Intelligence Artificielle

Russell & Norvig

3ème édition, 2010, éd. Pearson (1199 pages).

Chapitre 26 Fondements philosophiques

Où *l'on considère ce que penser veut dire, et où l'on se demande si des artefacts peuvent penser et même s'ils doivent le faire.*

Plan du Chapitre

26.1 lA faible: les machines peuvent-elles agir intelligemment?

26.2 lA forte: les machines peuvent -elles réellement penser?

26.3 Éthique et risques du développement de l'intelligence artificielle. (voir Texte 3)

Résumé

Ce chapitre traite des sujets suivants:

* Les philosophes emploient les termes d'lA **faible** pour désigner l'hypothèse selon laquelle les machines pourraient se comporter intelligemment, et d'lA **forte** pour celle qui énonce que de telles machines auraient des esprits réels (et non des esprits simulés).

* Alan Turing a récusé la question « Les machines peuvent-elles penser ? » et l'a remplacée par un test comportemental. Il a anticipé de nombreuses objections à la possibilité de machines pensantes. Peu de chercheurs en lA prêtent attention au test de Turing: ils préfèrent se concentrer sur les performances de leurs systèmes appliqués à des tâches pratiques plutôt que sur leur capacité à imiter les humains.

* À notre époque, on s'accorde généralement à penser que les états mentaux sont des états du cerveau.

* Les arguments pour et contre l'lA forte ne sont pas concluants. Dans le courant dominant, les chercheurs sont peu nombreux à penser que quoi que ce soit de significatif dépende du résultat du débat.

* La conscience demeure un mystère.

* Nous avons identifié six menaces potentielles que **l'lA** et les technologies associées constituent pour la société. Nous avons conclu que certaines d'entre elles sont improbables ou diffèrent peu de celles qui sont associées à des technologies «non intelligentes». L'une en particulier mérite qu'on s'y attarde: des machines ultra-intelligentes pourraient nous entraîner vers un futur très différent du présent que nous connaissons et celui-ci n'est pas nécessairement enviable...

[Fin du résumé]

Introduction

Les philosophes, qui existaient bien avant les ordinateurs, essaient depuis longtemps de résoudre des problèmes qui touchent à l'intelligence artificielle. Comment l'esprit fonctionne-t-il ? Est-il possible que des machines agissent intelligemment à la façon des humains? Dans l'affirmative, auraient -elles un esprit, réel et conscient? Quelles sont les implications éthiques de l'existence de machines intelligentes?

Posons tout d'abord la terminologie. L'assertion selon laquelle les machines pourraient *agir comme si* elles étaient intelligentes est qualifiée d'hypothèse IA **faible** par les philosophes, et celle selon laquelle les machines qui se comportent ainsi sont douées d'une pensée *réelle (*et non simplement *simulée)* s'appelle l'hypothèse IA forte. La plupart des chercheurs en lA tiennent l'hypothèse faible pour acquise et se soucient fort peu de l'hypothèse forte: tant que leurs programmes fonctionnent, peu importe qu'il s'agisse d'une simulation ou d'une réalité. Mais tous les chercheurs en lA devraient se soucier des implications éthiques de leur travail.

26.1 lA faible: les machines peuvent-elles agir intelligemment?

Le projet de programme de l'atelier d'été de 1956 qui définissait le champ de l'intelligence artificielle (Mcarthy *et al.,* 1955) énonçait que «Chaque aspect de l'apprentissage ou de toute autre caractéristique de l'intelligence peut en principe être si précisément décrit qu'il est de construire une machine pour le simuler.» Dans cette optique, l'lA faible était considéré comme possible. D'autres ont affirmé que l'lA faible était impossible: « L'intelligence artificielle *poursuivie dans le culte du computationalisme* n'a pas même l'ombre d'une chance d'obtenir des résultats durables.» (Sayre, 1993)

Bien évidemment que l'IA soit impossible ou non dépend de la façon dont on la définit. A la section 1.1, nous avons vu que c'était le meilleur programme agent dans une architecture donnée. Formulée ainsi, l'IA est par définition possible: pour toute architecture numérique comprenant **k bits de stockage** pour les programmes, il existe exactement 2^k **programmes agents**, et il suffit pour trouver le meilleur de les énumérer et de les tester tous. C'est peut-être infaisable si la valeur de k est importante, mais les philosophes s'occupent de théorie, non de pratique.

Page 1076

Notre définition de l'IA fonctionne bien pour le problème technique qui consiste à trouver le bon agent pour une architecture donnée. Nous sommes donc tentés de terminer là cette section en répondant à la question posée dans son titre par l'affirmative. Mais les philosophes s'intéressent à la comparaison de deux architectures - celle de l'humain et celle de la machine. De plus, ils l'ont toujours posée non en termes de **maximisation de l'utilité** attendue, mais sous la forme « **Les machines peuvent -elles penser ?** »

L'informaticien **Edsger Dijkstra** (1984) a dit que «La question de savoir si *les machines peuvent penser* ... avait à peu près autant de sens que celle de savoir si *les sous-marins peuvent nager».* Les dictionnaires définissent généralement *nager* comme «Se déplacer dans l'eau en bougeant les membres ou les nageoires ». et la plupart des gens conviennent qu'un sous- marin, étant dépourvu de membres, ne peut nager. Les dictionnaires définissent également *voler* comme « Se déplacer dans l'air au moyen d'ailes ou de parties équivalentes à des ailes» et la plupart des gens pensent que les avions, étant dotés de telles parties, peuvent voler. Toutefois, aucune de ces deux questions ni de leurs réponses n'a grand-chose à voir avec la conception des avions ou des sous-marins, mais concerne en revanche l'usage des mots dans notre langue. (Le fait que les sous-marins nagent *vraiment* en russe ne fait qu'amplifier ce point.) La possibilité pratique de «machines pensantes» nous accompagne depuis une cinquantaine d'années, ce qui est trop court pour que

des locuteurs s'accordent sur le sens du mot« penser» - faut-il pour cela un« cerveau» ou simplement des «parties équivalentes à un cerveau» ?

Dans le célèbre article « **Computing Machinery and Intelligence**» (1950), **Alan Turing** suggérait qu'au lieu de demander si les machines peuvent penser, on devrait se demander si elles peuvent réussir un test d'intelligence comportementale - que l'on nomme depuis test de Turing. Pour tester un programme, on le met en situation de converser pendant cinq minutes *(via* des messages saisis en ligne) avec un interrogateur. Celui-ci doit alors deviner si la conversation a eu lieu avec un programme ou avec une personne; le programme réussit le test s'il trompe l'interrogateur 30 % du temps. Selon les conjectures de Turing, il aurait dû être possible avant l'an 2000 de programmer suffisamment bien un ordinateur doté de 10^9 unités de mémoire pour qu'il réussisse le test. Il avait tort: les programmes ne parviennent pas encore à tromper un juge averti.

En revanche, certaines personnes ont *bien* été bernées, ignorant qu'elles pouvaient bel et bien converser avec un ordinateur. Le programme ELIZA et les chatbots Internet comme MGONZ (Humphrys, 2008) et NATACHATA ont abusé leurs correspondants à maintes reprises; le chatbot CVBERLOVER a attiré l'attention des autorités en raison de son penchant à amener les internautes à divulguer suffisamment d'informations personnelles pour que leur identité puisse être usurpée. Le concours pour le prix Lœbner, qui se tient chaque année depuis 1991, est le plus ancien mettant en œuvre un test du type de celui de Turing, et il a permis de mieux modéliser les erreurs de frappe humaines.

Turing lui-même examina de nombreuses objections possibles quant à l'existence de machines intelligentes, ce qui comprenait virtuellement toutes celles qui ont été soulevées durant le demi-siècle qui suivit la publication de son article. Nous allons en étudier quelques- unes.

Page 1077

26.1.1 L'argument de l'incapacité

L'«argument de l'incapacité» consiste à prétendre qu'«une machine ne pourra jamais faire X ». Parmi les exemples de X, **Turing** cite les suivants:

> « Être gentil, inventif, beau, amical; prendre des initiatives, avoir le sens de l'humour, discerner le bien du mal, commettre des erreurs, tomber amoureux, aimer 1es fraises à la crème, séduire quelqu'un, tirer des leçons de l'expérience, employer les mots correctement, être le sujet de sa propre pensée, avoir des comportements aussi diversifiés qu'un humain, faire quelque chose de réellement nouveau, »

Rétrospectivement, certains d'entre eux sont relativement faciles - nous sommes tous familiers des ordinateurs qui « commettent des erreurs» Nous connaissons bien également une technologie centenaire qui a prouvé son aptitude à • séduire quelqu'un»: l'ours en peluche. L'expert en échecs David Levy prédit que, d'ici 2050, les gens s'éprendront couramment de robots humanoïdes (Levy, 2007), Quant au robot qui tombe amoureux, c'est un thème courant en littérature [l], mais l'on a assez peu réfléchi à la question de savoir si c'était en fait probable (Kim *et al.,* 2007), Il existe des programmes qui jouent aux échecs , aux dames, d'autres jeux, inspectent des pièces sur des chaînes de montage, pilotent des voitures et des hélicoptères, diagnostiquent des maladies et exécutent des centaines de tâches aussi bien ou que les humains. Certains ordinateurs ont fait des découvertes – de petite envergure mais significatives - en astronomie, en mathématiques, en chimie, en minéralogie, en biologie, en informatique et dans d'autres domaines, Chacune d'elles exigeait 1e niveau d'un expert humain.

Étant donné ce que nous savons maintenant des ordinateurs, leur aptitude à résoudre les problèmes combinatoires tels que les échecs n'a rien de surprenant. Mais certains algorithmes s'acquittent également tout aussi bien de tâches qui nécessitent apparemment un jugement humain, ou, selon les termes de Turing, la capacité de «tirer des leçons de l'expérience » et de «discerner le vrai du faux». Dès 1955, Paul Meehl (voir Grove et Meehl, 1996) étudiait les processus de prise de décision d'experts entraînés à des tâches

subjectives telles que prédire le succès d'un étudiant à un programme de formation ou la probabilité de récidive d'un criminel. Dans 19 cas sur 20, Meehl découvrit que des algorithmes d'apprentissage simples (comme la régression linéaire ou le modèle de Bayes naïf) effectuaient de meilleurs prédictions que les experts. *L'Educational Testing Service* utilise depuis 1999 un programme automatisé pour noter des millions de questions ouvertes au GMAT [2]. Le programme est d'accord avec les examinateurs humains dans 97% des cas, à peu près comme entre deux examinateurs humains (Burstein *et al., 2001).*

Il est clair que les ordinateurs peuvent exécuter de nombreuses tâches aussi bien ou mieux que des humains,y compris celles dont nous pensons qu' elles nécessitent beaucoup d'intuition et de compréhension. Cela ne signifie pas bien sûr qu'il font appel pour ce faire à l'intuition ou à la compréhension - celles-ci ne font pas partie de 1eur comportement, et nous avons traité de ces questions ailleurs - mais il apparaît que la conception immédiate des processus mentaux nécessaires pour produire un comportement donné est souvent fausse.

Page 1078

Bien entendu, il est également vrai qu'il existe de nombreuses tâches pour lesquelles les ordinateurs n'excellent pas (pour employer un euphémisme), notamment celle du t**est de Turing: mener une conversation à bâtons rompus.**

26.1.2 L'objection mathématique

Depuis les travaux de **Turing** (1936) et de **Gödel** (1931), on sait que certains systèmes formels sont, en principe, incapables de répondre à certaines questions mathématiques. Le **théorème d'incomplétude** de **Gödel** (voir section 9.5) en constitue le plus fameux exemple. En bref, pour tout système axiomatique formel *F* capable de formaliser l'arithmétique, il est possible de construire une phrase de Gödel *G(F)* qui possède les propriétés suivantes:

- *G(F)* est une phrase de *F,* mais ne peut être prouvée dans *F.*
- Si F est consistant, alors *G(F)* est vraie.

Des philosophes, comme **J. R. Lucas** (1961) ont affirmé que ce théorème prouve que le machines sont mentalement inférieures aux humains parce que ce sont des systèmes formel limités par le théorème d'incomplétude - elles ne peuvent établir la vérité de leur propre phrase de **Gödel**- tandis que les humains ne souffrent pas de telles limites. Cette affirmation est à l'origine de dizaines d'années de controverses qui ont été commentées dans de nombreux textes, notamment deux ouvrages du mathématicien sir Roger **Penrose** (1989, 1994) qui la reprend en lui ajoutant quelques nouveautés (par exemple, l'hypothèse selon laquelle les *humains* seraient différents parce que le fonctionnement de leur cerveau reposerait sur la **gravité quantique**). Nous n'examinerons que trois des problèmes liés à cette affirmation.

Premièrement, le théorème d'incomplétude de **Gödel** ne s'applique qu'aux systèmes formels assez puissants pour faire de l'arithmétique. Ceux-ci comprennent les machines de Turing, et l'affirmation de Lucas se fonde en partie sur l'assertion selon laquelle les ordinateurs sont des machines de Turing. C'est là une bonne approximation, mais elle n'est que partiellement vraie. **Les machines de Turing sont infinies**, alors que les ordinateurs sont finis: on peut donc définir tout ordinateur comme un (très grand) système régi par la logique propositionnelle, laquelle n'est pas sujette au théorème d'incomplétude de Gödel.

Deuxièmement, un agent ne doit pas « avoir honte» de ne pas pouvoir établir la vérité d'une phrase alors que d'autres agents le peuvent. Considérez la phrase suivante:

J. R. Lucas ne peut pas toujours affirmer que cette phrase est vraie.

Si Lucas affirmait que cette phrase est vraie, il se contredirait: en conséquence, il ne peut pas toujours l'affirmer et elle doit donc être vraie. Nous avons ainsi démontré qu'il existe une phrase dont Lucas ne peut pas toujours assurer qu'elle est vraie alors que d'autres personne (et des machines) le peuvent. Mais cela ne diminue en rien notre estime pour Lucas. Prenons un autre exemple: aucun être humain ne pourrait calculer la somme d'un milliard de nombres de dix chiffres durant sa vie entière, mais un ordinateur en est capable en quelques seconde Pourtant, nous n'y voyons pas une limite fondamentale de la pensée humaine. Puisque l'humanité se comportait intelligemment des milliers d'années avant d'inventer les mathématique, il est improbable que le raisonnement mathématique joue plus qu'un rôle périphérique dans ce qu'être intelligent veut dire.

Troisièmement, et surtout, même si l'on convient que les ordinateurs ne peuvent prouver, rien ne permet d'affirmer que les humains ne sont pas pareillement limités. Il est bien trop facile de montrer rigoureusement qu'un système formel ne peut pas faire X. et d'affirmer que les humains le *peuvent* à leur propre manière informelle sans fournir au preuve de cette assertion. Certes, il est impossible de *prouver* que les humains ne sont pas

sujets au théorème d'incomplétude de **Gödel**. dans la mesure où toute preuve rigoureuse nécessiterait une formalisation du talent humain prétendument impossible à formaliser, se réfutant ainsi elle-même. Il nous reste donc à faire appel à l'intuition que les humains sont en quelque sorte capables de faits surhumains en matière d'intelligence mathématique. Cet appel se retrouve dans des arguments tels que «Nous devons supposer notre propre cohérence, faute de quoi aucune pensée n'est possible.. (Lucas, 1976). Or **les humains sont précisément connus pour leur incohérence**, Si cette propriété se manifeste dans le raisonnement de tous les jours, elle est également présente dans la pensée mathématique, si rigoureuse soit-elle. Un exemple célèbre est celui de la coloration d'une **carte en quatre couleurs**. En 1879, Alfred Kempe publiait une preuve qui fut largement acceptée et contribua à le faire élire *Fellow of the Royal Society.* Toutefois, en 1890, Percy Heawood signala une faille et la conjecture resta sans preuve jusqu'en 1977.

26.1.3 **L'argument de l'informalité**

L'une des objections les plus influentes et les plus tenaces au projet de l'lA a été envisagée, par Turing: c'est 1'« argument de l'informalité du comportement». En substance, le comportement humain serait beaucoup trop complexe pour qu'un ensemble de règles simples puissent en rendre compte, et, comme les ordinateurs ne savent rien faire d'autre que d'appliquer des ensembles de règles, ils ne pourraient pas générer de comportements aussi intelligents que ceux: des humains. L'incapacité de tout capturer dans un ensemble de règles logiques se nomme en lA **le problème de la qualification**.

Le principal avocat de cette optique a été le philosophe **Hubert Dreyfus**, qui a publié une série de critiques de l'intelligence artificielle souvent reprises: *What Computers Can't Do* (1972),la suite *What Computers Still Can't Do* (1992), et, *avec* son frère Stuart, *Mind Over Machine (1986).*

La position qu'ils critiquent est maintenant qualifiée de GOFAI *(Good Old-Fashioned AI, ou* «bonne vieille lA»), terme forgé par le philosophe John Haugeland (1985). GOFAI est censée prétendre que tout comportement intelligent peut être capturé par un système qui raisonne logiquement en partant d'un ensemble de faits et de règles qui décrivent un domaine. Elle correspond donc à l'agent logique le plus simple décrit au chapitre 7. Dreyfus a raison de dire que les agents logiques sont vulnérables au problème de la qualification. Comme nous l'avons vu au chapitre 13, les systèmes de raisonnement probabiliste sont mieux adaptés aux domaines «ouverts », La critique de Dreyfus n'est donc pas dirigée contre les ordinateurs *en soi* mais plutôt contre une programmation particulière. Il est toutefois raisonnable de supposer qu'un ouvrage intitulé *What First-Order Logical Rule-Based Systems Without Learning Can't Do*[3] aurait sans doute eu moins d'impact.

Du point de vue de Dreyfus, l'expertise humaine exige bien la connaissance de certaines règles, lesquelles ne constituent cependant qu'un « arrière- plan» ou un « contexte holistique » dans lequel nous opérons. Il donne l'exemple du comportement social qu'il est de bon ton d'adopter en matière d'échange de cadeaux:: «Normalement, on répond simplement dans les circonstances appropriées en offrant un cadeau équivalent.. Nous possédons apparemment «un sens immédiat de ce qu'il faut faire et de ce à quoi il convient de s 'attendre". Il en irait de même dans le contexte des échecs: «Un simple maître se demandera peut-être ce qu'il doit faire, mais un grand maître voit simplement que la configuration des pièces exige

un déplacement précis . la bonne réponse s'impose à son esprit.» Il ne fait aucun doute qu'une grande partie des processus mentaux du donateur ou du grand maître se déroulent à **un niveau de conscience qui échappe à l'introspection**, mais cela ne veut pas dire que ces *processus n'existent pas. La question importante à laquelle Dreyfus ne répond pas* est celle de savoir *comment le* bon *déplacement* s'impose à l'esprit du grand maître. *Cela rappelle le* commentaire de **Daniel Dennett** (1984) :

> « C'est un peu comme si les philosophes devaient s'autoproclamer experts en explication des méthodes des prestidigitateurs, puis, quand on leur demanderait comment le magicien exécute le truc de la femme coupée en deux, ils expliqueraient que c'est évident: le magicien ne la scie pas réellement en deux, il en donne simplement l'illusion: « Mais comment fait-il *cela?* . demandons-nous. « Ce n'est pas notre rayon », répondent les philosophes. »

Dreyfus et Dreyfus (1986) proposent un processus d'acquisition de l'expertise en cinq étapes, commençant par l'application de règles (du même type qu'en GOFAI) et se terminant par la capacité de sélectionner instantanément des réponses correctes. En émettant cette proposition, Dreyfus et Dreyfus se transforment, de fait, de critiques de l'IA en théoriciens de l'IA : ils préconisent une architecture de réseau connexionniste organisée en une vaste « bibliothèque de cas », tout en soulignant quelques problèmes. Heureusement, ces derniers ont tous été résolus, avec un succès partiel dans certains cas, total dans d'autres. Ces problèmes sont les suivants:

1. Il est impossible de bien généraliser à partir d'exemples sans connaître le contexte. Selon ces auteurs, personne n'a la moindre idée de la façon d'incorporer la connaissance du contexte au processus d'apprentissage d'un réseau connexionniste. En fait il existe des techniques qui permettent aux algorithmes d'apprentissage d'exploiter des connaissances *a priori* (voir chapitres 19 et 20). Toutefois, ces techniques reposent sur la disponibilité des connaissances sous forme explicite, ce que Dreyfus et Dreyfus nient énergiquement. Selon nous, c'est là une bonne raison de repenser sérieusement les modèles des réseaux connexionnistes actuels, afin qu'ils *puissent* tirer parti des connaissances acquises *a priori,* à la manière des autres algorithmes d'apprentissage.

2. L'apprentissage connexionniste est une forme d'apprentissage supervisé (voir chapitre 18) qui nécessite que soient identifiées au préalable des entrées pertinentes et des sorties correctes. En conséquence, prétendent-ils, il ne peut fonctionner de façon autonome sans l'aide d'un instructeur humain. En réalité, l'**apprentissage non supervisé** (voir chapitre 20) et l'**apprentissage par renforcement** (voir chapitre 21) en sont capables.

3. Les performances des algorithmes d'apprentissage ne sont pas bonnes quand les attributs sont nombreux et, si l'on en choisit un sous-ensemble, « il n'existe aucun moyen connu d'ajouter de nouveaux attributs si l'ensemble courant se montre incapable de rendre compte des faits appris ». En réalité, les nouvelles méthodes telles que les machines à vecteurs support gèrent très bien de grands ensembles d'attributs. Avec l'introduction de grands ensembles de données fondés sur le Web, de nombreuses applications dans des domaines tels que le traitement du langage (Sha et Pereira, 2003)et la vision artificielle (Viola et Jones, 2002a) gèrent en permanence des millions d'attributs. Nous avons vu au chapitre 19 qu'il est également possible de gérer méthodiquement de nouveaux attibuts, même s'il reste beaucoup à faire.

4- Le cerveau est capable d'orienter ses capteurs de façon à rechercher les informations pertinentes et de les traiter pour en extraire les aspects appropriés à la situation. Mais Dreyfus et Dreyfus soutiennent que «Actuellement, aucun détail de ce mécanisme n'est connu, ni ne fait même l'objet d'une hypothèse susceptible de guider la recherche en lA.» En fait, le domaine de la vision active, étayé par la théorie de la valeur de l'information (voir chapitre 16), s'intéresse précisément à ce problème de l'orientation des capteurs, et certains robots incorporent déjà les résultats théoriques obtenus. Le voyage de 132 miles de STANLEY dans le désert (voir page 30) a été rendu en grande partie possible par un système de capteurs actifs de ce type.

En somme, nombre des problèmes soulevés par Dreyfus - connaissance pratique du contexte, problème de la qualification, incertitude, apprentissage, formes compilées de prise de décisions - sont, bien sûr, très importants, et sont désormais intégrés à la conception des agents intelligents standard. Selon nous, cela prouve les progrès de l'TA et non son impossibilité.

L'un des arguments les plus forts de Dreyfus et Dreyfus concerne la nécessité de considérer des agents en situation et non des moteurs d'inférence logique désincarnés. Un agent dont la compréhension de «chien» ne provient que d'un ensemble limité de phrases logiques comme «*Chien(x) => Mammifère(x)* » est désavantagé par rapport à un agent qui a vu des chiens courir, a joué à la balle avec eux et s'est fait lécher par l'un d'eux. Comme le dit le philosophe *Andy* Clark (1998), « Les cerveaux biologiques sont d'abord et surtout les systèmes de commande des corps biologiques. Les corps biologiques évoluent et agissent dans le riche contexte du monde réel.» Pour comprendre comment les agents humains (ou animaux) fonctionnent, il faut prendre en compte l'agent dans son entier, et pas seulement le programme agent. En fait, la méthode dite de la **cognition incarnée** soutient que cela n'a aucun sens de considérer le cerveau séparément: la cognition a lieu dans un corps, qui s'insère lui-même dans un environnement. Il faut donc étudier le système comme un tout: le cerveau augmente son raisonnement en se référant à l'environnement, comme le fait le lecteur qui perçoit (et crée) des signes sur le papier pour transférer des connaissances. Dans le programme de la cognition incarnée, la robotique, la vision et les autres capteurs deviennent centraux, non périphériques.

26.2 lA forte: les machines peuvent-elles réellement penser?

De nombreux philosophes ont soutenu qu'une machine qui réussirait le test de Turing ne penserait toujours pas *réellement* mais ne ferait que *simuler* la pensée. Là encore, Turing avait prévu l'objection. Il cite un discours du Pr Geoffrey Jefferson (1949) :

> Tant qu'un ordinateur ne pourra pas écrire un sonnet ou un concerto parce qu'il a pensé et éprouvé des émotions et non en arrangeant des symboles au hasard, nous ne pourrons accepter qu'une machine égale un cerveau - autrement dit, non seulement qu'elle écrit mais qu'elle sait qu'elle écrit.

Turing qualifie cette position d'argument de la **conscience:** la machine doit être consciente de ses propres états mentaux et de ses actions. Si la conscience est un sujet important, l'objection essentielle de Jefferson relève en réalité de la **phénoménologie,** l'étude de l'expérience directe: la machine doit réellement ressentir des émotions. D'autres se focalisent sur **l'intentionnalité** - en d'autres termes, sur la question de savoir si les prétendus désirs, croyances et autres représentations portent réellement sur une entité du monde réel.

La réponse de Turing à cette objection est intéressante. Il aurait pu arguer que les machines *peuvent* être conscientes (ou avoir une phénoménologie, ou avoir des intentions). À la place, il soutient que la question est tout aussi mal posée que si l'on demandait: « Les machines peuvent-elles penser?» D'ailleurs, pourquoi devrait-on exiger des machines des standards supérieurs à ceux que l'on attend des humains? Après tout, dans la vie ordinaire, les états mentaux d'autrui ne sont pas directement observables. Néanmoins, dit Turing,

«Au lieu de nous battre continuellement sur ce point, nous admettons généralement par convention sociale que tout le monde pense».

Turing soutient que Jefferson accepterait d'étendre cette convention sociale aux machines si seulement il en rencontrait qui agissent intelligemment. Il cite le dialogue suivant, qui fait maintenant tellement partie de la tradition orale de l'lA qu'il est impossible d'en faire l'économie:

> HUMAIN: À la première ligne de votre sonnet, vous écrivez: «Te comparerai-je à un jour d'été ».
> «Un jour de printemps» n'irait-il pas aussi bien, sinon mieux?
>
> MACHINE: Le vers serait faux.
>
> HUMAIN: Et « un jour d'hiver»? Le vers serait juste.
>
> MACHINE: Oui, mais personne n'a envie qu'on le compare à un jour d'hiver.
> HUMAIN Diriez-vous que M. Pickwick vous rappelle Noël?
>
> MACHINE: En quelque sorte.
>
> HUMAIN: Pourtant, oël est un jour d'hiver, et je ne pense pas que M. Pickwick récuserait la comparaison.
>
> MACHINE: Je trouve que vous n'êtes pas sérieux. Par «jour d'hiver», on entend habituellement un jour d'hiver ordinaire, pas un jour spécial comme Noël.

On peut facilement imaginer un futur dans lequel de telles conversations avec des machines seront monnaie courante et qu'il deviendra coutumier de ne plus faire de distinction linguistique entre pensée « naturelle» et pensée « artificielle». Une transition similaire s'est produite durant les années qui ont suivi 1848, quand Frederick Wôhler a synthétisé l'urée artificielle pour la première fois. Avant cet événement, la chimie organique et la chimie minérale étaient deux mondes distincts et beaucoup pensaient qu'aucun processus ne pourrait jamais convertir des composés minéraux en composés organiques. Une fois la synthèse accomplie, les chimistes convinrent que l'urée artificielle *était* de l'urée puisqu'elle possédait les bonnes propriétés physiques. Ceux qui avaient postulé l'existence d'une propriété intrinsèque possédée par une substance organique qu'une substance de synthèse ne pourrait jamais avoir étaient confrontés à l'impossibilité d'imaginer un test capable de révéler la supposée déficience de l'urée artificielle.

En ce qui concerne la pensée, nous n'avons pas encore atteint 1848, et certains sont convaincus que la pensée artificielle, si impressionnante soit -elle, ne sera jamais réelle. Par exemple, le philosophe John Searle (1980)) avance l'argument suivant:

> Personne ne suppose qu'une simulation de tempête sur ordinateur va le tremper jusqu'aux os ... Pourquoi donc quelqu'un de sain d'esprit irait-il supposer qu'une simulation informatique de processus mentaux abrite réellement des processus mentaux? (pages 37 – 38).

S'il est aisé de convenir que les simulations de tempête sur ordinateur ne mouillent pas, la transposition de cette analogie aux simulations informatiques de processus mentaux est moins évidente. Après tout, une simulation hollywoodienne utilisant des arroseurs et des

générateurs de vent mouille *bien* les acteurs, et une simulation de tempête dans un jeu vidéo mouille *bien* les personnages simulés. La plupart des gens acceptent sans problème l'idée que la simulation informatique d'une addition est une addition et que la simulation informatique d'une partie d'échecs est une partie d'échecs, En fait, on dit généralement qu'on *implémente* une addition ou un jeu d'échecs, non qu'on les *simule.* Un processus mental ressemble-t-il plutôt à une tempête ou plutôt à une addition?

La réponse de Turing -la convention sociale - suggère que le problème finira par disparaître de lui-même une fois que les machines auront atteint un certain niveau de sophistication. Cela aurait pour effet de

dissoudre la différence entre l'IA forte et l'IA faible. Contre cet argument, on peut insister sur le fait qu'il existe un problème *factuel* en jeu: les humain *ont* de vrais esprits, alors que les machines pourraient en avoir ou non. Pour traiter ce problème factuel, il faut comprendre comment il se fait que les humains ont de vrais esprits, et pas seulement des corps qui génèrent des processus neurophysiologiques. Les efforts des philosophe pour résoudre ce **problème du corps et de l'esprit** concernent directement la question de savoir si les machines pourraient avoir de vrais esprits.

Le problème du corps et de l'esprit a été étudié par les philosophes grecs et par diverses écoles de pensée hindouistes, mais le premier à l'analyser en profondeur au XVIIe siècle fut le philosophe et mathématicien français **René Descartes**, Dans ses *Méditations métaphysiques* (1641), il considérait l'activité mentale de penser (un processus sans étendue spatiale ni propriétés matérielles) et les processus physiques du corps, et concluait que les deux devaient exister dans des domaines séparés - théorie que nous qualifierions maintenant de **dualiste**. Le problème auquel font face les dualistes est la question de savoir comment l'esprit peut contrôler le corps s'il s'agit réellement de deux entités distinctes. Descartes supposait que c'était la glande pinéale qui servait d'intermédiaire, ce qui ne fait que poser une nouvelle question : comment l'esprit contrôle- t - il la glande pinéale?

La théorie **moniste** de l'esprit, souvent appelée **matérialisme**, évite ce problème en affirmant que l'esprit n'est pas séparé du corps, et que les états mentaux *sont* des états physiques. La plupart des philosophes de l'esprit modernes professent une forme ou une autre de matérialisme, et le matérialisme autorise, au moins en principe, la possibilité d'une IA forte, Mais le problème des matérialistes est d'expliquer comment les états physiques - en particulier les configurations moléculaires et les processus électrochimiques du cerveau – peuvent être simultanément des **états mentaux**, comme souffrir, apprécier un hamburger, savoir qu'on monte à cheval ou croire que Vienne est la capitale de l'Autriche.

26.2.1 Le fonctionnalisme et l'expérience du cerveau dans la cuve

Les philosophes matérialistes ont tenté d'expliquer ce que cela signifie de dire qu'une personne - et, par extension, un ordinateur - est dans un état mental donné, En particulier, ils se sont concentrés sur les **états intentionnels**, Ce sont des états - tels que croire, savoir, désirer, craindre, etc, - qui se rapportent à un aspect particulier du monde externe, Par exemple, savoir qu'on est en train de manger un hamburger est une croyance *à propos* du hamburger et de ce qui lui arrive.

Si les matérialistes ont raison, il s'ensuit que la description correcte de l'état d'une personne est *déterminée* par son état cérébral, En conséquence ,si je suis actuellement concentré sur le fait de manger un hamburger de manière consciente, mon état cérébral du moment est une instance de la classe des états mentaux “savoir qu'on est en train de manger un hamberger”. Bien entendu, les configurations spécifiques de tous les atomes de

mon cerveau ne sont pas essentielles: il existe de nombreuses configurations de celui-ci ou de celui d'autres personnes, qui appartiendraient à la même classe d'états mentaux. Le point capital est que le même état cérébral ne pourrait pas correspondre à un état mental, fondamentalement distinct, comme savoir qu'on est en train de manger une banane.

La simplicité de ce point de vue est mise en défaut par certaines expériences de pensée..
Imaginez, si vous le voulez bien, que votre cerveau ait été extrait de votre corps à la naissance et placé dans une cuve merveilleusement conçue. La cuve lui permet de subsister, de grossir et de se développer. En même temps, il reçoit d'une simulation informatique des signaux électroniques relatifs à un monde entièrement fictif, et les signaux moteurs qu'il émet son- interceptés et utilisés pour modifier ladite simulation de façon appropriée [4]. En fait, la vie simulée que vous vivez est une réplique exacte de celle que vous auriez vécue si votre cerveau n'avait pas été placé dans la cuve, y compris la manducation simulée

de hamburgers simulés.. Ainsi, vous pourriez avoir un état cérébral identique à celui de quelqu'un qui mange réellement un hamburger réel, mais il serait littéralement faux de dire que vous avez l'état mental «savoir qu'on mange un hamburger». Vous n'êtes pas en train de manger un hamburger, vous n'avez même jamais rencontré un hamburger, et, en conséquence, vous ne pourriez pas avoir un tel état mental.

Cet exemple semble contredire la position selon laquelle les états cérébraux déterminent les états mentaux. Une façon de résoudre ce dilemme consiste à dire que le contenu des états mentaux peut être interprété de différents points de vue. L'optique du contenu large l'interprète du point de vue d'un observateur externe omniscient, qui a accès à l'intégralité de la situation et peut distinguer les différences présentes dans le monde. Dans ce cas, le contenu des états mentaux englobe à la fois l'état cérébral et l'histoire de l'environnement, En revanche, l'optique du contenu étroit ne prend en compte que l'état cérébral. Le contenu étroit des états cérébraux d'un vrai mangeur de hamburger et celui du cerveau dans la cuve «mangeant» un hamburger simulé seraient les mêmes dans les deux cas.

Le contenu large est parfaitement approprié si l'on a pour buts d'attribuer des états mentaux à d'autres dont on partage le monde, de prédire leur comportement probable e ses effets, etc. C'est dans ce contexte que notre langage ordinaire sur le contenu mental a évolué. En revanche, si l'on se préoccupe de la question de savoir si les systèmes d'IA pensent réellement et ont réellement des états mentaux, c'est le contenu étroit qui est approprié: dire que le fait qu'un système d'IA pense réellement dépend des conditions externes à ce système n'a tout simplement aucun sens. Le contenu étroit est également pertinent si l'on réfléchit à la façon de concevoir des systèmes d'IA ou de comprendre leur fonctionnement, parce que c'est le contenu étroit d'un état cérébral qui détermine ce que sera le (contenu étroit du) prochain état cérébral. Cela conduit naturellement à l'idée que ce qui compte pour un état cérébral- ce qui lui fait avoir un type de contenu mental et non un autre - est son rôle fonctionnel dans le fonctionnement mental de l'entité concernée.

26.2.2 Le fonctionnalisme et l'expérience du cerveau remplacé

Pour le fonctionnalisme, un état mental est une condition causale intermédiaire entre une entrée et une sortie. Selon cette théorie, deux systèmes quelconques dotés de processus mentaux isomorphes auraient les mêmes états mentaux. En conséquence, un programme informatique pourrait présenter les mêmes états mentaux qu'une personne. Bien entendu.

[4] Cette situation est peut -être familière aux lecteurs qui ont vu le film *Matrix.*

nous n'avons pas encore dit ce que signifie réellement «isomorphe ». mais l'on suppose qu'il existe un niveau d'abstraction en dessous duquel une mise en œuvre particulière n'aura pas d'importance.

Les thèses du fonctionnalisme sont illustrées des plus clairement par **l'expérience du cerveau remplacé,** Cette expérience de pensée a été introduite au milieu des années 1970 par le philosophe **Clark Glymour** et étudiée par **John Searle** (1980), mais on l'associe plus communément aux travaux du roboticien Hans Moravec (1988). Voyons en quoi elle consiste. Supposez que la neurophysiologie ait fait des progrès tels que les comportements d' entrée/ sortie et la connectivité de tous les neurones du cerveau humain soient parfaitement compris. Supposez, en outre, que l'on puisse construire des équipements électroniques microscopiques qui miment ce comportement et qu'il soit possible de les interfacer de façon homogène avec le tissu neural, Enfin, supposez qu'une technique chirurgicale miraculeuse puisse remplacer certains neurones par les dispositifs électroniques correspondants sans interrompre le fonctionnement du cerveau. **L'expérience consiste à remplacer progressivement tous les neurones d'un cerveau par des dispositifs électroniques**.

Nous nous intéressons à la fois au comportement externe et au vécu interne du sujet, durant et après l'opération. Selon la définition de l'expérience, son comportement externe doit demeurer inchangé par rapport à ce qui serait observé si l'opération n'avait pas eu lieu [5] . Un tiers aurait sans doute du *mal* à certifier

la présence ou l'absence de conscience, mais le sujet devrait au moins pouvoir être conscient. Les avis sont nettement partagés. Moravec, chercheur en robotique et **fonctionnaliste**, est convaincu que sa conscience n'en serait aucunement affectée. **Searle**, philosophe et tenant du **naturalisme biologique**, est convaincu que sa conscience s'évanouirait:

> « Totalement stupéfait, vous vous rendez compte que *vous* êtes en train de perdre tout contrôle sur votre comportement externe. Par exemple, lorsque les médecins testent votre vision, vous les entendez dire « Nous tenons devant vous un objet rouge. Dites-nous ce que vous voyez.» Vous voulez vous exclamer «Je ne vois rien, je suis complètement aveugle. » Mais vous vous entendez dire, sans pouvoir aucunement maîtriser votre voix: «Je vois un objet rouge devant moi» ... Votre expérience consciente sombre lentement dans le néant tandis que *votre* comportement observable reste le même ». (Searle, 1992)

L' on peut faire mieux que s'appuyer sur l'intuition. Première remarque: pour qu'un comportement externe demeure le même alors que le sujet perd graduellement conscience, il faudrait que la volition de ce dernier soit annihilée instantanément et totalement; sinon, la disparition de la conscience se refléterait dans le comportement externe - «Au secours, je disparais! » ou toute autre formulation du même ordre. La suppression instantanée de la volition résultant du remplacement des neurones un à un semble impossible à soutenir.

Deuxièmement, considérez ce qui se passerait si nous interrogions le sujet sur son expérience consciente durant la période au cours de laquelle il ne reste plus de vrais neurones. Selon les conditions de l'expérience, nous obtiendrons des réponses telles que «Je me sens bien. Mais je dois dire que je suis un peu surpris parce que je croyais à l'argument de Searle. » Ou bien nous pourrions titiller le sujet *avec* une baguette pointue et observer la réponse: *«Aie,* ça fait mal. » Maintenant, dans le cours normal des choses, le sceptique peut rejeter de tels résultats issus d'un programme d'IA en arguant qu'il s'agit d'un pur procédé. Bien sûr, il est

assez facile d'utiliser une règle du type «si le capteur 12 retourne "Fort" alors dire "Aïe ." Mais comme nous avons reproduit les propriétés fonctionnelles d'un cerveau humain normal, nous supposons que le cerveau électronique ne contient pas de tels artéfacts. Il faut une explication des manifestations de conscience produites par ce cerveau électronique qui ne fasse appel qu'aux propriétés fonctionnelles des neurones, *Et cette explication doit également s'appliquer au vrai cerveau, qui possède les mêmes propriétés fonctionnelles. Il n'existe,* semble-t-il, que trois conclusions possibles:

1. Les mécanismes causaux de la conscience qui génèrent ce type de réponse dans un cerveau normal sont toujours à l'œuvre dans la version électronique, qui est de ce fait consciente,

2. Les événements mentaux conscients du cerveau normal n'ont pas de connexion causale avec le comportement et sont absents du cerveau électronique, qui n'est donc pas conscient.

3. L'expérience est impossible et donc spéculer à son propos n'a aucun sens.

Même si nous ne pouvons pas exclure la deuxième possibilité, elle réduit la conscience à ce que les philosophes appellent un rôle **épiphénoménal** - quelque chose qui se produit mais ne projette, pour ainsi dire, aucune ombre sur le monde observable, Qui plus est, si la conscience est réellement épiphénoménale, il est impossible que le sujet dise «Aie» *parce que ça fait mal*- autrement dit, à cause de l'expérience consciente de la douleur. En lieu et place, le cerveau doit contenir un second mécanisme, **inconscient** celui-là, qui est responsable du «Aie».

Patricia Churchland (1986) souligne que les arguments fonctionnalistes qui opèrent au niveau du neurone peuvent également s'appliquer au niveau d'une unité fonctionnelle plus large - un groupe de neurones, un module mental, un lobe, un hémisphère ou la totalité du cerveau. Cela signifie que, si vous acceptez l'idée que l'expérience du cerveau remplacé montre que le cerveau de remplacement est conscient, vous devez également croire que la conscience persiste quand on remplace le cerveau entier par un circuit qui met à jour son état et projette les entrées sur les sorties *via* une immense **table de correspondance**. C'est une idée déconcertante pour tous ceux (y compris Turing lui-même) qui ont l'intuition que les tables de correspondances ne sont pas conscientes - ou au moins que les expériences conscientes générées durant la consultation d'une table ne sont pas les mêmes que celles générées par un système descriptible (même dans un sens informatique simpliste) comme générant et accédant à des croyances, des introspections, des buts, etc. Cela suggérerait que l'expérience de la prothèse mentale ne peut pas remplacer la totalité du cerveau d'un coup si elle doit guider efficacement les intuitions, mais cela ne signifie pas qu'elle doive remplacer un atome à la fois comme **Searle** veut nous le faire croire.

26.2.3 Le naturalisme biologique et la chambre chinoise

Le **fonctionnalisme** a été très fortement mis en question par le **naturalisme biologique** de **John Searle** (l980), selon lequel les états mentaux sont des **fonctionnalités émergentes** de haut niveau causées par des processus physiques de bas niveau *dans les neurones,* et que ce sont les propriétés (non spécifiées) de ces neurones qui importent. En conséquence, il est impossible de dupliquer des état mentaux simplement parce qu'un programme possède la même structure fonctionnelle et le même comportement en termes d'entrées-sorties : il faudrait que le programme s'exécute sur une architecture dotée du même pouvoir causal

que les neurones. Pour étayer son point de vue, Searle décrit un système hypothétique qui exécute clairement un programme et réussit le test de Turing, mais qui, tout aussi clairement (selon Searle) ne *comprend* rien à ses entrées ni à ses sorties. Il conclut que le fait d'exécuter le programme approprié (autrement dit d'obtenir les bons résultats) n'est pas une condition *suffisante* pour être un esprit.

Le système est constitué d'un humain qui ne comprend que l'anglais, équipé d'un recueil de règles rédigées en anglais et de plusieurs piles de feuilles de papier - certaines vierges, d'autres portant des inscriptions indéchiffrables. (L'humain joue donc le rôle de l'unité centrale, le recueil de règles est le programme et les piles de feuilles sont les unités de stockage.) Le système se trouve à l'intérieur d'une pièce munie d'une petite ouverture vers l'extérieur. Par l'ouverture apparaissent des feuilles de papier couvertes de symboles cryptiques. L'humain trouve les règles d'appariement des symboles dans le recueil et suit les instructions. Il peut s'agir d'écrire des symboles sur de nouvelles feuilles de papier, de trouver des symboles dans les piles, de ré agencer les piles, etc. Finalement, les instructions auront pour effet qu'un ou plusieurs symboles seront transcrits sur un morceau de papier qui sera retransmis au monde extérieur.

Jusque-là, tout va bien. Mais, de l'extérieur, nous voyons un système qui accepte des entrées sous forme de phrases chinoises et génère en chinois des réponses qui sont aussi « intelligentes» que celles de la conversation imaginée par Turing [6]. Searle argumente alors ainsi: la personne dans la pièce ne comprend pas le chinois (c'est une donnée). N'étant constitués que de papier, ni le recueil de règles ni les piles de feuilles ne comprennent le chinois. En conséquence, il n'y a là aucune compréhension du chinois. *Donc, selon Searle, l'exécution du programme correct ne génère pas nécessairement de la compréhension.*

Comme Turing, Searle a envisagé et entrepris de réfuter un certain nombre de réponses à cet argument. Plusieurs commentateurs, notamment John McCarthy et Robert Wilensky, ont proposé ce que Searle appelle la réponse systémique *(system reply).* L'objection est la suivante: demander si l'humain dans la pièce comprend le chinois équivaut à demander si l'unité centrale peut extraire des racines cubiques. La réponse est négative dans les deux cas, et, dans les deux cas, selon la réponse systémique, le système entier *possède* la capacité en question. Bien sûr, si l'on demande à la chambre chinoise si elle comprend le chinois, la

réponse sera affirmative (en chinois courant). Selon la « convention sociale» de Turing, cela devrait suffire. La réponse de Searle se borne à réitérer que l'humain ne comprend pas et que le papier ne peut pas comprendre: il n'y a donc pas de compréhension. Il semble s'appuyer sur l'argument selon lequel une propriété d'un tout doit résider dans l'une de ses parties. Pourtant, l'eau est humide, même si ni H_2 ni 0 ne le sont.

La véritable argumentation de Searle repose sur les quatre axiomes suivants (Searle, 1990) :

4.Les programmes informatiques sont formels (syntaxiques).

5.L'esprit humain a des contenus mentaux (sémantiques).

6.La syntaxe en soi n'est ni constitutive, ni suffisante pour la sémantique.

7.Le cerveau est la cause de l'esprit.

Des trois premiers axiomes, il conclut que les programmes ne sont pas suffisants pour les esprits. Autrement dit, un agent exécutant un programme *pourrait* être un esprit, mais le seul

[6]. Le fait que les piles de papier pourraient contenir des milliards de pages et que la génération des réponses pourrait demander des millions d'années n'a aucune incidence sur la structure *logique* de l'argument. L'un des buts de la formation philosophique est de développer un sens aigu du caractère pertinent ou non d'une objection.

fait d'exécuter le programme n'en fait pas *nécessairement* un esprit. Du quatrième, il déduit que « tout autre système capable de causer des esprits aurait des pouvoirs causaux (au moins) équivalents à ceux du cerveau». De là, il infère que tout cerveau artificiel devrait dupliquer les pouvoirs causaux du cerveau, non se contenter d'exécuter un programme particulier, et que les cerveaux humains ne produisent pas de phénomènes mentaux pour la seule raison qu'ils exécutent un programme.

Or les axiomes sont sujets à controverse. Par exemple, les axiomes 1 et 2 s'appuient sur une distinction non spécifiée entre la syntaxe et la sémantique qui semble étroitement apparentée à la distinction entre le contenu étroit et le contenu large. D'un côté, on peut voir les ordinateurs comme des machines qui manipulent des symboles syntaxiques; de l'autre, on peut les voir comme manipulant du courant, ce que font essentiellement les cerveaux (pour ce que nous en savons actuellement). Il semble donc que l'on pourrait aussi bien dire que les cerveaux sont syntaxiques.

En supposant que l'on interprète généreusement les axiomes, la conclusion que les programmes ne suffisent pas pour créer un esprit en découle *bien.* Mais cette conclusion est insatisfaisante- tout ce que Searle démontre, c'est que si l'on nie explicitement le fonctionnalisme (ce que fait son axiome 3), on ne peut pas nécessairement conclure que des **non-cerveaux** sont des esprits. Comme c'est relativement raisonnable, voire presque tautologique, tout le débat revient à savoir si l'axiome (3) est acceptable. Selon Searle, l'intérêt de l'argument de la chambre chinoise est de permettre d'intuiter l'axiome (3). La réaction publique montre que cet argument se comporte comme ce que **Daniel Dennett** (1991) appelle une **:** Il amplifie les premières intuitions, de sorte que les naturalistes biologiques sont plus convaincus de leurs positions, et que les fonctionnalistes sont convaincus que seul l'axiome 3 n'est pas fondé, ou que l'argument de Searle en général n'est pas convaincant. L'argument excite les combattants mais n'a pas fait grand chose pour changer l'opinion de qui que ce soit. Searle ne se décourage pas et a récemment commencé à qualifier **la chambre chinoise de réfutation de l'lA forte** et non plus seulement d'« argument» (Snell, 2008).

Même ceux qui acceptent l'axiome 3, et donc l'argument de **Searle**, ne peuvent s'appuyer que sur leurs intuitions pour décider quelles entités sont des esprits. L'argument prétend montrer que la chambre chinoise n'est pas un esprit *en vertu du fait qu'elle exécute le programme* mais il reste muet sur la façon de décider si la chambre (ou un ordinateur, un autre type de machine ou un extraterrestre) est un esprit *en vertu d'une autre raison.* Searle Iui-même affirme que certaines machines ont un esprit: **les humains sont des machines biologiques** qui en possèdent un. Selon lui, les humains peuvent ou non exécuter quelque chose

de semblable à un programme d'lA, mais, s'ils le font, ce n'est pas pour cela qu'ils sont des esprits. Il en faut plus pour faire un esprit: d'après lui, quelque chose d'équivalent aux pouvoirs causaux des neurones individuels. Ce que sont ces pouvoirs demeure non spécifié. Notons toutefois que les neurones ont évolué pour jouer des rôles *fonctionnels:* des créatures dotées de neurones apprenaient et décidaient bien avant que la conscience n'entre en scène. Ce serait une remarquable coïncidence si de tels neurones se trouvaient générer de la conscience en raison de pouvoirs causaux sans rapport avec leurs capacités fonctionnelles. Après tout, ce sont les capacités fonctionnelles qui dictent la survie d'un organisme.

Dans le cas de la chambre chinoise, Searle s'appuie sur l'intuition, non sur la preuve: regardez cette pièce: où est l' esprit là-dédans ? Mais l'on pourrait avancer le même argument à propos du cerveau: regardez cette collection de cellules (ou d'atomes), fonctionnant en aveugle selon les lois de la biochimie (ou de la physique) – où est l'esprit là-dedans ? Pourquoi un morceau de cerveau peut-il être un esprit quand un morceau de foie ne le peut pas ? cela demeure le grand mystère.

Page 1089

26.2.4 La conscience, les qualia et le fossé explicatif

Une question traverse toutes les discussions sur l'AI forte -l'éléphant dans amphithéâtre, pour ainsi dire: celle de la conscience. On décompose souvent la conscience en aspects tels que l'**entendement** et la **conscience de soi**. Celui sur lequel nous allons nous concentrer est ***l'expérience subjective****:* comment se fait -il que l'on *ressente* quelque chose dans certains **états cérébraux** (par exemple lorsqu'on mange un hamburger), alors qu'il n'y a probablement aucune **sensation** associée à d'autres **états physiques** (par exemple lorsqu'on est une pierre). Le terme technique pour désigner la nature intrinsèque des expériences est ***qualia*** (du mot latin signifiant approximativement **« les qualités des êtres »)**.

Les *qualia* présentent une difficulté pour les explications fonctionnalistes de l' esprit, parce que des *qualia* différents pourraient être impliqués dans des processus causaux autrement isomorphes. Ainsi, considérons l'expérience de pensée du spectre inversé, dans laquelle l'expérience subjective d'une personne X lorsqu'elle voit des objets rouges est la même que ce que le reste d'entre nous ressent en voyant des objets verts, et inversement. X continue à qualifier les objets rouges de « rouges » , s'arrête lorsque le feu passe au rouge et convient que la rougeur des feux rouges est plus intense que celle du soleil couchant. Pourtant, l'expérience subjective de X est tout simplement différente.

Les *qualia* ne sont pas seulement problématiques pour le fonctionnalisme, mais aussi pour toute la science. Supposez, pour les besoins de l'étude, que la recherche scientifique sur le cerveau soit terminée: nous avons découvert que le processus neural P12 dans le neurone N 117 transforme la molécule *A* en une molécule *B,* etc., etc. Il n'existe simplement aucune forme de raisonnement actuellement acceptée qui permette de déduire de ces découvertes que l' entité qui possède ces neurones a une expérience subjective donnée. Ce **fossé explicatif** a conduit certains philosophes à conclure que les êtres humains sont tout bonnement incapables de comprendre réellement leur propre conscience. D'autres, notamment **Daniel Dennett** (1991), évitent ce fossé en déniant l'existence des *qualia,* et les attribuent à une confusion philosophique.

Turing lui-même concède que la question de la conscience est difficile mais nie qu' elle soit très pertinente pour la pratique de l'lA: «Je ne veux pas donner l'impression de penser que la conscience n'est pas un mystère ... Mais je ne pense pas que ce mystère doive nécessairement être résolu avant de pouvoir répondre à la question qui nous occupe dans cet article.» Nous sommes d'accord avec **Turing**: ce qui nous intéresse, c'est de **créer des programmes qui se comportent intelligemment**. Nous ne sommes pas équipés pour entreprendre un autre projet qui consisterait à les rendre conscients, et nous serions incapables d'en prédire le succès.