien en quête de l'âme, *La Recherche*, 287, 62-77.
- SHASTRI L., AJJANAGADDE V. (1993), From simple association to systematic reasoning : A connectionist representation of rules, variables and dynamic bindings using temporal synchrony, *Behavioral and Brain Sciences*, 16, 417-494.
- SMOLENSKY P. (1988), On the proper treatment of connexionism, *Behavioral and Brain Sciences*, 11, 1-74.
- TASSIN J.P. (1995), Le rêve naît du sommeil, *Pour la science*, 214, 22-23.
- VARELA F.J. (1995), Resonant Cell Assemblies. A new approach to cognitive functions and neuronal synchrony, *Biological Research*, 28, 81-99.
- VON DER MALSBERG C., BIENENSTOCK E. (1986), Statistical coding and short-term synaptic plasticity : A scheme for knowledge representation in the brain, in E. Bienenstock, F. Fogelman & G. Weisbuch (Eds), *Disordered systems and biological organization*, Berlin, Springer-Verlag, 247-272.