Conscience sans Cortex

Michel Dojat, Manik Bhattacharjee, Christian Graff

Chapitre 8
Conscience sans Cortex

Michel Dojat, Manik Bhattacharjee, Christian Graff

La conscience comme fonction émergente du cerveau humain

Jusque dans la terminologie zoologique, *Homo sapiens*[1], la conscience a été pour une large part considérée comme une spécificité humaine. Débordant de ce courant dominant de la pensée occidentale, les approches philosophiques initiées dans la deuxième moitié du XX^{ème} siècle, l'antispécisme d'un côté (Singer, 1975) et la singularité informatique de l'autre (Chalmers, 2010; Good, 1965), partagent de plus en plus la conviction que certains animaux et certaines machines jouissent de facultés comparables à celles d'un humain conscient. Les enjeux moraux sont de taille. Pendant plusieurs siècles, le débat est resté souterrain pour l'ensemble des scientifiques français. Dans la droite ligne de Descartes, la physiologie expérimentale héritée de Claude Bernard (Bernard, 1865) a considéré le corps humain comme un mécanisme semblable à celui d'autres animaux. Le philosophe mathématicien, dans une vision dualiste séparant matière et esprit, accordait, mais à l'humain seul, une parcelle divine, l'âme. Sa vision de l'âme restait en accord avec la religion du Dieu incarné, créateur d'un Homme à son image, libre et supérieur à la nature et aux entités qui la composent. Ces questions spirituelles sont restées hors du champ mono-disciplinaire du scientifique et hors de portée de ses outils. Il s'en est donc tenu à l'animal-machine. Cependant la place privilégiée de l'homme a été relativisée par la théorie de l'évolution des espèces (Darwin, 1859). La neuropsychologie, appuyée par la neuroimagerie, a mis directement en lien des réseaux corticaux communs à différentes espèces et certaines facultés cognitives et affectives. Par ailleurs, sur l'aspect fonctionnel, les capacités des machines actuelles, en termes de capacité mémorielle et de traitement de l'information, les rapprochent plus des organismes vivants que des horloges des siècles passés.

Descartes proposait que, contrairement à l'homme, pour des machines « qui eussent des organes et la figure d'un singe, ou de quelque animal sans raison, nous n'aurions aucun moyen pour reconnaître qu'elles ne seraient pas en tout de même nature que les animaux » (Descartes, 1637). Il soulignait ainsi la dichotomie entre l'Homme et de simples mécanismes naturels ou artificiels. Le célèbre canard de Vaucanson, un automate reproduisant les mouvements d'un canard, aurait été construit relativement à cette posture, à laquelle se sont opposés d'autres penseurs, comme Perrault dans son « Avertissement » introduisant « La mechanique des animaux » (Perrault, 1680)[2], puis Diderot (Diderot and (Le Rond) D'Alembert, 1751-72) qui accordait volontiers une âme aux animaux dans la célèbre Encyclopédie. A leur suite, associant âme et conscience, nous reconsidérons dans quelle mesure, dans le contexte des neurosciences et de l'informatique du XXI^{ème} siècle, la conscience resterait ou non une spécificité humaine. Et si non, à quel type de conscience les entités sans cortex pourraient-elles prétendre ?

Nous nous plaçons d'entrée dans une perspective matérialiste, moniste, partagée par la majorité des neuroscientifiques (voir enquête, annexe 1). Dans cette perspective,

---

[1] du latin *homo* : être humain, et *sapiens* : intelligent, sage, raisonnable.

[2] Perrault estime d'ailleurs qu'un animal est « un être qui a du sentiment et qui est capable d'exercer les fonctions de la vie par un principe que l'on appelle âme » ; et que cette dernière « se sert des organes du corps, qui sont de véritables machines, comme étant la principale cause de l'action de chacune des pièces de la machine ».

l'esprit n'existe pas en-dehors de la matière qui en est l'origine. La conscience évoquée ici ne se réfère pas, comme l'âme de Descartes à Dieu. Ici, la conscience est considérée comme une *faculté fonctionnelle* sous-tendue par une *structure* adéquate, comme d'autres facultés, processus ou procédures de nos organes (locomotion, communication, reproduction, …). La question de l'attribution d'une conscience à des agents non-humains remplissant ces fonctions est ainsi nécessairement largement ouverte. Les formes de conscience humaine évoquées dans les chapitres précédents émergent de la structuration spécifique de notre système nerveux, en particulier du cortex cérébral. Nous pouvons examiner si ces structures neurones spécifiques sont aussi présentes chez d'autres animaux (voir Encart 1). Nous pouvons aussi définir des caractéristiques fonctionnelles à la conscience (voir Encart 2) et rechercher si elles s'expriment chez d'autres agents naturels ou artificiels. La conscience émergerait alors de substrats qui seraient différents, mais possédant, pour les agents naturels, des principes d'organisation équivalents qui restent à découvrir, et pour les agents artificiels qui auront été bio-inspirés à partir de l'humain.

Si elle aboutit, l'émergence d'une conscience conférée à un système artificiel s'inscrirait dans la continuité de toutes sortes d'innovations technologiques. Ainsi, la représentation artistique du monde (art pariétal, peinture, sculpture) réside dans l'utilisation de techniques spécifiques qui reproduisent certaines de nos capacités (percevoir une scène, la mémoriser). La photographie s'est appropriée encore plus efficacement ces fonctions biologiques : l'appareil photo focalise l'information lumineuse par ses lentilles, comme l'œil des vertébrés avec sa cornée et son cristallin ; il possède une matrice sensible (le film argentique ou le capteur numérique), comme son modèle naturel la rétine, capable de transcoder l'énergie lumineuse en trace chimique ou signal électrique ; l'image photographique garde la trace d'un motif spatial, comme l'hippocampe cérébral, support de la mémoire. De même, la force motrice musculaire a trouvé un équivalent fonctionnel dans les machines thermiques puis les moteurs électriques, sans employer ni protéines ni calcium. Dans certains contextes d'optimisation pour la résolution de problèmes, les algorithmes génétiques utilisés en informatique se sont inspirés des principes darwiniens de l'évolution biologique. Et enfin, les réseaux de neurones artificiels sont, par construction, bio-inspirés, (voir Encart 3). A l'instar de l'aéronautique qui produit la fonction « voler » sans battement des ailes, se conçoit ainsi une conscience fonctionnellement équivalente à celle de notre organisme qui, sans nécessairement imiter directement l'implémentation neuronale biologique, ferait usage de dispositifs et de mécanismes originaux.

L'histoire naturelle est jalonnée d'épisodes semblables qui ont précédé ces « innovations » de l'humanité. Par convergence évolutive, des fonctions très équivalentes ont émergé d'organes construits de tissus distincts, en différents points du buisson de l'évolution : le vol des insectes et celui des oiseaux, le marquage des lieux par les fourmis et par les chiens, et bien sûr l'œil caméral des vertébrés et celui des céphalopodes comme le poulpe. Des facultés fonctionnelles équivalentes peuvent donc être produites par des systèmes différant non seulement par la matière dont ils sont construits, mais aussi par l'architecture qui unit leurs composants. L'apparition d'autres consciences que la nôtre est-elle ainsi concevable ?

Encart 1 : **Prérequis structurels**. Les primates non humains, et même l'ensemble des mammifères possèdent un néocortex de structure lamellaire en six couches caractéristiques. C'est la dynamique de circulation d'influx dans ce cortex qui est le support d'une théorie dominante de la conscience humaine. En comparant la dynamique neuronale de personnes humaines déclarées conscientes ou inconscientes, Tononi en identifie des marqueurs physiologiques (voir Chapitre XX Les Théories). Dans une telle perspective, un cerveau de mammifère éveillé peut sans doute répondre par « oui » à la fatidique question : « Esprit (ou plutôt, Conscience), es-tu là » ? On sauterait ainsi le fossé creusé par Descartes (Descartes, 1637) au nom de la probabilité et de la vraisemblance : « elles [les bêtes] auraient une âme immortelle aussi bien que nous ; ce qui n'est pas vraisemblable, à cause qu'il n'y a point de raison pour le croire de quelques animaux, sans le croire de tous, et qu'il y en a plusieurs trop imparfaits pour pouvoir croire cela d'eux, comme sont les huîtres, les éponges, etc … »[3].

Approche fonctionnelle

Selon cette démarche, une première étape consiste à définir les propriétés d'une conscience fonctionnelle, pour pouvoir examiner ensuite dans quelle mesure les machines ou les animaux peuvent en témoigner. Une proposition est présentée dans l'Encart 2. Elle compte huit critères de traitement de l'information, de la perception à l'introspection. Remarquons que cette proposition est non seulement anthropomorphique (forcément), mais plus spécifiquement, ontologiquement et épistémologiquement, définie par des auteurs dont la pensée est marquée par le maniement d'outils informatiques et de concepts cybernétiques. Ainsi Seth et Tsakaris convoquent la « théorie du contrôle », de « l'énergie libre » et du « codage prédictif », pour montrer que l'expérience même du soi, se comprend en termes de régulation prédictive et d'inférences introspectives. Là où Descartes voyait en l'animal-machine, une impossibilité à l'émergence de « l'âme », Seth et Tsakaris retournent le problème considérant que c'est cette nature même de traiteur d'information, de « beast machine », qui conduit à l'émergence de sa propre individualité (Seth and Tsakiris, 2018), un soi conscient. Ainsi, les prérequis fonctionnels que nous retenons pour l'attribution d'une conscience (encart 2) seront sans doute retrouvés chez nos machines intelligentes plutôt que chez nos animaux de compagnie.

Cependant, les attributs de l'âme ou de la conscience qui sont avancés dans d'autres environnements culturels, témoignent d'autres préoccupations et empruntent d'autres formulations. Les amis des animaux seront plus sensibles à d'autres dimensions, comme l'affect, privilégié aussi par Descartes[4]. Celui-ci propose six « passions (primitives) de l'âme » : l'admiration, l'amour, la haine, le désir, la joie et la tristesse (auxquelles on peut rajouter les « actions de l'âme »). Plus récemment, Birch et al. (Birch et al., 2020) proposent cinq dimensions pour évaluer les états de conscience chez les animaux : la richesse perceptuelle de l'environnement (*P-richness*), qui qualifie la finesse de la discrimination ; la richesse de l'expérience affective (*Evaluative-richness*), la capacité

[3] Dans la lettre à Morus du 5 février 1649, Descartes dit aussi qu'il est plus probable de concevoir les vers de terre, les moucherons ou les chenilles comme des machines que de leur attribuer une âme immortelle (AT, V, 277).

[4] Descartes n'a pas fait l'erreur de négliger l'importance des émotions dans la prise de décision humaine que lui attribue faussement Damasio (Damasio, 1992).

d'intégration de l'information ponctuelle (*Integration at a time*) et d'intégration au cours du temps (*Integration across time*) et finalement la conscience de soi (*self-consciousness*).

Encart 2 : **Prérequis fonctionnels**. A l'aune de la cybernétique du siècle précédent, avançons quelques fonctionnalités essentielles pour un agent « conscient ».

1. **Percevoir** et mettre l'information à disposition de l'attention ;
2. **Gérer des objectifs :** avec des priorités -les plus essentielles étant de survivre et se reproduire - qui guident le choix des actions ;
3. **Focaliser l'attention** pour attribuer les ressources cognitives en temps réel ;
4. **Analyser et mettre en relation** les informations perçues ; ce qui peut se traduire sous forme d'émotions (Scherer, 1999) ;
5. **Planifier les actions et prendre des décisions** : en fonction de l'état interne (notamment émotionnel), des objectifs, des perceptions analysées (incluant les dimensions spatiales et temporelles), et des plans d'actions possibles et de leurs effets attendus ;
6. **Agir sur le monde et communiquer** (dans le cas des êtres sociaux) ;
7. **Apprendre** pour s'adapter et mieux agir dans le futur ;
8. **Introspection** : pour apprendre de son passé, analyser ses propres comportements, comprendre le comportement des autres comme une variante de son propre fonctionnement, se décentrer et ainsi mieux appréhender sa propre place dans le monde (voir Chapitre XX Métaconnaissance).

Autres consciences naturelles

Il semble que les sept premiers points fonctionnels soient satisfaits par les animaux en général. Des capacités cognitives de haut niveau et de conscience phénoménale sont observables chez les oiseaux, notamment les perroquets et les corvidés (Gunturkun and Bugnyar, 2016; Nieder et al., 2020). Par exemple, un geai bleu « se » voyant dans un miroir tente d'essuyer une tache sous son bec comme il le fait de celle qu'il aurait détectée directement sur son aile. Plus encore, alors que la structure du système nerveux du poulpe est radicalement différente de la nôtre (Young, 1971), des études approfondies démontrent des capacités cognitives similaires à celles des mammifères (Schnell et al., 2021). Ayant sauté le premier fossé qui nous séparait des autres animaux, puis parcouru le chemin qui nous éloignait des non-mammifères, passerait-on la barrière qui nous sépare des invertébrés, démunis d'un système nerveux central ? C'est en partie ce qu'a fait le législateur qui, en France, impose comme pour les vertébrés, des précautions particulières dans les expérimentations infligées aux céphalopodes en général[5]. Les autres invertébrés sont laissés sur le bord du chemin, malgré les découvertes de longue date des comportementalistes animaliers. Ainsi Turner au début du siècle dernier décrit comment une fourmi, isolée sur une île dans une flaque d'eau, organise la construction d'un pont en assemblant des objets de matériaux différents (Galpayage Dona and -Chittka, 2020). T. Seeley montre que chez les abeilles la prise de décision collective, par exemple la sélection du meilleur site pour la colonie lors de l'essaimage, émerge de la confrontation de différentes propositions individuelles sur lesquelles les habitantes vont se positionner (Seeley, 2011).

[5] Directive 201063/UE du parlement européen et du conseil du 22 sept 2010 relative à la protection ces animaux utilisés à des fins scientifiques.

Le modèle mathématique proposé par Tononi appliqué au système nerveux du poulpe, permettrait peut-être de comparer chez cette espèce, un individu éveillé et un individu anesthésié. Les critères retenus (voir Chapitre XXX Théories) pour reconnaitre un état conscient sont basés sur une évaluation de l'entropie (degré d'organisation) dans la configuration des connexions actives entre les différentes parties du système nerveux central humain. Chez ce mollusque aquatique, l'absence d'hémoglobine (nécessaire pour l'effet BOLD de l'IRMf), entre autre, limite pour le moment l'usage des techniques d'investigations par imagerie pour mettre en évidence les réseaux nerveux actifs dans ces deux états. Que dire alors du système nerveux d'une fourmi ?

Autres consciences artificielles

Aujourd'hui, nul ne peut ignorer les capacités cognitives des ordinateurs, en particulier depuis la montée en puissance des algorithmes d'apprentissage machine. Ceux-ci les conduisent à exceller dans des jeux de stratégie complexe, comme le jeu de go (Silver et al., 2017). Ils font même preuve d'intuition pour adapter dynamiquement leur stratégie en fonction des informations incomplètes disponibles, typique du jeu de poker sans limite (Moravčík et al., 2017). Ces algorithmes aux performances impressionnantes sont *bio-inspirés* (voir Encart 3). Réciproquement, les théories neuroscientifiques actuelles de la conscience humaine s'appuient sur des modèles computationnels du traitement de l'information ; et dans ce contexte hybride l'émergence de la conscience chez la machine, i.e. l'ordinateur, ne pose alors pas de problème conceptuel insurmontable (Dehaene et al., 2017). L'informatique se présente ainsi comme la technologie la plus à même de servir de base à une conscience artificielle puisque que le traitement de l'information inspiré du cerveau humain en est le principe originel.

Cependant, on perçoit bien qu'il y a là, entre conscience artificielle et naturelle, un mécanisme d'auto-référence, à la manière des boucles étranges de Hofstadter (Hofstadter, 1979), où un système ne se définit qu'en faisant explicitement référence à lui-même (les exemples de Hofstadter sont les dessins en boucle de Escher, les suites de Bach ou la logique mathématique qui conduit au théorème d'incomplétude de Gödel). En effet, dès les premiers travaux sur la théorie de l'information et la naissance de l'informatique, les scientifiques ont fait le rapprochement entre cerveau et machine de Von Neuman. Ainsi, le père de l'architecture actuelle des ordinateurs écrit un livre publié à titre posthume « Brain & Machine » (Von Neuman, 1958), et N. Wiener, père de la cybernétique et source d'inspiration du behaviourisme, publie « Nerve, Brain and Memory models» (Wiener and Shadé, 1963). Les pères fondateurs de l'Intelligence Artificielle J. McCarthy, M. Minsky et A. Newell, s'intéressent à rendre les machines conscientes de leur états mentaux (McCarthy, 1995), s'interrogent sur leur libre arbitre (McCarthy, 2000), sur l'émergence de l'esprit via les interactions entre agents rationnels (Minsky, 1986) ou proposent un modèle computationnel de la cognition (Newell, 1994).

Les principales théories neuroscientifiques actuelles sur la conscience (voir Chapitre XXX les Théories ») sont issus en droite ligne de ces travaux (Newman et al., 1997). La théorie de l'espace neuronal global (« Neuronal Global Workspace ») considère la conscience comme une forme de traitement global de l'information disponible dans un espace commun accessible à l'ensemble des processus mentaux (Baars, 2002; Dehaene et al., 2014; Mashour et al., 2020). Il n'y a donc pas, par principe, d'empêchement à ce qu'une machine soit consciente (Dehaene et al., 2017). D'autant que cette théorie se base directement sur des travaux menés en intelligence artificielle qui considéraient que la cognition émergeait de processeurs (ou agents) distribués qui écrivaient sur un tableau (« *blackboard* ») visible par l'ensemble des agents (Minsky, 1986; Newell, 1994). En

pratique, la réalisation de telles machines n'était pas sans poser des problèmes de contrôle pour favoriser l'accès à cet espace global à certains agents en fonction du contexte et permettre la prise de décision finale en temps réel (Hayes-Roth, 1982). A noter que la prise de décision mise en place chez les abeilles pour la gestion de situations complexes n'a jusqu'ici pas été simulée informatiquement[6] (Seeley, 2011). Pour la théorie de l'information intégrée (« Integrated Information Theory ») (Tononi, 2004), de faibles valeurs de Φ, mesure de la conscience phénoménale, peuvent être produites dans des systèmes beaucoup plus simples que le cerveau humain, pour autant qu'ils contiennent des unités logiques en interaction réciproque. Cette condition est certainement remplie par les systèmes nerveux de nombreux animaux, mais aussi par des cellules vivantes isolées, et même à un degré bien moindre par des circuits électroniques simples.

Nos prérequis fonctionnels (voir encart 2) sont en bonne partie remplis par les machines actuelles :

1. Percevoir : les algorithmes de reconnaissance d'images (détection de visages, de personnes, d'objets) et de son (reconnaissance vocale ou musicale) ont beaucoup progressé notamment depuis l'introduction de l'apprentissage profond (voir encart 3) et l'amélioration des capteurs (LIDAR, caméra haute définition, microphones en réseau …) ;
2. Gérer des buts : en robotique ils sont ordonnés par un score attribué dynamiquement de façon à élaborer un plan d'action contextuel (voir point 5) ;
3. Focaliser l'attention : les robots et machines actuels ont généralement un champ d'application restreint qui limite la complexité de l'attribution des ressources. On peut cependant considérer que les heuristiques suivies, par exemple par un programme de jeu d'échec pour choisir quelles actions il effectuera dans le temps imparti (à défaut d'une analyse exhaustive), sont une forme de sélection attentionnelle ;
4. Analyser : la modélisation des connaissances et du raisonnement, de type inductif ou déductif, a été beaucoup explorée en intelligence artificielle avec notamment l'introduction des ontologies et l'extension des logiques formelles (logiques temporelles, floues, …) et l'usage de langages adaptés, à base de règles et objets, de la programmation logique (langages Prolog, LISP), par contraintes, ou les langages de requêtes pour les données liées [SPARQL, Web sémantique] ;
5. Planifier les actions et prendre des décisions : peut être vue comme un problème d'optimisation. Tous les robots et systèmes autonomes prennent des décisions (agir ou pas), en fonction du contexte, des résultats attendus de leur(s) action(s) et du temps disponible ;
6. Agir sur le monde et communiquer : un agent artificiel dans un jeu vidéo ou un robot, peut exercer le même répertoire d'actions qu'un être humain : se déplacer, saisir des objets, tourner la tête, monter un escalier, émettre des sons ou exprimer une émotion ; il peut communiquer via une interface spécifique (écran, parole synthétique, etc.)
7. Apprendre : l'apprentissage est un domaine essentiel de l'intelligence artificielle depuis son origine. Il repose généralement sur l'effet observé d'une action face à une situation donnée de façon à pouvoir adapter sa réponse future dans une

[6] On estime à 1 million le nombre de neurones chez l'abeille versus 100 Milliards chez l'homme. Une colonie contient environ 40000 individus. La prise de décision individuelle humaine semble donc être encore plus difficile à modéliser.

situation similaire. Actuellement, les capacités de généralisation et d'adaptabilité sont cependant réduites.

8. Introspection : l'introspection et la conscience de soi sont encore peu explorées en pratique, bien que ces capacités soient essentielles pour qu'une machine soit reconnue comme consciente. Nous pouvons noter qu'actuellement les machines réalisent des tâches à l'insu de leur utilisateur (mise à jour, interrogation de serveur) qui imposent la reconnaissance automatique de l'état courant et la nécessité de le faire évoluer.

Si l'on prend un point de vue fonctionnaliste sur la conscience, on peut écarter certaines objections comme l'incapacité d'une machine à être autre chose qu'une imitation de la conscience humaine, ce qui est indémontrable, de la même façon qu'on ne peut pas démontrer que tout individu, autre que soi, n'est pas une simple imitation mécanique de notre conscience. Dans cette acception, la conscience a un rôle, une utilité justifiée par son émergence et sa sélection au cours de l'évolution des organismes biologiques. Certains critères, notamment le huitième (introspection), s'ils sont définis clairement, représentent un défi à relever pour les concepteurs de machine intelligente.

Cependant, il peut être aisé de demander à quelqu'un de définir un quelconque objet, puis à un artisan ou un ingénieur d'en modeler les formes, les couleurs, les densités, la mécanique… répondant aux critères et attributs mentionnés. Il y a fort à parier qu'au vu du résultat, même si rationnellement le prototype satisfait parfaitement à la première définition, (d'une certaine manière le premier cahier des charges à l'ingénieur), même pour celui qui l'aura rédigée, « ceci n'est pas une pipe » ! Une fois les premiers critères satisfaits, de nouveaux critères s'imposeront. Dans une logique d'implémentation informatique, les critiques du nouveau résultat obtenu peuvent être prises en compte de façon réitérée pour une version beta, puis beta+ … Le terme « intelligence artificielle » a longtemps été ridiculisé. Ainsi, nous avons connu les débats autour de la capacité à jouer aux échecs, à identifier des visages, à reconnaître une langue ou un texte, de la capacité à le traduire (Searle, 1983), à imaginer de nouvelles formes, etc. Les gardiens du temple rassuraient le peuple sur la supériorité de l'intelligence humaine en effet longtemps inégalée. Selon Chapouthier et Kaplan (Chapouthier and Kaplan, 2011), la course est sans fin : dès lors qu'un critère est satisfait pour montrer le défaut de la machine sur une qualité humaine, après un « oui mais… », un nouveau défi peut être lancé. Après des années de défi au jeu de go, qui nécessite une forme d'intuition au-delà du raisonnement pur, la machine s'est imposée face au plus grand maître (voir le match AlphaGo vs Lee Sedol en 2016, (Silver et al., 20)). Ainsi, force est de constater qu'en 2021, l'intelligence artificielle est largement reconnue comme une intelligence, pour part équivalente, pour part supérieure même à celle des individus humains. A quel moment pourrons-nous accepter de même l'existence d'une conscience artificielle, même différente de la conscience humaine ? Certains maintiendront qu'il lui manquera toujours le libre arbitre.

Une conscience artificielle suffisamment complexe, influencée par de nombreux facteurs (ses capacités, ses objectifs, ses émotions, sa mémoire, etc.) et évoluant continuellement par apprentissage aura un comportement complexe tantôt prévisible, parfois inattendu, tout en restant parfaitement déterministe si l'on connaissait tous les facteurs impliqués. De même, la psychologie sociale, la sociologie, l'anthropologie démontrent que notre comportement est influencé par les normes sociales, le milieu socio-économique fréquenté, les expériences de vie. Comme on ne peut pas connaître l'ensemble des facteurs influençant un individu (génétiques, épigénétiques, mécanismes inconscients, diététiques, parcours de vie, état émotionnel, motivations, normes

personnelles et sociales, capacité à se projeter dans le futur, etc. …), il est difficile de prédire son comportement, bien que des données statistiques existent sur la corrélation entre ces facteurs et le comportement individuel. On peut donc défendre l'idée que le libre arbitre, souvent associé exclusivement à la conscience humaine, est une illusion qui cache un comportement déterministe dont les facteurs d'influence et les mécanismes inconscients nous échappent (Bignetti, 2014; Soon et al., 2008). Être libre de prendre des décisions impose qu'elles soient prises par soi-même et adaptées à une situation dynamique. Ce que nous apprend l'étude des insectes c'est que cela n'implique pas d'en être nécessairement conscient (Heisenberg, 2009), à moins d'accorder une conscience à un *groupe* d'insectes. Ainsi, on peut construire un cadre formel pour introduire le « libre arbitre même pour les robots » (McCarthy, 2000).

Encart 3 : **Réseaux de neurones**

Le principe remonte aux années soixante du siècle dernier, et est issu du courant connexionniste de la nouvelle discipline de l'époque, l'Intelligence Artificielle. Il s'agit de mimer le fonctionnement cérébral en regroupant des neurones formels. Un neurone formel est une entité simple dont la sortie dépend d'une fonction (échelon, sigmoïde, …) et de la pondération de ses entrées. Un réseau de neurones est constitué de couches successives dont les sorties sont connectées entre elles. Lors d'une phase dite « d'apprentissage supervisé », des exemples (e.g. des images d'insectes et d'arachnides) sont présentés à l'entrée du système. Une fonction d'erreur est estimée pour comparer la sortie obtenue (e.g. insecte vs arachnide) avec celle attendue (e.g. arachnide). Les poids associés à chaque couche de neurones sont alors adaptés pour réduire la fonction d'erreur. Une phase dite de « Validation » permet sur quelques exemples de raffiner le cas échéant les poids. Finalement, la phase dite de « test » permet d'évaluer les performances du réseau sur un nouveau jeu d'exemples. Un système à une couche, tel le perceptron (Rosenblatt, 1958), permet de séparer linéairement deux classes. En pratique, la classification d'images n'est pas une opération linéaire et plusieurs couches (on parle de couches cachées, d'où l'appellation de *deep-learning*, pour les couches situées entre l'entrée et la sortie) sont donc nécessaires. Dès leur introduction, les réseaux multicouches s'avéraient difficiles à programmer et cet axe de recherche a été rapidement abandonné (Minsky and Papert, 1969). Il est revenu en force au début du 21ème siècle grâce à la conjonction de trois facteurs majeurs : la puissance accrue des ordinateurs, la disponibilité de grande masse de données pour disposer de beaucoup d'exemples pour estimer un modèle généralisable et une méthode efficace d'ajustement des poids lors de l'apprentissage (méthode dite de back-propagation, (Rumelhart et al., 1986)). A noter que plutôt que d'associer à chaque neurone artificiel un pixel de l'image, ce qui est très computationnellement coûteux, Y. Le Cun et al. ont proposé, d'associer une information locale obtenue dans un voisinage (Le Cun et al., 1989). Ils utilisent pour cela des filtres convolutifs variables à chaque couche (CNN : Convolutional Neural Network). C'est une approche bio-inspirée du système visuel des mammifères issus des travaux de Fukushima sur le NéoCognitron (Fukushima, 1980).

En conclusion, la question n'est sans doute pas de savoir si une machine peut être consciente (« no more interesting that the question of whether a submarine can swim » E. Dijkstra) mais plutôt d'explorer si la machine est un bon modèle pour nous aider à comprendre les conditions d'émergence d'états de conscience ; quelles sortes de systèmes sont conscients et d'autres non conscients et pourquoi ; et en quoi ces états de conscience favorisent ou nuisent à la résolution de certaines tâches complexes et

participent de la définition de l'individualité des entités vivantes. L'accentuation des efforts pour comprendre la conscience (Michel et al., 2019) et l'étude des architectures informatiques implémentant des modèles de traitement sensori-moteur et cognitif (Schrimpf et al., 2018), testables sur des jeux de données partagées, sont les clés pour affiner nos théories actuelles sur la spécificité de l'*Homo Sapiens*.

Contributions

Les trois auteurs ont élaboré ce chapitre collectivement.
Les autres membres du collectif CARMEN ont relu ce chapitre et suggéré quelques modifications.

Références

Baars, B., 2002. The conscious access hypothesis: origins and recent evidence. Trends Cognitive Sciences 6, 47-52.
Bernard, C., 1865. Introduction à l'étude la médecine expérimentale. Flammarion, Paris , Fr (ed. 1984).
Bignetti, E., 2014. The functional role of free-will illusion in cognition: "The Bignetti Model". Cognitive Systems Research 31-32, 45-60.
Birch, J., Schnell, A.K., Clayton, N.S., 2020. Dimensions of Animal Consciousness. Trends Cogn Sci 24, 789-801.
Chalmers, D.J., 2010. The Singularity: A Philosophical Analysis. Journal of Consciousness Studies
17, 7–65.
Chapouthier, G., Kaplan, F., 2011. L'homme, l'animal et la machine: perpétuelles redéfinitions CNRS editions Paris, Fr.
Darwin, C., 1859. L'origine des espèces. Flammarion Paris, Fr (1992).
Dehaene, S., Charles, L., King, J.R., Marti, S., 2014. Toward a computational theory of conscious processing. Curr Opin Neurobiol 25, 76-84.
Dehaene, S., Lau, H., Kouider, S., 2017. What is consciousness, and could machines have it? Science 358, 486-492.
Descartes, R., 1637. Discours de la méthode. Vrin Paris, Fr (1984).
Diderot, D., (Le Rond) D'Alembert, J., 1751-72. L'Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers.
Fukushima, K., 1980. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36, 193–202.
Galpayage Dona, H., -Chittka, L., 2020. Charles H. Turner, pioneer in animal cognition. Science 370, 530-531.
Good, I.J., 1965. Speculations Concerning the First Ultraintelligent Machine. In: Alt, F.L., Rubinoff, M. (Eds.), Advances in Computers. Academic Press
, New York City, pp. 31-88.
Gunturkun, O., Bugnyar, T., 2016. Cognition without Cortex. Trends Cogn Sci 20, 291-303.
Hayes-Roth, B., 1982. A blackboard architecture for control. Artificial Intellligence 26, 251-321.
Heisenberg, M., 2009. Is free will an illusion. Nature 459, 164-165.
Hofstadter, D., 1979. Gödel, Escher, Bach: an Eternal Golden Braid. Basic Books.

Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1989. Handwritten Digit Recognition with a Back-Propagation Network NIPS, pp. 396-404.
Mashour, G.A., Roelfsema, P., Changeux, J.P., Dehaene, S., 2020. Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron 105, 776-798.
McCarthy, J., 1995. Making robots conscious of their mental states. MachineIntelligence 15, Oxford Univ.
McCarthy, J., 2000. Free will - even for robots. Stanford Univ.
Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., Carrasco, M., Chirimuuta, M., Chun, M., Cleeremans, A., Dehaene, S., Fleming, S.M., Frith, C., Haggard, P., He, B.J., Heyes, C., Goodale, M.A., Irvine, L., Kawato, M., Kentridge, R., King, J.R., Knight, R.T., Kouider, S., Lamme, V., Lamy, D., Lau, H., Laureys, S., LeDoux, J., Lin, Y.T., Liu, K., Macknik, S.L., Martinez-Conde, S., Mashour, G.A., Melloni, L., Miracchi, L., Mylopoulos, M., Naccache, L., Owen, A.M., Passingham, R.E., Pessoa, L., Peters, M.A.K., Rahnev, D., Ro, T., Rosenthal, D., Sasaki, Y., Sergent, C., Solovey, G., Schiff, N.D., Seth, A., Tallon-Baudry, C., Tamietto, M., Tong, F., van Gaal, S., Vlassova, A., Watanabe, T., Weisberg, J., Yan, K., Yoshida, M., 2019. Opportunities and challenges for a maturing science of consciousness. Nat Hum Behav 3, 104-107.
Minsky, M., 1986. The Society of Mind. Simon & Schuster, New York.
Minsky, M., Papert, S., 1969. Perceptrons: An Introduction to Computational Geometry. The MIT Press, Cambridge, MA.
Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., Bowling, M., 2017. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. Science 356, 508-513.
Newell, A., 1994. Unified theory of cognition. Harvard University Press.
Newman, J., Baars, B., Cho, S.-B., 1997. A Neural Global Workspace Model for Conscious Attention. Neural Netw 10, 1195-1205.
Nieder, A., Wagener, L., Rinnert, P., 2020. <Corvidé.pdf>. Science 369, 1626-1629.
Perrault, C., 1680. La Méchanique des animaux. Coignard, Paris (Fr).
Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65, 386-408.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533-536.
Scherer, K.R., 1999. Appraisal theory. Handbook of cognition and emotion. John Wiley & Sons, pp. 637–663.
Schnell, A.K., Amodio, P., Boeckle, M., Clayton, N.S., 2021. How intelligent is a cephalopod? Lessons from comparative cognition. Biol Rev Camb Philos Soc 96, 162-178.
Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D.L.K., DiCarlo, J.J., 2018.
Searle, J., 1983. Minds, brains and programs. In: Hofstadter, D.R., Dennett, D.C. (Eds.), The Mind's I: fantasies and reflections on self and soul. Penguin Books, Hardmonsworth, pp. 353-373.
Seeley, T.D., 2011. Honeybee democracy. Princeton University Press.
Seth, A.K., Tsakiris, M., 2018. Being a Beast Machine: The Somatic Basis of Selfhood. Trends Cogn Sci 22, 969-981.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of Go without human knowledge. Nature 550, 354-359.

Singer, P., 1975. Animal Liberation. HarperCollins, NY, USA.
Soon, C.S., Brass, M., Heinze, H.J., Haynes, J.D., 2008. Unconscious determinants of free decisions in the human brain. Nat Neurosci 11, 543-545.
Tononi, G., 2004. An information integration theory of consciousness. BMC Neurosci 5, 42.
Von Neuman, J., 1958. The computer and the Brain. Yale University Press, New Haven.
Wiener, N., Shadé, J., 1963. Nerve, Brain and Memory Models. Elsevier, New York.
Young, J.Z., 1971. Anatomy of the nervous system of Octopus vulgaris. . Clarendon Press, Oxford (UK).