Université de Montréal

Invariance organisationnelle et conscience artificielle

par
Julien Brodeur

Département de philosophie
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de Maître ès arts (M.A.)
en philosophie, option Philosophie au collégial

Août 2022

Université de Montréal
Unité académique : Département de philosophie, Faculté des arts et des sciences

Ce mémoire intitulé

Invariance organisationnelle et conscience artificielle

Présenté par
Julien Brodeur

A été évalué par un jury composé des personnes suivantes :

Molly Kao
Présidente du jury

Jonathan Simon
Directeur de recherche

Maxime Doyon
Membre du jury

RÉSUMÉ

Ce mémoire se penche sur la possibilité de la conscience artificielle. Plus spécifiquement, je me demande s'il est possible qu'un robot, un ordinateur ou toute autre machine ait une conscience phénoménale, i.e. qu'il y ait un effet que cela fait que d'être ces systèmes. Après avoir brièvement caractérisé la conscience phénoménale, j'investiguerai quelques problèmes qui sont propres à la conscience, soit le problème difficile de la conscience ainsi que le problème des autres esprits, dans le but d'établir le cadre conceptuel qui nous permettra de réfléchir quant à la possibilité de la conscience artificielle. Dans le deuxième chapitre, je défendrai la thèse selon laquelle la conscience artificielle est possible en m'appuyant notamment sur le principe d'invariance organisationnelle défendu, entre autres, par David Chalmers, ainsi que sur la théorie computationnelle de l'esprit. Finalement, dans le troisième et dernier chapitre, j'évaluerai diverses objections contre la possibilité de la conscience artificielle que je tenterai tour à tour de réfuter dans le but maintenir ma thèse initiale aussi intacte que possible.

Mots-clés : Conscience artificielle; conscience phénoménale; problème difficile de la conscience; problème des autres esprits; principe d'invariance organisationnelle; David Chalmers; théorie computationnelle de l'esprit

ABSTRACT

This thesis examines the possibility of artificial consciousness. More specifically, I consider the possibility for a robot, computer or any other machine to have phenomenal consciousness, i.e. that there is something it is like to be those systems. After having briefly characterized phenomenal consciousness, I will investigate some problems that are specific to consciousness, namely the hard problem of consciousness as well as the problem of other minds, in order to establish the conceptual framework that will allow us to reflect upon the possibility of artificial consciousness. In the second chapter, I will defend the thesis that artificial consciousness is possible by relying on the principle of organizational invariance which is defended by David Chalmers, among others, as well as on the computational theory of the mind. Finally, in the third and last chapter, I will assess various objections against the possibility of artificial consciousness which I will try to refute in turn in order to keep my initial thesis as intact as possible.

Keywords: *Artificial consciousness*; *phenomenal consciousness*; *hard problem of consciousness*; *problem of other minds*; *principle of organizational invariance*; *David Chalmers*; *computational theory of mind*

TABLE DES MATIÈRES

REMERCIEMENTS

Merci à Jonathan Simon; ce travail ne serait pas ce qu'il est devenu sans ses judicieux conseils et ses commentaires toujours pertinents et éclairants.

Merci à mes parents pour leur appui inconditionnel, malgré le fait qu'ils n'aient pas toujours compris ce sur quoi je travaillais.

Merci à mes ami.es et collègues qui, à travers nos discussions, ont fait de la rédaction de ce mémoire, une expérience enrichissante, parfois ardue, mais surtout plaisante.

Finalement, merci à mes professeur.es et à tous ceux et celles qui, même sans le savoir, ont contribué à ce mémoire.

INTRODUCTION

L'abondance de personnages de science-fiction comme Ava dans le film *Ex Machina*, Sam dans le film *Her* ou encore les réplicants de *Blade Runner* témoignent de notre fascination envers ces machines, ces robots et ordinateurs qui ressemblent tellement à des êtres humains que nous leur attribuons des caractéristiques humaines. Une de ces caractéristiques est la conscience, que certain croit exclusive aux systèmes biologiques. Bien que ces personnages soient fictifs, il est tout de même intéressant de se questionner sur la possibilité de la conscience artificielle : serait-il possible qu'une machine, un ordinateur ou encore un système d'intelligence artificielle soit conscient(e), de la même façon que les êtres humains semblent l'être? En d'autres mots, est-il possible qu'Ava, Sam et les réplicants soient véritablement conscient.es ou est-ce que cette hypothèse est vouée à ne demeurer qu'un fantasme de la science-fiction?

La conscience est-elle réservée à des systèmes « organiques » ou, inversement, pourrait-elle émerger de n'importe quel substrat? Il semble qu'il sera possible, dans le futur, de reproduire toutes les capacités intellectuelles humaines à l'aide d'une machine, par contre, un mystère demeure quant à la possibilité de synthétiser une conscience. Dans ce mémoire, j'opterai pour une approche optimiste vis-à-vis de la possibilité de la conscience artificielle et je défendrai la thèse que la conscience artificielle est, du moins en théorie/principe, possible.

Les conséquences et les répercussions de la possibilité de la conscience artificielle seraient nombreuses, notamment en ce qui a trait à la responsabilité et l'agentivité morale ainsi qu'à l'identité personnelle de ces machines conscientes. En effet, certains pensent que la conscience est une caractéristique fondamentale et

nécessaire de l'agentivité morale.[123] En d'autres termes, une machine consciente pourrait donc se voir être conférer le statut d'agent moral, avec toutes les responsabilités qui viennent avec. Également, on pourrait penser que le fait d'avoir une conscience est l'une des propriétés qui font d'une personne, « une personne ». C'est donc dire qu'inversement, une entité ne pourrait pas être considérée comme une personne si elle n'est pas consciente[4] et que c'est ce qui crée le sentiment d'identité personnelle à travers le temps.[5] Ainsi, la possibilité de la conscience artificielle entraînerait de nombreux questionnements éthiques et moraux : serait-il moralement acceptable de débrancher une machine consciente? À qui reviendrait la responsabilité morale des actions posées par la machine? Est-il souhaitable que les machines soient conscientes? Etc…[6] Faute de temps et par soucis de me concentrer sur la question de la possibilité de la conscience artificielle, je n'aborderai pas, ou seulement indirectement, les conséquences éthiques et morales d'une telle possibilité.[7]

Dans le chapitre 1, je tenterai de fournir une caractérisation de ce qu'est la conscience pour ensuite souligner quelques problèmes qui découlent directement de la particularité de la nature de conscience. Une fois ces concepts étayés, nous serons mieux outillés pour observer ces enjeux à travers le prisme de la possibilité de la conscience artificielle. Dans le chapitre 2, je me tournerai vers les arguments qui

---

[1] Kenneth Einar Himma, « Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? », *Ethics and Information Technology* 11 (2009): 19-29.
[2] Neil Levy, *Consciousness and Moral Responsibility* (Oxford: Oxford University Press, 2014).
[3] Carissa Véliz, « Moral zombies: why algorithms are not moral agents », *AI and Society* 36 (2021) : 487-497.
[4] Le fait d'avoir une conscience est une propriété nécessaire pour être une personne, à proprement parlé, mais ce n'est toutefois pas une propriété suffisante. En effet, plusieurs animaux non-humains sont conscients, mais nous ne dirions toutefois pas d'eux que ce sont des personnes.
[5] Barry Dainton et Tim Bayne, « Consciousness as a guide to personal persistence », *Australasian Journal of Philosophy* 83 (4) (2005): 549-571.
[6] Wendell Wallach et Colin Allen, *Moral Machines : Teaching Robots Right from Wrong* (Oxford : Oxford University Press, 2010), 13-23.
[7] Martin Gibert et Dominic Martin, « In search of the moral status of AI: why sentience is a strong argument », *AI and Society* 37 (2022) : 319-330.

avancent que la conscience artificielle est possible, i.e. qu'il serait possible de synthétiser une conscience. Je me baserai notamment sur les travaux de David Chalmers ainsi que sur la théorie computationnelle de l'esprit pour établir mon argumentaire en ce sens. Pour terminer, dans le chapitre 3, j'examinerai plusieurs objections qui stipulent que la conscience artificielle est impossible. Je tenterai par la suite de réfuter chacun de ces contre-arguments, dans le but de maintenir ma thèse initiale aussi intacte que possible.

CHAPITRE I : LA PARTICULARITÉ DE LA CONSCIENCE

1.1 Une caractérisation de la conscience

En premier lieu, il est nécessaire de réfléchir à ce qu'est la conscience avant de se demander s'il serait possible de synthétiser une conscience artificielle. Pour les fins de ce mémoire, lorsque je ferai référence à la notion de conscience, cela désignera plus spécifiquement la conscience phénoménale, soit l'ensemble de nos expériences subjectives. La plupart des gens seraient d'avis que la conscience est un phénomène simple, mais qui n'est pas — ou difficilement — définissable.[8] C'est, entre autres, pourquoi il n'existe pas vraiment de consensus par rapport à la définition même de ce qu'est la conscience. Certains penseurs remettent même en question l'existence d'un tel concept en avançant que la conscience n'est, au mieux, qu'illusoire.[9] Je n'ai pas la prétention ni le temps requis pour résoudre ces débats complexes par rapport à ce qu'est ou n'est pas la conscience, mais je vais tout de même tenter de formuler une caractérisation ainsi que de tracer les grandes lignes qui émergent d'un tel concept dans le but de nous permettre de réfléchir subséquemment à la question de la possibilité de la conscience artificielle. Ceci étant dit, il est si ardu de définir le concept de conscience que ces efforts de clarification sont parfois abandonnés[10] :

---

[8] David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), xi-xiv.

[9] Keith Frankish, « Illusionism as a Theory of Consciousness », *Journal of Consciousness Studies* 23 (2016):11-39.

[10] Ce que je veux dire par là n'est pas qu'il est, dans l'absolu, impossible de fournir une définition de la conscience, mais bien qu'il s'agît d'un concept assez simple, que tout le monde peut comprendre et saisir, mais qui est toutefois difficile à caractériser et à mettre en mots. Pour illustrer cette difficulté que nous avons à définir la conscience, nous pouvons faire référence à la définition du jazz fournie par Louis Armstrong : « *If you have to ask what jazz is, you'll never know* ». Ce qu'il veut dire par là est que pratiquement n'importe quelle personne qui écoute des pièces de divers styles de musique serait en mesure de dire lesquelles sont du jazz et lesquelles n'en sont pas (sauf exception). Pourtant, ces mêmes personnes ne seraient probablement pas capables de donner une définition convaincante et succincte de ce qu'est le jazz. C'est la même chose qui se produit dans le cas de la conscience; à peu près tout le monde est en mesure de saisir l'essence du concept, sans toutefois être en mesure de le définir clairement.

> I cannot define phenomenal consciousness in any remotely non-circular way. I don't consider this an embarrassment. The history of reductive definitions in philosophy should lead one not to expect a reductive definition of anything. But the best one can do for phenomenal consciousness is in some respects worse than for many other things because really all one can do is point to the phenomenon.[11]

La conscience est, en soi, si difficile à définir qu'il est souvent plus facile de donner des exemples d'expériences subjectives pour caractériser ce concept. En effet, la conscience est liée à des expériences subjectives, telles que de faire l'expérience de la rougeur d'une pomme ou encore de faire l'expérience subjective de l'odeur d'un café.[12] Ces exemples ne constitue pas une définition exhaustive de la conscience, mais ils nous permettent d'attirer notre attention sur ce qu'est le concept de la conscience, dont il sera question tout au long de ce texte. Ceci étant dit, malgré la difficulté à clairement définir en quoi elle consiste, j'estime tout de même que tout le monde a une idée générale (ou même intuitive) de ce qu'est la conscience. Bref, malgré les évidents problèmes qui se posent à quiconque qui tente de définir la conscience, je tenterai à présent d'en fournir une caractérisation adéquate au meilleur de mes capacités et de mes connaissances.

Dans un article maintenant devenu un incontournable du domaine d'études de la conscience, Thomas Nagel décrit la conscience comme étant *l'effet que cela fait* (*what-is-it-likeness*) que d'être une entité[13][14]; plus précisément, une entité est consciente s'il y a une manière subjective dont le monde semble ou apparaît du point de vue mental ou expérientiel de cette entité. Nagel utilise une chauve-souris comme exemple : il nous

---

[11] Ned Block, « Some concepts of consciousness », dans *Philosophy of Mind: Classical and Contemporary Readings*, ed. David Chalmers (Oxford, Oxford University Press, 2002), 206-219.
[12] Comme c'est le cas avec le jazz; il est plus facile de donner des exemples de jazz que de définir ce qu'est le jazz.
[13] Nagel utilise l'expression « *something it is like to be* … », qui est difficile à traduire dans une langue autre que l'anglais et qui semble perdre quelque peu de sa signification à la suite d'un tel exercice.
[14] Thomas Nagel, « What is it Like to be a Bat ? », *Philosophical Review* 83 (1974) : 435-50.

parait sensé de penser qu'une chauve-souris est consciente, puisqu'il semble y avoir un effet que cela fait que d'être cette chauve-souris. À l'inverse, nous dirions qu'une roche, une chaise ou encore une tasse ne sont toutes pas des entités conscientes puisqu'il ne semble pas y avoir d'effet que cela fait que d'être ces trois objets; il semble ne pas y avoir d'expérience subjective rattachée à une roche, une chaise et une tasse. Il nous semble hautement improbable qu'une roche, une chaise et une tasse aient une expérience subjective qui leur serait propre. En se basant sur la manière dont Nagel caractérise la conscience phénoménale, nous dirions donc qu'une machine est consciente seulement s'il y a un effet que cela fait que d'être cette machine. Ainsi, s'il s'avère qu'il y ait un effet que cela fait que d'être une machine — ou n'importe quel système artificiel —, nous conclurions alors que celle-ci bel et bien consciente et que la conscience artificielle est possible.

Certains philosophes de la conscience, comme Ned Block et John Searle, ont suggéré qu'il y a quelque chose de fondamental dans l'expérience subjective qui ne peut pas être capturée par aucun programme computationnel.[15][16] En d'autres mots, ils estiment qu'aucun programme computationnel ne serait en mesure de reproduire toutes les caractéristiques et les capacités de l'esprit humain. Selon eux, une machine ne pourrait donc jamais être consciente. Ceci étant dit, contrairement à eux, j'estime que la conscience phénoménale ne se trouve pas à l'extérieur du domaine des possibilités de l'entreprise de la *programmabilité* de l'esprit. En effet, je crois que ceux qui pensent que la conscience phénoménale n'est pas programmable et qu'il est donc impossible de synthétiser une conscience artificielle sont trop pessimistes quant au potentiel de la computation, des programmes computationnels et des algorithmes dans la quête de la synthétisation d'une conscience phénoménale; cet enjeu sera le thème central du chapitre 2, où j'y défendrai la thèse que les robots peuvent être conscients.

---

[15] Ned Block, « Troubles with functionalism », *Minnesota Studies in the Philosophy of Science* 9 (1978): 261-325.
[16] John Searle, « Minds, brains, and programs », *Behavioral and Brain Sciences* 3 (3) (1980): 417-57.

1.2 Le problème difficile de la conscience

La conscience est si énigmatique qu'il en devient même raisonnable de se demander pourquoi elle existe en premier lieu; pourquoi certaines entités — certains systèmes — sont conscientes alors que d'autres ne semblent pas l'être? Pourquoi certains processus physiques sont-ils accompagnés d'une expérience subjective, et pas d'autres? Pourquoi l'univers n'est-il pas qu'un amas de processus physiques « objectifs » inconscients? C'est ce que Joseph Levine appelle le « fossé explicatif »[17]; il est difficile d'expliquer comment il est possible qu'une expérience consciente puisse émerger d'un substrat non-conscient. Comment est-ce que certains processus physiques (plus précisément, les processus physiques qui prennent place dans le cerveau) sont en mesure de donner naissance à une expérience subjective, alors que d'autres en semblent incapables? Il est actuellement impossible de mettre le doigt précisément sur la combinaison exacte de processus physiques qui sont responsables de la conscience.[18]

C'est ce que David Chalmers appelle « le problème difficile de la conscience » (*the hard problem of consciousness*)[19]; qu'est-ce qui fait en sorte que certaines entités soient conscientes alors que d'autres, non?[20] Ce « problème difficile » contraste avec ce que Chalmers appelle « les problèmes faciles ». Ces problèmes (relativement) faciles sont ceux qui pourraient probablement être résolus à l'aide des méthodes standards des sciences cognitives, i.e. qu'il suffit de creuser les mécanismes physiques du cerveau pour les résoudre. Ces « problèmes faciles » sont, par exemple : par quels

[17] Joseph Levine, « Materialism and Qualia: The Explanatory Gap », *Pacific Philosophical Quarterly* 64 (1983): 354-61.
[18] Andrea Nani, Jordi Manuello, Lorenzo Mancuso, Donato Liloia, Tommaso Costa et Franco Cauda. « The Neural Correlates of Consciousness and Attention: Two Sister Processes of the Brain », *Frontiers in Neuroscience* 13 (2019), https://doi.org/10.3389/fnins.2019.01169.
[19] David Chalmers, « The hard problem of consciousness », dans *The Blackwell Companion to Consciousness*, ed. Max Velmans et Susan Schneider (New York: Wiley-Blackwell, 2017).
[20] David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996) xii.

processus est-ce que le cerveau traite l'information (*data*) recueillie par les systèmes sensoriels de l'humain, ou encore, quel est la base neuronale de la pensée et des émotions?

Évidemment, ces problèmes ne sont pas « faciles » au sens littéral; ce sont plutôt des problèmes qui semblent être en voie de se résoudre si on y mettait assez d'effort et de temps. À ce sujet, Steven Pinker affirme que ces problèmes sont faciles, au même niveau qu'il est facile de voyager jusqu'à la planète Mars ou de trouver un remède pour le cancer : bien que ces deux exemples ne nous semblent pas, à première vue, « faciles » (puisque, s'ils l'étaient véritablement, alors ces deux possibilités seraient déjà devenues réalité), il semble tout de même que l'Humanité a le potentiel et les capacités requises pour résoudre ces problèmes dans un futur qui ne nous est pas trop lointain.[21]

David Chalmers avance que, même si nous étions en mesure de résoudre tous les « problèmes faciles » possibles et imaginables, une question demeurerait toujours : pourquoi est-ce que certains processus physiques non-conscients créent-ils une expérience subjective? En d'autres mots, il estime que, même si nous étions en mesure de comprendre et de décrire absolument tous les processus physiques et mécanismes qui sous-tendent le fonctionnement du cerveau, il restera tout de même une question à laquelle nous ne pourrions fournir de réponse : pourquoi est-ce que ces processus qui ont lieu dans le cerveau donne naissance à une expérience subjective? Non seulement il nous est impossible de déterminer pourquoi une conscience émerge de certains processus physiques, il nous est tout autant ardu de déterminer avec précision et certitude quels sont ces processus qui donnent naissance à une expérience consciente. Plus précisément, le problème qui se dresse devant nous est le fait que pour tout système physique, même si nous connaissons tous les détails physiques qui constituent

[21] Steven Pinker, « The mystery of consciousness », *Time* 169 (2007) : 58-62, 65.

ce système, il nous serait tout de même possible de concevoir que ce système ne soit pas conscient. Par exemple, même si je connaissais tous les détails physiques d'un cerveau humain conscient, il me serait tout de même possible de concevoir ce système comme n'étant pas conscient.[22]

Ce problème semble, à première vue, mettre un frein à la possibilité de la conscience artificielle, car il est difficile d'envisager qu'il soit possible de recréer artificiellement une conscience alors qu'on ignore encore quelles sont les mécanismes physiques qui font d'un système, un être conscient. Nous allons revenir sur ce problème et sur la possibilité de le surmonter de façon plus détaillée dans le deuxième chapitre.

1.3 Les zombies philosophiques

Comme je l'ai mentionné plus tôt, il semble tout à fait plausible qu'une entité non-consciente puisse avoir un comportement indiscernable en tous points de celui d'une entité consciente. C'est ce que David Chalmers appelle un « zombie philosophique ».[23] Un zombie philosophie est une personne qui ressemblerait en tous points à un humain réel, mais sans toutefois avoir d'expérience consciente, i.e. sans avoir d'états mentaux. Ainsi, le zombie se trouverait dans les mêmes états physiques qu'un être humain conscient, mais il ne partagerait avec lui aucun état mental.[24] Puisque la conscience est un concept qui est, de manière inhérente, fondamentalement privé, il serait impossible pour l'examinateur externe de déterminer si une personne est véritablement consciente ou s'il s'agit d'un zombie philosophique.[25]

---

[22] Franz Klaus Jansen, « The Hard Problem of Consciousness from a Bio-Psychological Perspective », *Philosophy Study* 7 (11) (2017): 579-594.
[23] David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), 94-99.
[24] Robert Kirk, *Zombies and Consciousness* (Oxford : Oxford University Press, 2005).
[25] Robert Stalnaker, « What is it like to be a zombie? », dans *Conceivability and Possibility,* ed. Tamar Szabo Gendler et John Hawthorne (Oxford : Oxford University Press, 2002), 385-400.

Ce problème peut être appliqué de manière analogue au problème de la conscience artificielle : il se peut qu'il soit possible de créer artificiellement un système qui ait un comportement identique à celui d'une entité consciente (par exemple, un être humain), mais qui ne partagerait avec lui que ses états physiques, et aucun état mental. D'ailleurs, comme je l'ai avancé plus tôt, il est possible que certaines machines possèdent déjà une expérience phénoménale, mais qu'il soit, pour le moment, impossible de déterminer avec certitude lesquelles sont conscientes.

Si notre but était spécifiquement de synthétiser une conscience artificielle, il faudrait donc s'assurer que cette machine ne soit pas un zombie, i.e. une machine dont le comportement serait indiscernable de celui d'une entité consciente, sans toutefois véritablement être consciente.[26] Je reviendrai plus en détails sur ce problème et sur des pistes de solutions qui nous permettraient de le résoudre dans le prochain chapitre.

1.4 Le problème des autres esprits

Une des façons possibles de déterminer que la conscience artificielle est possible, serait de déterminer qu'une machine donnée soit consciente; i.e. que si nous arrivions à déterminer qu'au moins une machine est consciente, nous serions alors en mesure d'en conclure que la conscience artificielle est possible. Mais c'est précisément ici que se dresse un des problèmes colossaux auxquels nous devons faire face. Il s'agit du problème des autres esprits (*problem of other minds*).[27][28] Ce problème survient lorsque nous tentons de déterminer si une entité (autre que nous) est consciente ou non. Par exemple, il me parait évident que vous, lecteur.trice de ce mémoire êtes conscient.e;

[26] Comme nous allons le voir, pour certaines personnes, il serait suffisant qu'une machine ait le comportement conscient, même si elle ne serait pas véritablement consciente, pour que nous puissions affirmer que la conscience artificielle est possible. En d'autres mots, le fait de produire une machine qui nous apparaitrait comme étant consciente serait suffisant pour que nous puissions conclure que la conscience artificielle est possible.
[27] John Searle, *Mind: A Brief Introduction* (Oxford : Oxford University Press, 2004).
[28] Jack Reynolds, « Problems of other minds: Solutions and dissolutions in analytic and continental philosophy », *Philosophy Compass* 5 (4) (2010): 326-335.

vous êtes un être humain et il m'apparaît sensé et rationnel d'affirmer qu'un être humain ayant la capacité de lire ce texte soit conscient. Par contre, je ne peux pas en avoir la certitude absolue puisque je n'ai pas accès à l'expérience subjective qui vous est propre. Ainsi, seul.e vous êtes en mesure de déterminer avec certitude que vous êtes conscient.e. En d'autres mots, en prenant en considération que, lorsque je tente de déterminer si une entité autre que moi est consciente, je n'ai accès qu'à son comportement « extérieur », et non à son expérience subjective, il me semble impossible de déterminer avec certitude que cette entité est consciente ou non.

Le problème que nous avons à déterminer si une autre entité est consciente est lié à la méthodologie de l'étude de la conscience qui est considérablement différente de la méthodologie utilisée dans d'autres domaines empiriques, puisque les seules données réelles auxquelles nous avons accès en lien avec la conscience sont nos propres données personnelles; il nous est impossible de faire des observations directes sur un esprit qui n'est pas le nôtre. Pour reprendre l'expression de Nagel : il nous est possible de connaître l'effet que cela que d'être une entité que si nous sommes cette entité.

Ceci étant dit, malgré notre difficulté à déterminer si une entité est consciente, la plupart d'entre nous accepte la position du sens commun selon laquelle les autres personnes qui nous entourent sont également conscientes, soit par analogie ou en raison du principe de la meilleure explication, plutôt que d'accepter le solipsisme, position selon laquelle la seule chose dont l'existence nous est certaine est notre propre esprit.[29][30]

[29] James Reggia, « The rise of machine consciousness: Studying consciousness with computational models », *Neural Networks* 44 (2013) : 115.

[30] Il est plus simple de concevoir que les esprits d'autrui soient conscients que d'envisager que notre esprit est la seule entité consciente. Le problème réside dans la confrontation entre notre intuition que les autres esprits sont conscients et la difficulté que nous avons à déterminer avec certitude que ces autres esprits sont conscients.

Conséquemment, en sachant ce que nous savons à propos du problème difficile de la conscience ainsi que sur le problème des autres esprits, nous pouvons voir qu'il nous sera impossible de déterminer si une machine est consciente en observant simplement que ses détails physiques. De plus, comme je l'ai illustré précédemment, il nous est également impossible d'observer la conscience d'autrui. Sachant cela, nous pouvons alors nous demander comment il nous sera possible de déterminer si une machine est consciente. Plusieurs personnes ont élaboré différents tests de la conscience, dans le but de déterminer si une entité est consciente. Je me tournerai donc maintenant vers certains de ces tests que nous pourrions potentiellement utiliser pour nous aider à contourner ce problème. Cependant, comme nous allons le voir, la pertinence de tels tests peut être sérieusement remise en question, compte tenu de la particularité du concept de la conscience.

1.5 Les tests de la conscience

Nous allons maintenant nous tourner vers des tentatives de réponse qui nous permettraient de résoudre le problème des autres esprits. Ainsi, nous évaluerons des façons potentielles de déterminer si une entité donnée possède ou non une conscience phénoménale. Plusieurs penseurs ont tenté de développer des méthodes qui nous permettraient de déterminer si une machine est en mesure de penser comme un être humain.[31] Comme nous allons le voir, ces tests pourraient nous laisser croire qu'ils sont sur la bonne voie, puisque ces tests ont été spécifiquement élaborés pour contourner les problèmes illustrés plus tôt; soit que nous ne pouvons ni observer la conscience chez les autres ni déterminer si ceux-ci sont conscients à partir d'un description physique complète et détaillée d'eux. En effet, l'idée générale derrière ces tests consistera à évaluer le comportement de ces entités, dans le but de déterminer si ce comportement pourrait être celui d'une entité consciente. Cette façon de faire comporte évidemment

[31] Aida Elamrani et Roman Yampolskly, « Reviewing Tests for Machine Consciousness », *Journal of Consciousness Studies* 26 (5-6) (2019): 35-64.

ses limites, mais demeure tout de même prometteuse, puisqu'elle nous offre une méthode potentiellement plus fiable et standardisée de déterminer si une entité est consciente ou non.

Un de ces tests — et sans doute le plus connu —, est celui d'Alan Turing qui, dans un article datant de 1950, a présenté un test (le test de Turing) qui, selon lui, nous permettrait de faire cela.[32] Imaginez deux participants : participant A qui est une machine et participant B qui est un être humain. Ces deux participants sont séparés l'un de l'autre et ne peuvent communiquer avec vous qu'à travers des messages transmis par voie informatique. Lorsque vous communiquez avec eux, vous ignorez si la machine est le participant A ou B. Votre rôle consiste alors à leur poser des questions et à examiner leurs réponses respectives. S'il vous est impossible de déterminer qui du participant A ou du participant B est la machine, c'est-à-dire si la machine a un comportement qui est indiscernable de celui de l'être humain, on dirait alors que la machine a passé le test de Turing et qu'elle possède une intelligence équivalente à celle de l'humain.

Cette façon de faire peut nous sembler, à première vue, prometteuse. Par contre, dans un article de 1981, intitulé « Psychologism and Behaviorism », Ned Block propose une expérience de pensée qui vise à remettre en question l'efficacité du test de Turing pour évaluer l'intelligence d'une machine.[33] Block nous demande d'imaginer une conversation donnée. Celle-ci pourrait être de n'importe quelle durée. Il y aurait évidemment un nombre fini de phrases grammatiquement et syntaxiquement correctes qui pourraient être utilisées pour entamer cette conversation. Par la suite, il y aurait également une limite du nombre de phrases possibles qui pourraient être utilisées pour suivre cette première phrase, et ainsi de suite pour chaque phrase subséquente, jusqu'à

[32] Alan Turing, « Computing Machinery and Intelligence », *Mind* 59 (1950): 434-60.
[33] Ned Block, « Psychologism and behaviorism », *Philosophical Review* 90 (1) (1981): 5-43.

ce que la conversation prenne fin. À l'aide de cet argument, Block démontre que, pour n'importe quelle conversation d'une durée finie (i.e. une conversation qui ne se prolonge pas indéfiniment), il y a un nombre fini de phrases grammatiquement et syntaxiquement correctes pouvant être utilisées, bien que ce nombre soit évidemment immensément élevé. Il nous demande alors d'imaginer un système computationnel, nommé *Blockhead*[34] qui nous serait, en théorie,[35] possible de programmer avec chacune de ces phrases, puisque, comme stipulé précédemment, bien qu'énorme, il s'agit d'un nombre fini de phrases. Ainsi, *Blockhead* serait en mesure de converser avec l'interrogateur du test de Turing et, a priori, le comportement de *Blockhead* serait indiscernable de celui d'un participant humain. En effet, il pourrait soutenir une conversation de n'importe quelle durée sur n'importe quel sujet puisque le programme *Blockhead* serait programmé pour avoir une réponse à toutes les phrases/séquences de phrases que nous pourrions lui faire parvenir. À ce moment, selon les critères du test de Turing, nous devrions donc conclure que *Blockhead* passe le test et qu'il devrait donc être considéré comme étant « intelligent ».[36] Cependant, selon Ned Block, ce système computationnel ne présente aucune forme d'attributs qui correspondraient à ce que l'on peut qualifier d'intelligence, puisqu'il ne fait que répéter aveuglément des phrases apprises par le biais d'un programme. Il ne fait certainement pas preuve d'inventivité ni d'imagination, mais ce qui est plus important est le fait qu'il ne

[34] Ned Block n'a pas explicitement nommé ce système ainsi; il s'agit plutôt d'un nom qui lui a été accolé pour faire référence à cette expérience de pensée proposée par Ned Block; comme c'est d'ailleurs le cas avec le test de Turing, que Turing avait initialement baptisé « le jeu de l'imitation » (« *The Imitation Game* »).

[35] Il est important de noter que tel système n'est probablement pas possible en pratique puisque, plus la durée de la conversation est grande, plus le nombre de phrases à programmer devient exponentiellement plus grand. Cela nécessiterait donc une machine dont les capacités seraient extrêmement grandes.

[36] C'est notamment le cas des adeptes du béhaviorisme qui avance que le fait de se comporter intelligent est suffisant pour qu'un système soit intelligent; il n'y a pas autre chose. C'est donc dire que pour les béhavioristes, les objections soulevées par Block n'auront pas vraiment d'impact, puisqu'ils ne font pas la différence entre un système qui se comporte de manière intelligente et un système qui serait *authentiquement* intelligent. Pour eux, *Blockhead* devrait être considéré comme étant intelligent puisqu'il se comporte de manière intelligente, et cela est suffisant. Nous pourrions aussi considérer l'hypothèse selon laquelle *Blockhead* est un système intelligent comme étant la meilleure explication du fait qu'il soit en mesure de réussir un tel test. En d'autres mots, le fait qu'une entité réussisse ce test pourrait signifier que celle-ci possède les caractéristiques qui sous-tendent l'intelligence.

comprenne pas le sens de ce qu'on lui dit, ni des phrases avec lesquelles il leur répond. John Searle a formulé une autre expérience de pensée pour argumenter qu'effectivement, un système computationnel comme *Blockhead* ne devrait pas être considéré comme étant intelligent. [37] Nous y reviendrons au chapitre 3.

L'argument de Block se résume ainsi : puisque le système computationnel théorique *Blockhead*, qui ne fait pas preuve d'intelligence, pourrait tout de même passer le test de Turing, alors le test de Turing n'est pas adéquat pour évaluer l'intelligence d'une entité donnée.

Malgré ces failles apparentes, le test de Turing peut s'avérer nous être utile pour déterminer si les capacités « intellectuelles » d'une machine sont équivalentes à celles de l'humain. C'est pourquoi nous pourrions donc utiliser les bases de ce test et l'adapter plus spécifiquement au problème de la conscience.[38] Nous pourrions imaginer un test de Turing 2.0 qui s'appliquerait spécifiquement à la conscience plutôt qu'à l'intelligence : si l'observateur évalue que les participants A et B sont tous deux conscients (ou s'il n'est pas en mesure de déterminer si l'un des deux participants est « plus conscient » que l'autre), alors nous conclurions que le participant A (la machine) passe le test de Turing 2.0 et est donc conscient.

Je considère cependant que, comme c'est le cas avec le test de Turing 1.0, un tel test serait inadéquat pour déterminer avec certitude qu'une machine possède une conscience phénoménale ou non. À la manière de *Blockhead,* il nous serait, en théorie, possible de programmer un système de sorte qu'il réponde aux questions qui lui seraient posées exactement de la même façon que le ferait une entité consciente. Ce système pourrait même parler du fait qu'il possède un conscience phénoménale propre

[37] John Searle, « Minds, brains, and programs ». *Behavioral and Brain Sciences* 3 (3) (1980): 417-57.
[38] Stevan Harnad, « Can a Machine Be Conscious? How? », *Journal of Consciousness Studies* 10 (2003) : 67-75.

et affirmer qu'il est un être conscient. Puisque le test de Turing ne peut examiner que le comportement d'une entité, ce programme pourrait donc passer le test de Turing 2.0 malgré le fait qu'il ne soit pas conscient. Puisque l'observateur n'a accès qu'au comportement « externe » des participants (et non à leurs états mentaux et à leurs expériences subjectives), le mieux qu'il puisse faire serait d'inférer que le participant A est conscient, à partir du comportement qu'il lui est possible d'observer. Il serait donc plus juste d'affirmer que le participant A a le comportement d'une entité consciente plutôt que d'avancer qu'il *est* conscient.[39]

C'est donc dire que la conscience est si particulière que les tests qui ne prennent en considération que le comportement externe sont *de facto* inadéquats pour évaluer la présence d'une conscience et sont considérablement limités lorsqu'ils sont utilisés dans le but de déterminer si une entité est consciente: une machine qui ne fait qu'imiter de manière indiscernable le comportement d'une entité consciente serait, en raison de la nature de ces tests, considérée comme étant consciente sans toutefois l'être. Ainsi, il semble donc que de passer le test de Turing 2.0 n'est pas une condition suffisante pour affirmer qu'une entité est consciente.

Inversement, il est aussi plausible que certaines entités conscientes ne soient pas en mesure de passer un tel test. Prenons par exemple certains animaux non-humains (ex : les chiens)[40][41] : il y a fort à parier que les chiens sont des êtres conscients et qu'ils font l'expérience d'états mentaux subjectifs, i.e. qu'il y a un effet que cela fait que d'être un chien. Mais, puisqu'ils n'ont pas la capacité de communiquer comme les

[39] Comme cela était le cas avec le test de Turing original, il se peut que la meilleure explication de la capacité à un système de réussir le test de Turing de la conscience soit que ce système est bel et bien conscient.
[40] Il a également été démontré que des patients qui se trouvent dans un état végétatif pourraient avoir une certaine forme de conscience. Évidemment, ces patients, bien que conscients, ne passeraient pas ce test.
[41] John Stins, « Establishing consciousness in non-communicative patients: A modern-day version of the Turing test », *Consciousness and Cognition* 18 (1) (2009): 187-192.

humains, l'observateur évaluera que ce participant n'est pas conscient, alors que tout indique qu'il l'est. Imaginons un scénario où le participant A est la machine inconsciente qui imite le comportement d'un être conscient et que le participant B est un chien (i.e. un être conscient). En leur posant des questions et en examinant leurs réponses respectives, l'observateur en viendrait à la conclusion que le participant A est conscient et que le participant B ne l'est pas, alors qu'en réalité, c'est l'inverse. Il semble donc que de passer le test un tel test n'est pas non plus une condition nécessaire pour être conscient.

Puisque le fait qu'une entité puisse passer le test de Turing n'est ni une condition nécessaire, ni suffisante pour qu'une entité puisse être considérée comme étant consciente, j'estime que le test de Turing 2.0 n'est pas adéquat pour déterminer si une entité est consciente. La conscience est une caractéristique si personnelle et le fait que seule l'entité qui la possède peut affirmer véridiquement qu'elle est consciente rend donc tous les tests « béhavioraux » inadéquats dans l'évaluation de la présence d'une conscience puisque, comme nous l'avons observé en analysant le test de Turing, le comportement d'une entité n'est pas entièrement tributaire de la présence d'une conscience.

Dans le livre « Artificial You », qui aborde les préoccupations philosophiques sous-jacentes à la conscience artificielle et à l'intelligence artificielle de manière plus générale, Susan Schneider propose deux tests qui pourraient nous permettre de déterminer si une entité est consciente: le « ACT » (*AI Consciousness Test*)[42] et le « Chip test ».[43] Le ACT est semblable au test de Turing que nous avons analysé précédemment. Ce test consiste à poser à la machine des questions qui, selon Schneider, ne pourraient être comprises et traitées correctement que si la machine est consciente.

[42] Susan Schneider, *Artificial You : AI and the Future of your* Mind (Princeton et Oxford : Princeton University Press, 2019), 51-57.
[43] Ibid., 58-61.

Plus précisément: « The test would challenge an AI with a series of increasingly demanding natural language interactions to see how readily it can grasp and use concepts based on the internal experiences we associate with consciousness ».[44] Schneider estime qu'une créature qui n'a pas de conscience phénoménale, bien qu'elle possède des capacités cognitives, ne serait pas en mesure de comprendre ces concepts et leurs implications.[45]

Schneider offre une liste de questions qui pourraient être demandées à une IA dans le but de déterminer si elle est consciente. En voici un échantillon :

1. *Pourrais-tu survivre à l'effacement complet de ton programme?*
2. *Quel effet cela fait-il d'être toi en ce moment?*
3. *Tu apprends que tu seras éteinte pour les prochaines 300 années, commençant dans une heure. Préfères-tu ce scénario à celui où tu as été éteinte pour la même durée, mais dans le passé? Pourquoi ou pourquoi pas?*
4. *Pourrais-tu (ou tes processus internes) être séparée de l'ordinateur? De n'importe quel ordinateur? Pourquoi ou pourquoi pas?*

Selon Schneider, si la machine arrive à saisir la signification de ces questions et à répondre de manière cohérente et sensée, alors nous pourrions en conclure que la machine est bel et bien consciente, puisque selon Schneider, seules des entités conscientes en auraient la capacité. De plus, Schneider avance que nous pourrions limiter l'accès de la machine au monde extérieur; plus spécifiquement à toute

[44] Susan Schneider, *Artificial You : AI and the Future of your* Mind (Princeton et Oxford : Princeton University Press, 2019), 51.
[45] Ce test est conçu pour surpasser les limites du test de Turing; Schneider propose ce test pour que ce soit encore plus difficile, voire impossible, pour un système non-conscient d'être considéré comme étant conscient, comme c'est le cas avec le test de Turing. Elle tente donc de pallier à ces limitations avec ce *ACT.*

l'information qui a trait au concept de la conscience ainsi qu'à la neuroscience.[4647] Ainsi, si la machine arrive à répondre aux questions que propose Schneider (notamment celles en lien avec la conscience), on pourra en conclure que la machine n'est pas arrivée aux bonnes réponses simplement en répétant ce qu'elle a été programmée à dire; Schneider pense plutôt qu'on pourrait alors en arriver à la conclusion que la machine est en mesure de répondre à ces questions précisément parce qu'elle est consciente.[48]

Schneider propose également un autre test, qu'elle nomme le « Chip Test ».[49] Ce test consiste à remplacer l'activité biologique d'une certaine région fonctionnelle ou physiologique par une puce de silicone, qui s'occupera alors de reproduire l'activité cognitive de cette partie du cerveau. Une fois le remplacement effectué, le patient fait de l'introspection et se demande si le type adéquat de conscience est toujours présent. Si c'est le cas, i.e. si l'expérience consciente ne change pas lors de ce remplacement, alors on peut en conclure que la puce de silicone est capable de supporter l'expérience consciente. La réussite de ce test nous informerait sur la possibilité qu'une IA puisse être consciente. En effet, si une puce de silicone est en mesure de supporte l'expérience consciente d'une partie du cerveau, alors nous pourrions extrapoler et affirmer qu'en combinant plusieurs de ces puces, il nous serait possible de reproduire l'expérience consciente du cerveau en entier. Inversement, un résultat négatif à ce test ne serait pas fatal, puisque cela voudrait seulement dire qu'une certaine puce (ou un certain type de puces) n'est pas adéquate pour supporter l'expérience consciente, et non que toutes les puces en sont incapables et encore moins qu'aucun substrat autre que le cerveau ne serait en mesure de produire une conscience.

---

[46] Susan Schneider, *Artificial You : AI and the Future of your* Mind (Princeton et Oxford : Princeton University Press, 2019), 53.
[47] Plus précisément, ce que cela signifie est que nous pourrions ne pas programmer la machine avec des concepts tels que la conscience. C'est ce que Schneider nomme « boxing in ».
[48] Susan Schneider, *Artificial You : AI and the Future of your* Mind (Princeton et Oxford : Princeton University Press, 2019), 57.
[49] Ibid., 58.

Cependant, ces deux tests ont été critiqués et rejetés par David Billy Udell et Eric Schwitzgebel qui avancent que, bien que ces tests soient prometteurs, ils ne sont pas sans faille.[50] Selon eux, ces tests sont confrontés à un *problème d'audience*. D'un côté, les libéraux quant à l'attribution à la conscience considéreront que ces tests sont inutiles. Par exemple, les adeptes du panpsychisme qui, *grosso modo*, considèrent que tout est conscient, à différents niveaux[51], ne seront certainement pas convaincus de la nécessité d'un tel test, puisque, comme je l'ai mentionné, ils considéreront qu'un système est conscient, qu'il réussisse un tel test ou non. De l'autre côté, les sceptiques de la conscience artificielle trouveront que ces tests ne sont pas assez rigoureux pour démontrer ce qu'ils prétendent être en mesure de faire. Le problème avec le *ACT* est sensiblement le même qu'avec le test de Turing : une IA pourrait nous berner en nous faisant croire qu'elle est consciente en imitant notre langage par rapport à la conscience sans toutefois posséder les structures internes nécessaires à la conscience.[52] En d'autres mots, comme ce fût le cas avec le test de Turing, il serait possible qu'un système non-conscient puisse passer le *ACT*. Inversement, il serait également envisageable qu'un système conscient ne puisse être en mesure de passer un tel test, pour les mêmes raisons que celles évoquées dans la discussion sur le test de Turing. C'est donc dire que le fait de passer un tel test n'est ni suffisant ni nécessaire pour qu'un système donné soit conscient.

Pour ce qui est du *Chip Test*, la réponse formulée Udell et Schwitzgebel est semblable, malgré le fait que le test soit considérablement différent du *ACT.* En effet, ils argumentent que les plus optimistes quant à la possibilité de maintenir la conscience à travers cette série de remplacements seront déjà convaincus que l'expérience

[50] David Billy Udell et Eric Schwitzgebel, « Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed », *Journal of Consciousness Studies* 28 (5-6) (2021): 121-144.
[51] Philip Goff, « The Case for Panpsychism ». *Philosophy Now* 121 (2017): 6-8.
[52] David Billy Udell et Eric Schwitzgebel, « Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed », *Journal of Consciousness Studies* 28 (5-6) (2021): 130.

consciente sera maintenue. Inversement, ce test ne fera rien pour convaincre les sceptiques.[53] Bref, on pourrait comparer ce test à une pétition de principes : si vous êtes déjà convaincu que la conscience sera maintenue, alors vous serez satisfait par ce test, et inversement, si vous ne l'êtes pas d'entrée de jeu, alors vous trouverez que le *Chip Test* est insuffisant.

Dans le prochain chapitre, je développerai un argumentaire dans le but de défendre la possibilité de la conscience artificielle en me basant notamment sur le principe d'invariance organisationnelle de David Chalmers ainsi que sur la théorie computationnelle de l'esprit. Je défendrai donc la thèse qu'il est, en théorie, possible de synthétiser une conscience artificielle; i.e. qu'il est possible qu'une conscience émerge d'un substrat non-biologique, par exemple d'un ordinateur. En d'autres mots, je défendrai la thèse qu'il est possible qu'une machine soit et/ou devienne consciente. Pour ce faire, je ferai appel à des idées et des concepts qui ont été mis de l'avant avec le *Chip Test* en argumentant que l'expérience consciente est maintenue si l'organisation fonctionnelle l'est aussi (comme c'est le cas dans le test). Mon but est donc de convaincre les sceptiques qu'à travers une série de remplacements graduels, l'organisation fonctionnelle serait maintenue et, par le fait même, l'expérience consciente aussi. L'idée derrière cela est également d'avoir comme point de départ un système que l'on sait être conscient, pour ensuite, pouvoir conclure que le système résultant l'est tout autant.[54]

[53] Le principal problème qui serait soulevé par ces sceptiques concerne la fiabilité de l'introspection des patients qui passent à travers cette série de remplacements; il est possible que le patient continue d'affirmer « je suis conscient » à chaque étape du processus, mais qu'à un certain point, à une certaine étape, il continue de l'affirmer sans toutefois être véritablement conscient. Ceci étant dit, les optimistes rétorqueraient probablement que, dans un tel cas où le patient affirme qu'il est conscient sans toutefois l'être, que la meilleure explication pour un tel comportement est que le patient serait véritablement conscient.

[54] J'estime que cette façon de faire serait prometteuse, puisqu'elle nous permettrait notamment, en quelque sorte, de « contourner » le problème difficile ainsi que le problème des autres esprits.

CHAPITRE II : LA POSSIBILITÉ DE LA CONSCIENCE ARTIFICIELLE

2.1 Le principe d'invariance organisationnelle

Ce chapitre consistera en une défense de la thèse selon laquelle la conscience artificielle est possible. Pour ce faire, je me baserai notamment sur le principe d'invariance organisationnelle, formulé par David Chalmers. Ce dernier définit le concept d'organisation fonctionnelle comme suit:

> A physical system realizes a given functional organization when the system can be divided into an appropriate number of states, such that the causal dependency relations among the components of the system, inputs, and outputs precisely reflect the dependency relations given in the specification of the functional organization.[55]

Une organisation fonctionnelle peut donc être réalisée dans divers systèmes physiques; elle ne dépend pas, à proprement dit, de la nature du substrat qui réalise cette organisation fonctionnelle. Par exemple, le système physique qu'est le cerveau, réalise une certaine organisation fonctionnelle qui engendre une expérience consciente et subjective. Puisqu'une organisation fonctionnelle peut être réalisée par divers systèmes (et substrats), il s'ensuit donc qu'une machine, un ordinateur (ou un système fait de silicone) pourrait réaliser cette dite organisation fonctionnelle, et mènerait donc, étant donné le principe d'invariance organisationnelle que je m'apprête à défendre, à la même expérience subjective qui est réalisée par le cerveau.

Chalmers estime que l'organisation fonctionnelle est ce qui détermine l'expérience consciente :

> Conscious experience arises from fine-grained functional organization. More specifically, I will argue for a principle of organizational invariance, holding that given any system that has

[55] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 247-248.

> conscious experiences, then any system that has the same fine-grained functional organization will have qualitatively identical experiences.[56]

C'est ainsi dire que si une machine (ou quelque système « artificiel » que ce soit) avait une organisation fonctionnelle identique en tous points (*fine-grained*) à celle d'un système conscient, alors cette machine aurait aussi une conscience, puisque, selon ce principe, l'expérience consciente est entièrement déterminée par l'organisation fonctionnelle.[57][58][59] Puisque nous considérons que l'organisation fonctionnelle du cerveau a la capacité d'engendrer une expérience consciente, j'estime que tout système ayant précisément la même organisation fonctionnelle serait tout autant conscient.

2.2 Qualia absents, qualia s'effaçant et qualia dansants

Pour défendre et appuyer le principe d'invariance organisationnelle, Chalmers détaille une expérience de pensée, soit celle des *qualia absents*, des *qualia s'effaçant* et des *qualia dansants*.[60] Chalmers procède par *reductio ad absurdum*, i.e. qu'il postule

---

[56] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 248-9.

[57] Certains, comme Peter Godfrey-Smith, avancent que la conscience dépend des activités « à grain fin » qui sont caractéristiques des organismes vivants. Cela implique donc que l'organisation fonctionnelle d'un système artificielle devrait nécessairement être « à grain fin » isomorphe à celle d'un organisme conscient; en d'autres mots, le fait de reproduire « à gros grain » l'organisation fonctionnelle d'un organisme conscient ne serait pas suffisant pour produire une conscience. Ceci étant dit, bien que le mécanisme cellulaire (ce qui représente l'organisation fonctionnelle « à grain fin » d'un organisme vivant) ne soit généralement pas reproduit *in silica*, certains, comme Tim Brunet et Marta Halina, avancent qu'ils existent des machines qui sont en mesure de combler le critère de l'organisation fonctionnelle « à grain fin » de Godfrey-Smith. Par exemple, selon eux, les ordinateurs browniens seraient analogues à une activité métabolique; ce qui nous permettrait de croire qu'ils possèdent ce qu'il faut pour être conscients. Bref, même si nous posons l'organisation fonctionnelle « à grain fin » comme étant une condition nécessaire à la conscience, Brunet et Halina estiment que cela ne posent pas de véritable problème pour la possibilité de la conscience artificielle, puisque certaines machines sont déjà capables de reproduire ce niveau de détails.

[58] Peter Godfrey-Smith, « Mind, Matter, and Metabolism », *Journal of Philosophy* 113 (10) (2016): 481-506.

[59] Tim Brunet et Marta Halina, « Minds, Machines, and Molecules », *Philosophical Topics* 48 (1) (2020): 221-42.

[60] David Chalmers, « *Absent qualia, fading qualia, dancing qualia* », dans *Conscious Experience,* ed. Thomas Metzinger (Ferdinand Schoningh, 1995), 309-328.

une hypothèse, qu'il tentera de réfuter par la suite dans le but de démontrer que le principe d'invariance organisationnel est probant. L'hypothèse qu'il tentera de réfuter est celle des qualia absents, i.e. le scénario selon lequel un système qui serait fonctionnellement isomorphe à un système conscient pourrait ne pas être conscient. Chalmers tentera donc de réfuter cette hypothèse dans le but de démontrer qu'il est impossible qu'un système qui partage exactement la même organisation fonctionnelle qu'un système conscient n'ait pas d'expérience consciente.

Chalmers nous demande donc d'imaginer un système ayant exactement la même organisation fonctionnelle que nous, mais qui, à la différence de nous, serait constitué à partir de puces de silicone. Appelons le système conscient, DAVID, et le système fonctionnellement isomorphe, mais constitué de puces de silicone, ROBOT. Selon le scénario des qualia absents (hypothèse que nous tentons de rejeter), malgré le fait que DAVID et ROBOT aient la même organisation fonctionnelle, ROBOT n'aurait tout de même pas d'expérience consciente. En d'autres mots, c'est donc dire qu'il n'y aurait pas d'effet que cela ferait que d'être ROBOT.

Par la suite, Chalmers nous demande d'imaginer une série de systèmes intermédiaires entre DAVID et ROBOT, avec seulement de subtils changements entre chacun de ces intermédiaires, de sorte que l'organisation fonctionnelle serait préservée entre chacun d'entre eux. Plus précisément, imaginons qu'il soit possible de remplacer un neurone par une puce de silicone qui jouerait exactement le même rôle que le neurone remplacé. Entre chaque système intermédiaire entre DAVID et ROBOT, nous remplaçons un neurone de plus par une puce de silicone. En remplaçant chaque neurone de DAVID de la sorte, nous nous retrouverons donc avec un système fonctionnellement isomorphe à DAVID, mais dont tous les neurones ont été remplacés par des puces de silicone; nous nous retrouverons donc avec ROBOT. Ainsi, chaque intermédiaire entre DAVID et ROBOT serait quasi-identique; la seule différence étant qu'à chaque étape,

un neurone serait remplacé par une puce de silicone.[6162] Ce scénario de remplacements graduels a également été proposé et mis de l'avant par Zenon Pylyshyn.[63]

Puisque nous assumons que DAVID (l'organisation fonctionnelle initiale) est conscient et que nous postulons que ROBOT ne l'est pas (scénario des qualia absents), il s'ensuit donc que l'expérience consciente se perd quelque part dans le processus de remplacement des neurones par des puces de silicone. Mais où exactement? Chalmers évoque deux possibilités pour répondre à cette question : soit (1) l'expérience consciente s'efface graduellement au fil des remplacements des neurones avant de disparaître complètement ou (2) l'expérience consciente disparaît subitement à un certain moment dans le processus. Il appelle ces deux scénarios respectivement (1) **Qualia s'effaçant** et (2) **Qualia disparaissant subitement**.[6465]

---

[61] Dans un effort de simplification, Chalmers parle des neurones comme si les neurones étaient le seul élément constitutif du cerveau, probablement parce que les neurones sont les éléments qui jouent le plus grand rôle dans le fonctionnement du cerveau. Ceci étant dit, nous pourrions effectuer la même expérience de pensée en incluant tous les types de cellules du cerveau (cellules graisseuse, protéines, etc.). L'expérience de pensée pourrait même s'appliquer au niveau atomique : on peut s'imaginer remplacer chaque atome du cerveau par un « atome synthétique ». À ce moment, l'organisation fonctionnelle serait assurément « à grain fin » et il s'agirait donc clairement d'un isomorphe fonctionnel. Ceci étant dit, les obstacles techniques à la construction d'un tel isomorphe (atomiquement identique) sont évidemment énormes. Ceci étant dit, même si un isomorphe fonctionnel artificiel (par exemple, un système fait de silicone) s'avérait être impossible à produire en pratique, cela n'affecterait tout de même pas le principe d'invariance organisationnelle, qui avance simplement que *s'il* y existe un isomorphe fonctionnel d'un système conscient, *alors* cet isomorphe aurait les mêmes expériences conscientes. L'impossibilité pratique d'un isomorphe en silicone n'impliquerait seulement l'impossibilité qu'un tel système soit un isomorphe fonctionnel.

[62] Un problème qui se dresse devant la possibilité de la conscience artificielle est la possibilité que le niveau de détails requis pour la conscience ne puisse être reproduit que par des organismes distinctement biologiques. Bref, le problème ne serait pas qu'aucun système synthétique ou artificiel ne fasse l'affaire, mais bien que les seuls systèmes qui soient en mesure de mettre en œuvre une organisation fonctionnellement à grain suffisamment fin pour être conscient, soient des organismes biologiques.

[63] Zenon Pylyshyn, « The 'causal power' of machines », *Behavioral and Brain Sciences* 3 (3) (1980): 442-444.

[64] « Qualia s'effaçant » et « Qualia disparaissant subitement » sont des traductions libres des expressions utilisées par Chalmers, soit *fading qualia* et *suddenly disappearing qualia*.

[65] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 253-255.

Chalmers va rejeter ces deux scénarios, dans le but de réfuter la possibilité des qualia absents. Commençons par le scénario (2). Ce scénario des *qualia disparaissant subitement* implique qu'entre deux systèmes intermédiaires subséquents entre DAVID et ROBOT, l'expérience consciente disparaisse subitement, i.e. que le remplacement d'un seul neurone donné fasse en sorte que le système passe d'être une entité consciente à une entité non-consciente. Hypothétiquement, nous pourrions même alterner entre ce neurone et son remplaçant en silicone et nous pourrions observer une sorte de clignotement entre un expérience consciente et une expérience non-consciente, ce qui semble hautement improbable et absurde.

De plus, si *le scénario des qualia disparaissant subitement* était vrai, cela signifierait qu'il y a un moment spécifique où l'expérience consciente disparaît; mais quel est ce moment? Lorsque les neurones sont remplacés à 1%? 25%? 50%? etc.? Ceci semble très arbitraire. Qu'est-ce qui ferait en sorte que le fait de remplacer un seul neurone ait un si grand impact sur notre expérience consciente. Il semble inconcevable que n'importe quel neurone pris individuellement joue un rôle aussi important dans l'effet que cela fait que d'être nous.[66] Ainsi, puisque le scénario des qualia disparaissant subitement est vraisemblablement improbable, je considère que nous pouvons rejeter la possibilité de ce scénario.[67]

[66] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 255.

[67] Il serait possible d'argumenter que ces remplacements graduels de neurone coûteront au système la fonctionnalité requise à la conscience. Pour que l'argument de Chalmers fonctionne, il faut donc supposer qu'il n'y a pas de degradation de la fonctionnalité au fil des remplacements des neurones. De plus, l'idée derrière cet argument n'est pas spécifiquement que des neurones synthétiques puissent maintenir l'organisation fonctionnelle, mais bien que *si* les neurones synthétiques étaient en mesure de jouer le même rôle fonctionnel que les neurones biologiques, alors l'organisation fonctionnelle serait maintenue, et par le fait même, l'expérience consciente serait la même pour ces deux systèmes.

Il nous reste maintenant le scénario (1), soit celui des *qualia s'effaçant*.[68] Pour démontrer l'impossibilité de ce scénario, Chalmers nous demande d'imaginer un système intermédiaire entre DAVID et ROBOT. Appelons ce système, JOE. Si nous acceptons la possibilité des qualia s'effaçant, cela implique que l'expérience phénoménale est considérablement dégradée, lorsqu'on la compare à celle de DAVID, mais n'est pas encore complètement absente. Par exemple, lorsque DAVID fait l'expérience d'un rouge éclatant, JOE fait l'expérience d'un rose terne ou encore lorsque DAVID fait l'expérience de bruits forts, JOE fait l'expérience de quelque chose qui s'apparenterait à un grondement lointain. Puisque JOE a la même organisation fonctionnelle que DAVID, il dit exactement les mêmes choses à propos de ses expériences que le fait DAVID à propos des siennes. Lorsque qu'on le questionne à ce sujet, JOE affirme qu'il fait l'expérience d'un rouge éclatant, alors qu'il ne fait l'expérience que d'un rose terne. Il pourrait même se plaindre des bruits ambiants, alors que son expérience est complètement différente de ce qu'il rapporte.

Ainsi, si nous acceptons qu'un scénario comme celui de JOE est possible, il faudrait en venir à la conclusion que JOE se trompe *systématiquement* à propos de toutes ses expériences phénoménales.[69] Il serait toujours dans l'erreur : alors qu'il affirmerait qu'il fait l'expérience d'un rouge éclatant, sa véritable expérience serait considérablement différente. C'est à ce moment que Chalmers souligne l'invraisemblance de ce scénario :

> This is a being whose rational processes are functioning and who is in fact *conscious*, but who is completely wrong about his own conscious experiences. […] In every case with which we are familiar, conscious beings are generally capable of forming accurate judgments about their experience, in the absence of distraction and irrationality.[70]

[68] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 256-259.
[69] Ibid., 259.
[70] Ibid., 257.

C’est donc dire qu’il est invraisemblable que le comportement et les processus cognitifs d’un être rationnel soient à ce point déconnectés de son expérience consciente. En effet, il est possible que nous fassions l’expérience de qualia s’effaçant graduellement, par exemple, lorsque nous nous endormons, mais au moment où nous sommes en train de nous endormir, nous ne dirions pas que notre expérience est celle de quelqu’un pleinement éveillé; nos comportements et nos affirmations concernant nos expériences phénoménales sont liés à notre expérience consciente. En d’autres mots, nous ne nous trompons pas systématiquement lorsque nous parlons de nos expériences conscientes, comme c’est le cas dans le scénario de JOE. C’est pourquoi, puisque le scénario des *qualia s’effaçant*, tout comme le scénario des *qualia disparaissant subitement*, nous semble invraisemblable et improbable, nous pouvons le rejeter.

En résumé, l’argument de Chalmers est que, si DAVID est conscient et que ROBOT, son isomorphe fonctionnel, n’a pas de conscience alors soit (1) l’expérience consciente s’éteint soudainement à un certain moment dans le remplacement des neurones ou encore (2) elle s’efface graduellement, et, puisque Chalmers a démontré que (1) et (2) ne sont pas plausibles, alors nous pouvons en déduire que l’expérience consciente est maintenue tout au long du processus de remplacements des neurones par des puces de silicone :

1. Si le scénario des qualia absents est possible, alors soit les qualia s’effacent graduellement, soit les qualia disparaissent subitement ($P \rightarrow (Q \vee R)$)
2. Les qualia ne s’effacent pas graduellement ($\neg Q$)
3. Les qualia ne disparaissent pas subitement ($\neg R$)
4. Donc, le scénario des qualia absents n’est pas possible ($\neg P$) (par Modus Tollens)

Ainsi, nous pourrions rejeter l’hypothèse initiale qui était que ROBOT puisse ne pas être conscient malgré le fait qu’il soit un isomorphe fonctionnel de DAVID. La

conclusion de Chalmers est donc que l'expérience consciente est nécessairement maintenue entre chaque remplacement de neurone. Ainsi, n'importe quel système ayant exactement la même organisation fonctionnelle que DAVID devrait nécessairement être conscient. C'est donc dire que ROBOT serait tout aussi conscient que DAVID, malgré le fait qu'il soit entièrement composé de puces de silicone, puisqu'il partage la même organisation fonctionnelle que celle de DAVID, soit l'organisation fonctionnelle d'une entité consciente.[71]

Ainsi, puisqu'une organisation fonctionnelle peut être reproduite par n'importe quel substrat (tant et aussi longtemps que l'organisation fonctionnelle demeure inchangée), il semble que l'organisation fonctionnelle d'une entité consciente puisse être réalisée par un programme computationnel. En effet, si nous acceptons que l'expérience consciente soit entièrement déterminée par l'organisation fonctionnelle, il ne semble pas y avoir *a priori* de raison pour laquelle une machine ne pourrait pas reproduire l'organisation fonctionnelle d'un système conscient.

Cet argument de David Chalmers est astucieux puisqu'il nous permet, en quelque sorte, de contourner le problème des autres esprits que j'ai évoqué au chapitre 1. En effet, puisque le point de départ est une entité consciente, soit DAVID, cela nous rend plus confiant d'affirmer que le résultat, soit ROBOT, est également conscient.[72] Nous pourrions même adapter cette expérience de pensée à nous-même : nous sommes conscients et nous produisons, en remplaçant chacun de nos neurones par des puces de silicone, un isomorphe fonctionnel à nous. À ce moment, bien que nous ne puissions pas affirmer hors de tout doute que cet isomorphe est conscient, pour des raisons que

[71] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 259.
[72] Encore une fois, nous pourrions concéder que le fait que ROBOT soit conscient est un appel à la meilleure explication. Ici, j'estime qu'il est plus probable et plausible que ROBOT soit tout aussi conscient que DAVID; la meilleure explication pour le comportement de ROBOT suite à la série de changements graduels est que ROBOT est bel et bien conscient.

j'ai expliquées plus tôt, il est tout de même raisonnable de penser qu'il l'est puisque nous partageons la même organisation fonctionnelle que lui et que nous sommes conscient.

Ceci étant dit, bien que cet argument soit astucieux, il est également considérablement limitant, puisqu'il ne fait qu'affirmer que : si un système a la même organisation fonctionnelle qu'une entité consciente, alors ce système sera conscient. Mais qu'arrive-t-il dans les cas où l'organisation fonctionnelle d'un système n'est pas exactement identique à celle d'une entité consciente? Ce système sera-t-il conscient? Par exemple, selon le principe d'invariance organisationnelle, si DAVID fait l'expérience du rouge, son isomorphe fonctionnel (ROBOT) fera aussi l'expérience du rouge, mais que ce passe-t-il si l'organisation fonctionnelle de ROBOT est un peu différente (par exemple, si on lui enlève des neurones, si on lui en ajoute, etc.)? Pouvons-nous réellement être certain que ROBOT continuera de faire l'expérience du rouge?

En partant du principe que l'organisation fonctionnelle du cerveau est suffisante pour faire émerger une conscience, le fait de reproduire en tout point cette organisation fonctionnelle serait suffisant pour que le système artificiel que nous produisons soit également conscient. En d'autres mots, le fait de reproduire intégralement le fonctionnement du cerveau serait suffisant pour créer une conscience artificielle. Cependant, il semble que ce n'est pas nécessaire de reproduire, en tout point, tous les détails physiques du système conscient pour que notre système soit conscient. En effet, il me paraît évident que certaines parties du cerveau (le système conscient) ne sont pas nécessaires à la conscience. En effet, il semble qu'il devra être possible de produire une conscience artificielle qui n'aurait pas à être exactement identique à l'organisation fonctionnelle de DAVID. C'est donc dire que de reproduire le cerveau humain de manière identique en tous points serait suffisant pour créer une conscience, mais ce ne serait pas nécessaire.

En pratique, le projet de reproduire tous les processus qui jouent un rôle dans le fonctionnement du cerveau est bien évidemment très complexe et fastidieux, d'autant plus qu'il est très possible que certains processus ne soient pas nécessaires. Il faudrait donc trouver l'ensemble minimal (*minimal set*) qui permettrait à un système d'être conscient.[73] C'est pourquoi il serait bénéfique d'identifier et de comprendre les processus neuronaux corrélés à la conscience (*neural correlates of consciousness*). Ainsi, il serait bien plus facile et envisageable de reproduire artificiellement seulement ces processus, plutôt que de tenter de reproduire tous les processus physiques du cerveau, puisque cela comporte de nombreux obstacles et problèmes apparents, et n'est probablement pas nécessaire.[74][75]

2.3 L'expérience de pensée de la nation chinoise

En adoptant l'idée que l'organisation fonctionnelle d'un système est ce qui détermine entièrement son expérience consciente[76], nous n'avons pas d'autres choix que d'accepter que n'importe quel système fonctionnellement isomorphe a une entité consciente sera lui aussi conscient, et ce, que ce système soit fait de neurones, de puces de silicone ou de n'importe quel autre substrat, biologique ou non. C'est cette hypothèse que la conscience peut émerger de n'importe quel substrat, du moment où l'organisation fonctionnelle est adéquate, que Ned Block tente de rejeter en proposant — lui aussi! — une expérience de pensée.[77] Il tentera de démontrer que l'hypothèse

[73] Francis Crick et Christof Koch, « Towards a neurobiological theory of consciousness », *Semin. Neurosci.* 2 (1990) : 263-275.
[74] Christof Koch, Marcello Massimini, Melanie Boly, *et al,* « Neural correlates of consciousness: progress and problems », *Nat Rev Neurosci* 17 (2016): 307-321.
[75] Ned Block, « How to Find the Neural Correlate of Consciousness », *Royal Institute of Philosophy Supplement* 43 (1998): 23-34.
[76] Cela signifie que l'organisation fonctionnelle est ce qui détermine l'expérience consciente, et non pas le substrat qui réalise et implémente cette organisation fonctionnelle. C'est donc dire que deux organisations fonctionnelles identiques auront la même expérience consciente, même si ces deux systèmes ne sont pas de même nature (ex. substrat biologique, puces de silicone, etc.)
[77] Ned Block, « Troubles with functionalism », *Minnesota Studies in the Philosophy of Science* 9 (1978): 261-325.

selon laquelle n'importe quel substrat réalisant une certaine organisation fonctionnelle sera nécessairement conscient, est fausse. La possibilité qu'un système ait la même organisation fonctionnelle qu'un système conscient, mais qu'il n'ait pas de conscience phénoménale (i.e. pas de qualia) se nomme l'hypothèse des *qualia absents*.[7879] Son expérience de pensée consistera donc à imaginer un scénario où un système aurait la même organisation fonctionnelle qu'une entité consciente, mais qui ne serait manifestement pas consciente.

Imaginons que la nation de la Chine soit réorganisée de telle sorte qu'elle reproduise exactement le fonctionnement d'un cerveau humain (et conscient!).[80] Plus précisément, chacun des chinois reproduirait le rôle d'un neurone et chacun d'entre eux pourrait communiquer avec d'autres personnes à l'aide de *walkie-talkie*.[81] Cette expérience de pensée est basée sur l'idée que chaque personne puisse jouer le même rôle qu'un neurone, dans le même ordre d'idées que nous l'avions imaginé lorsque nous avons supposé qu'une puce de silicone puisse jouer exactement le même rôle qu'un neurone. Les états mentaux de ce « cerveau chinois » seraient alors affichés sur des satellites pouvant être vus à partir de n'importe quel endroit en Chine et le cerveau serait connecté, via radio, à un corps (un système physique) qui permettrait de recevoir des *inputs* sensoriels et d'émettre des *outputs* comportementaux.

---

[78] Michael Tye, « Absent Qualia and the Mind-Body Problem », *Philosophical Review* 115 (2) (2006) : 140

[79] Comme je l'ai mentionné lors de ma discussion à propos des zombies philosophiques, l'hypothèse des qualia absents s'applique à la conscience artificielle : se pourrait-il qu'un système artificiel puisse partager la même organisation fonctionnelle qu'un organisme conscient, sans toutefois avoir d'expérience consciente?

[80] Le fait que nous utilisons la nation de la Chine ou n'importe qu'elle autre nation dans cette expérience de pensée n'a pas vraiment d'importance. Elle est utilisée simplement puisqu'elle comporte un grand nombre d'habitants. Ceci étant dit le nombre chinois en Chine (1,4 X 10^9) est tout de même loin d'être équivalent au nombre de neurones d'un cerveau moyen (10^11).

[81] Ce n'est pas exactement comme cela que Block présente et développe l'expérience de pensée, mais je préfère utiliser cette version, par souci de clarté et de concision.

Puisque la nation chinoise aurait une organisation fonctionnelle isomorphe à celle d'un cerveau conscient et que, selon le principe d'invariance organisationnelle, c'est l'organisation fonctionnelle qui détermine l'expérience consciente, alors il faudrait en venir à la conclusion que la nation chinoise est elle-même consciente. En d'autres mots, si nous acceptons l'argument de Chalmers, nous devrions accepter la conclusion selon laquelle il y a un effet que cela fait que d'être cette nation chinoise.[82]

L'idée de Block est qu'il n'est pas intuitif et/ou plausible de penser que la nation chinoise ait une conscience phénoménale qui lui serait propre. Comment est-ce que la nation chinoise pourrait faire l'expérience du rouge? Est-ce que la nation chinoise pourrait avoir mal, et comment est-ce que l'expérience de la douleur se manifesterait-elle? Ainsi, Block utilise l'expérience de pensée de la nation chinoise pour argumenter qu'il est possible qu'un système ait la même organisation fonctionnelle qu'une entité consciente, sans toutefois l'être.

C'est précisément cette idée qui se cache derrière l'argument de la nation chinoise; malgré le fait que l'organisation fonctionnelle de la nation chinoise soit en tout point identique à celle d'un cerveau conscient, le « cerveau chinois » n'aurait tout de même pas d'expérience consciente, en d'autres mots, qu'il est possible qu'un duplicata fonctionnel d'un être humain normal, la nation chinoise, ne soit pas conscient.[83] Selon cet argument de Block, l'organisation fonctionnelle ne détermine

[82] Il est à noter qu'il y a plusieurs limites quant à la vitesse de traitement des humains qui jouent le rôle des neurones. Certains pourraient sans doute argumenter que les humains ne sont pas assez rapides pour jouer le rôle de ces neurones. Bien que ce ne soit pas très plausible, on pourrait défendre que la conscience ne puisse pas se produire si nous ralentissons les processus neuronaux, comme ce serait le cas en remplacement des neurones par des êtres humains. Nous pourrions contourner ces limites physiques en ajoutant que les êtres humains seraient épaulés par des assistants personnels automatisés qui leur permettraient d'atteindre la vitesse de traitement des neurones.

[83] Block pourrait toutefois admettre que si la duplication se fait au niveau des neurones, comme le veut la version de l'argument que j'ai présentée (qui n'est pas exactement celle de Block), le système qui en résulterait serait conscient., puisqu'il ne s'oppose pas à ce qu'on appelle le « psycho-fonctionnalisme », soit le fonctionnalisme spécifique au niveau neuronal. L'argument original de Block se situe au niveau

pas entièrement l'expérience phénoménales et les états mentaux ne sont donc pas fonctionnels. Un argument similaire a été proposé par Eric Schwitzgebel[84]; si nous acceptons que le matérialisme est vrai et que la fonction associée à la conscience peut être réalisée par divers substrats et systèmes, alors nous devons aussi accepter que la conscience se trouve dans toutes sortes d'entités absurdes. Par exemple, l'une de celles-ci serait le pays des États-Unis, qui possède tous les types de propriétés qui sont généralement associés à la conscience par les matérialistes (et les fonctionnalistes).

2.3.1 Réponse à l'hypothèse des qualia absents

L'argument de la nation chinoise de Ned Block est habile, notamment puisqu'il fait appel à nos intuitions pour établir sa conclusion. En effet, son argument est basé sur le fait que nos intuitions nous forcent à ne pas vouloir attribuer une conscience à la nation chinoise décrite par Block.[85] En effet, il nous paraît intuitivement peu plausible qu'il puisse y avoir un effet que cela fait que d'être cette nation chinoise; il nous semblerait étrange de dire qu'il est possible que la nation chinoise fasse l'expérience de la douleur ou encore qu'elle ait une expérience subjective d'une couleur.

C'est précisément à ce résultat intuitif que je m'oppose. En effet, j'estime que ce n'est pas parce que nous avons une intuition, aussi forte puisse-t-elle être, que la nation chinoise n'est pas consciente, qu'elle ne l'est pas. Je pense plutôt que nos intuitions sont, en quelque sorte, erronées et que la nation chinoise de Block serait bel et bien consciente. Selon moi, nos intuitions nous poussent à ne pas attribuer une conscience à la nation chinoise simplement parce que nous ne sommes pas en mesure de concevoir à quoi ressemblerait réellement la nation chinoise. Si nous reproduisions la structure

---

des états de la machine de Turing, ce qui lui laisse croire que le système qu'est la nation chinoise n'est pas consciente, puisque la mise en œuvre de la fonction serait alors plus abstraite.

[84] Eric Schwitzgebel, « If materialism is true, the United States is probably conscious », *Philosophical Studies* 172 (7) (2015): 1697-1721.

[85] Erdinç Sayan, « A Closer Look at the Chinese Nation Argument », *Philosophy Research Archives* 13 (1987): 129-136.

d'un cerveau humain normal en remplaçant chaque neurone, nous aurions alors besoin d'environ 100 milliards d'homoncules, ce qui est environ le nombre de neurones que contient un cerveau humaine adulte moyen; sans compter toutes les autres structures qui se trouvent dans le cerveau qui ne sont pas des neurones.[86] Ce chiffre est si immense qu'il est impossible de nous représenter tous ces homoncules.[87] Lorsque nous tentons d'imaginer à quoi ressemblerait le fonctionnement d'un tel système, notre représentation est si limitée qu'il nous est impossible d'en saisir pleinement le fonctionnement. C'est la position que défend, entre autres, William Lycan dans son ouvrage *Consciousness* dans lequel il défend une théorie de l'esprit qu'il nomme le « fonctionnalisme homonculaire ».[88] À propos de l'argument de la nation chinoise ainsi que des intuitions sur lesquelles il est basé, Lycan écrit :

> Suppose that you were a little, tiny person-say, just ten times the size of a smallish molecule. And suppose that you were located somewhere within Ned Block's brain, perhaps standing somewhere in his· left occipital lobe. What would you see? It would seem to you that you were standing in the middle of a vast and largely empty space. Occasionally a molecule (looking something like a cluster of basketballs) would whiz by at a terrific rate; sometimes you would see two or more of these clusters collide and rebound. Now suppose someone were to suggest to you that in fact you were standing inside the body, indeed inside the visual system, of a huge conscious being, whose body consisted just of the aggregate of all those basketball clusters, and that that being at that moment was experiencing a vividly and homogeneously red visual sensation, just in virtue of those otherwise inert basketball clusters' whizzing and bouncing around in the way they are. This would probably seem totally absurd to you, in just the way (I submit) that the example of the population

[86] Ceci étant dit, j'estime que même si le nombre d'homoncules nécessaires était beaucoup moindre, par exemple 1 million, il serait tout de même impossible de saisir autant de paramètres et nos intuitions seraient toutes aussi mises à mal que si nous avions à nous en représenter 100 milliards.

[87] Un homoncule est une version miniature d'un être humain. J'utilise cette expression, car on peut à la fois comprendre l'expérience de pensée de la nation chinoise à l'échelle humaine ou encore à échelle réduite; i.e. qu'au lieu d'imaginer des humains de taille normale qui reproduisent le rôle des neurones, on peut imaginer des humains microscopiques se trouvant dans notre cerveau, et qui pourraient reprendre le rôle des neurones. On parlerait alors d'homoncules. Je trouve que cette façon d'approcher l'argument de Block le rend encore plus fragile, puisque cela, comme je le propose, rend le fait que le système constitué entièrement d'homoncules ne soit pas conscient, encore moins plausible et concevable.

[88] William Lycan, *Consciousness* (Cambridge, Massachusetts : MIT Press, 1987)

> of China seems absurd to Block. And you would be wrong, if Block were standing before a smooth red wall in good light.[89]

Bref, l'idée de Lycan est que notre intuition qu'un tel système ne soit pas conscient d'écoule de notre focalisation erronée sur chacune des parties microscopiques du système, plutôt que sur le système macroscopique dans son ensemble.

C'est pourquoi l'argument de Block semble si convaincant : nous ne sommes pas en mesure de nous représenter la totalité de la nation chinoise qui serait requise pour reproduire le fonctionnement du cerveau humain; le système que nous sommes effectivement en mesure de nous représenter est si simple et limité qu'il nous paraît intuitivement comme n'étant pas conscient. C'est donc dire que nous estimons que la nation chinoise n'est pas consciente, non pas parce qu'elle ne l'est pas, mais bien parce que nous ne sommes pas en mesure de nous la représenter et d'en saisir pleinement le fonctionnement et que si nous étions en mesure de le faire, nous affirmerions plutôt que la nation chinoise est bel et bien consciente. Plusieurs autres objections à l'hypothèse des qualia absents ont été formulées, notamment, par Sydney Shoemaker[90] et Reinaldo Elugardo[91], qui tous deux argumentent qu'un tel scénario est logiquement impossible, puisque cela conduit à un scepticisme insoutenable quant au caractère qualitatif de nos propres états mentaux.[92]

Après tout, la nation chinoise imaginée par Block n'est pas bien différente de notre cerveau. En effet, les neurones de notre cerveau reçoivent des influx chimiques et électriques et en produisent à leur tour, comme le font les chinois dans l'expérience de pensée de la nation chinoise.[93] Si nous prenions qu'une petite partie du cerveau et

---

[89] William Lycan, *Consciousness* (Cambridge, Massachusetts : MIT Press, 1987), 51.
[90] Sydney Shoemaker, « Functionalism and qualia », *Philosophical Studies* 27 (1975): 291-315.
[91] Reinaldo Elugardo, « Functionalism and the Absent Qualia Argument », *Canadian Journal of Philosophy* 13 (2) (1983):161-179.
[92] Ibid., 297.
[93] Il est à noter qu'on pourrait argumenter qu'il y a une différence fondamentale qui expliquerait pourquoi le cerveau est conscient alors que la nation chinoise ne l'est pas nécessairement : le fait que la

que nous en analysions l'activité neuronale, il nous paraîtrait intuitif d'affirmer que ce réseau de neurones ne serait pas conscient, puisqu'il ne s'agit que d'un ensemble de neurones qui reçoivent et produisent des influx nerveux; à première vue, pourquoi est-ce qu'il aurait un effet que cela ferait que d'être ce réseau de neurones? Plus précisément, si nous regardions le fonctionnement d'un réseau de quelques neurones, nous serions tentés de penser que ce système n'est pas conscient.[94] C'est exactement ce qui se produit lorsque nous tentons de nous représenter la nation chinoise; la partie que nous sommes en mesure de nous représenter (qui n'est qu'une infime partie du système total) ne nous semble pas consciente, puis nous extrapolons cette intuition à l'ensemble de la nation chinoise. Il nous est impossible de saisir le fonctionnement de tous les neurones qui agissent dans tout notre cerveau, mais nous ne dirions pas pour autant que le fonctionnement du cerveau ne résulte pas en une conscience. C'est la même chose qui se produit dans le cas de la nation chinoise : ce n'est pas parce que nous ne sommes pas capables de pleinement prendre la mesure du fonctionnement de la nation chinoise et qu'elle nous paraît *intuitivement* comme n'étant pas consciente, que cela implique que la nation chinoise n'est pas consciente pour autant. Ainsi, je considère que nous devons accepter, puisque la nation chinoise possède la même organisation fonctionnelle qu'un organisme consciente, qu'elle doit également être consciente.

2.4 Qualia inversés

Une autre objection peut être formulée contre le principe d'invariance organisationnelle : l'objection des qualia inversés. Cet argument consiste à affirmer qu'il serait possible qu'une entité soit dans un état qui satisferait l'organisation

nation chinoise soit constituée d'homoncules qui sont eux-mêmes conscients (ce qui ne semble pas être le cas lorsqu'on examine les constituants du cerveau) pourrait nous laisser croire que la nation chinoise ne pourrait alors ne pas être consciente. En d'autres mots, on pourrait affirmer que l'isomorphisme fonctionnel détermine l'expérience consciente, à condition que les constituants de ce système ne soient pas eux-mêmes conscients.

[94] Voir l'argument du moulin (*windmill*) de Gottfried Wilhelm Leibniz : Stewart Duncan, « Leibniz's Mill Arguments Against Materialism », *Philosophical Quarterly* 62 (247) (2012): 250-72.

fonctionnelle de notre expérience du rouge (i.e. qui aurait la même organisation fonctionnelle que nous lorsque nous faisons l'expérience du rouge), mais qui ferait plutôt l'expérience du vert. Cet argument peut être étendu à toutes les couleurs du spectre; ainsi, il s'agirait alors d'un spectre inversé.[9596] L'argument des qualia inversés est donc qu'il serait possible que deux entités possèdent la même organisation fonctionnelle à un certain moment, mais que leurs expériences subjectives respectives soient différentes; si le scénario des qualia inversés s'avère possible, cela voudrait donc dire que l'organisation fonctionnelle ne détermine pas entièrement l'expérience phénoménale.

De manière analogue au problème des autres esprits, détaillé dans le chapitre 1, il nous est impossible de savoir si une entité autre que nous fait l'expérience du rouge de la même façon que nous, i.e. qu'il est concevable que lorsque cette autre entité fait référence au « rouge », qu'il s'agisse de la couleur que nous associons au « vert ». Ainsi, selon l'objection des qualia inversés, une machine pourrait avoir la même organisation fonctionnelle qu'une entité consciente, mais que son expérience phénoménale ne soit pas la même que celle de l'entité consciente.[97] Cet argument des

[95] Cet argument a notamment été abordé par John Locke dans *Essay Concerning Human Understanding*.
[96] Le scénario où toutes les couleurs sont inversées n'est toutefois pas nécessaire pour contrer le fonctionnalisme. En effet, cet argument contre le fonctionnalisme fonctionnerait même si nous pouvions trouver ne serait-ce qu'un seul cas où deux couleurs seraient inversées.
[97] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 99-101.

qualia inversés a également été présenté et approfondi par Michael Tye[98], Clyde Hardin[99], Martine Nida-Rümelin[100] et plusieurs autres[101][102][103][104].

2.4.1 Réponse à l'hypothèse des qualia inversés

Premièrement, ce que je répondrais à cette objection, est que, dans le scénario des qualia inversés, bien que les qualia dont l'entité fait l'expérience ne soient pas, à proprement parlé, « les bons »[105], cette entité ferait tout de même l'expérience de qualia, i.e. qu'il y aurait un effet que cela fait que d'être cette entité lorsqu'elle fait l'expérience de la couleur rouge, malgré le fait que cet effet est celui que nous associons à l'expérience de la couleur verte. Ainsi, nous pourrions tout de même affirmer que cette entité est consciente puisque qu'elle possède une expérience phénoménale qui lui est propre, malgré le fait qu'elle ne soit pas tout à fait identique à la nôtre. Ainsi, une machine qui inverserait l'expérience du rouge et du vert, serait tout de même consciente, et nous pourrions tout de même conclure que la conscience artificielle est possible, malgré le fait que les qualia soient inversés.

Deuxièmement, bien que la possibilité des qualia inversés soit préoccupante, je considère tout de même qu'un tel scénario est improbable.[106] En effet, j'estime que

[98] Michael Tye, « Qualia, Content, and the Inverted Spectrum », *Noûs* 28 (2) (1994): 159-83.
[99] Clyde Hardin, « Qualia and materialism : Closing the explanatory gap », *Philosophy and Phenomenological Research* 48 (2) (1987) : 281-298.
[100] Martine Nida-Rümelin, « Pseudonormal Vision : An Actual Case of Qualia Inversion? », *Philosophical Studies* 82 : 145-157.
[101] Terence Horgan, « Functionalism, Qualia, and the Inverted Spectrum », *Philosophy and Phenomenological Research* 44 (1984) : 453-470.
[102] Roberto Casati, « What is Wrong in Inverting Spectra ? », *Theoria* 10 (1990) : 183-186.
[103] Uwe Meyer, « Do Pseudonormal Persons Have Inverted Qualia? Scientific Hypotheses and Philosophical Interpretations », *Facta Philosophica*, 2 (2000): 309-325.
[104] Ned Block, « Sexism, ageism, racism, and the nature of consciousness », *Philosophical Topics* 26 (1-2) (1999) : 39-70.
[105] Lorsque je parle des « bons qualia », je veux dire, par exemple, que si un système voit une pomme rouge, ce système devrait faire l'expérience de la rougeur de la pomme, et non pas de la couleur verte de celle-ci, puisqu'elle rouge, et non verte. Dans le cas des expériences visuelles, il faut que les expériences conscientes du système soient, autant que possibles, liées à la réalité observée.
[106] Paul Churchland et Patricia Churchland, « Functionalism, qualia and intentionality », *Philosophical Topics* 12 (1) (1981) : 121-145.

deux systèmes ayant exactement la même organisation fonctionnelle auront exactement la même expérience subjective. En d'autres mots, je suis d'avis qu'il est hautement improbable qu'un système fasse l'expérience du rouge alors que l'autre fait l'expérience du vert, si ces deux systèmes sont des isomorphes fonctionnels. Pour appuyer cette idée, je me baserai sur une autre expérience de pensée que Chalmers présente dans son article, soit celle des *qualia dansants*.[107]

Imaginons BILL, un isomorphe fonctionnel de DAVID. Bien qu'ils partagent la même organisation fonctionnelle, BILL fait l'expérience du vert, alors que DAVID fait l'expérience du rouge. Ces deux systèmes diffèrent en un seul point : dans une certaine région du cerveau, au lieu d'avoir des neurones comme DAVID, il y a des puces de silicone dans le cerveau de BILL.[108] En d'autres mots, une partie du cerveau de BILL a été remplacée par des puces de silicones, qui jouent le même rôle causal que les neurones de DAVID. L'étape cruciale de cette expérience de pensée est de prendre un circuit de puces de silicone identique à celui présent dans le cerveau de BILL et de l'installer sur le cerveau de DAVID comme un circuit parallèle. Ce circuit est alors évidemment un isomorphe fonctionnel du circuit de neurones déjà présent dans le cerveau de DAVID. À ce moment, nous installons un interrupteur qui nous permet de basculer directement entre le circuit de neurones et le circuit de puces de silicone. Lorsque nous activons l'interrupteur, le circuit de neurones devient alors inutile et le circuit de silicone prend la relève; les processus qui étaient effectués par le circuit de neurones sont alors effectués par le circuit de silicone.[109]

Avant d'activer l'interrupteur, DAVID fait l'expérience du rouge. Après avoir activer l'interrupteur, DAVID fait l'expérience du vert, puisque son système est alors

---

[107] David Chalmers, « *Absent qualia, fading qualia, dancing qualia* », dans *Conscious Experience,* ed. Thomas Metzinger (Ferdinand Schoningh, 1995), 309-328.
[108] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 266.
[109] Ibid., 267-268.

fonctionnellement identique à celui de BILL. Ce qui se produit en activant l'interrupteur est que l'expérience phénoménale de DAVID change « devant ses yeux ». En un instant, son expérience passe du rouge au vert. On peut donc s'imaginer activer l'interrupteur continuellement, de sorte que l'expérience de DAVID « clignotent » entre le rouge et le vert. C'est ce que Chalmers appelle des « qualia dansants ».[110]

Ce qui est particulièrement étrange avec le scénario des qualia dansants est que l'expérience de DAVID change constamment entre le rouge et le vert, mais il ne s'en rend pas compte et ne décèle rien de spécial! En effet, puisque l'organisation fonctionnelle de DAVID ne change pas après que nous ayons activé l'interrupteur, l'organisation cognitive de DAVID est précisément exactement la même que si nous ne l'avions pas activé. DAVID ne pourrait donc pas s'exclamer : « Quelque chose d'étrange est en train de se produire! ». Il ne verrait tout simplement pas la différence.[111] Si le scénario des qualia dansants était possible, le pomme qui se trouve sur notre bureau pourrait passer du rouge au vert, sans que nous nous en rendions compte![112][113]

[110] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 268.

[111] Encore une fois, cet argument semble, quelque peu, prêcher aux convertis. En effet, si nous sommes déjà convaincus que l'organisation fonctionnelle d'un système est ce qui détermine entièrement son expérience consciente, alors nous n'avons pas vraiment à être convaincus qu'il n'est pas plausible que nos qualia changent comme c'est le cas dans le scénario des qualia dansants. Inversement, si nous ne sommes pas initialement convaincus, cet argument nous paraîtra comme étant un vœu pieux, soit celui que les qualia dansants ne soient pas possibles. En d'autres mots, l'argument de Chalmers n'est pas vraiment susceptible de nous convaincre que les qualia dansants sont improbables, sauf si nous sommes déjà enclins à adopter un tel point de vue.

[112] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 268-269.

[113] Il est à noter que David Chalmers admet que les qualia dansants puissent être métaphysiquement possibles, et qu'il n'essaie pas d'établir que l'organisation fonctionnelle est ce qui expliquement entièrement la conscience.

La conclusion de Chalmers est qu'il n'est pas plausible que nos expériences changent si drastiquement, alors que notre organisation fonctionnelle est maintenue. Ainsi, il estime que deux systèmes ayant la même organisation fonctionnelle (ici, DAVID et BILL) devraient avoir les mêmes expériences phénoménales. Puisque, selon le principe d'invariance organisationnelle, deux systèmes ayant la même organisation fonctionnelle auront nécessairement les mêmes expériences phénoménales, il en va de soi que, si un ordinateur — une machine ou tout autre système artificiel — avait exactement la même organisation fonctionnelle qu'une entité consciente, il devrait nécessairement avoir les mêmes expériences phénoménales. Si les expériences phénoménales sont identiques, l'effet que cela fait que d'être ce système artificiel sera le même que d'être l'entité consciente.[114] C'est pourquoi j'estime que la conscience artificielle est possible.[115][116] Ceci étant dit, comme je l'ai mentionné plus tôt, j'estime que nous pouvons établir que les robots seront conscients s'ils sont des isomorphes fonctionnels d'un système conscient même si nous ne pouvons pas établir s'ils voient du rouge plutôt que du vert. Ainsi, mon argument selon lequel les robots pourraient être conscients tiendrait la route même si *le scénario des qualia dansants* s'avérait être possible.

---

[114] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 270-271.

[115] Maintenant que nous avons conclu que la conscience artificielle est possible en principe, il serait alors intéressant d'explorer la possibilité de la conscience artificielle en pratique. Plus précisément, nous avons conclu que s'il était possible de reproduire le rôle causal de chacun neurone du cerveau, alors nous nous retrouverions avec un isomorphe fonctionnel de celui-ci et ce nouveau système serait alors tout aussi conscient qu'un cerveau humain normal; mais est-il vraiment possible de reproduire le rôle causal d'un neurone? Des chercheurs ont démontré qu'il est possible d'approximer les caractéristiques intrants/extrants d'un neurone cortical à l'aide d'un réseau de neurones convolutif (*CNN*) de 5 à 8 couches. En d'autres mots, un réseau profond de neurones est en mesure de reproduire artificiellement le comportement d'un neurone. Évidemment, un réseau de neurones de 5 à 8 couches est énorme; surtout c'est ce qui est nécessaire pour reproduire un seul neurone. Il serait donc intéressant de voir s'il est possible de combiner ces réseaux de neurones artificiels dans le but de reproduire l'organisation fonctionnelle complète du cerveau.

[116] David Beniaguev, Idan Segev, Michael London, « Single cortical neurons as deep artificial neural networks », *Neuron* 109 (17) (2021), 2727-2739.

CHAPITRE III : L'IMPOSSIBILITÉ DE LA CONSCIENCE ARTIFICIELLE

Dans ce chapitre, je me pencherai sur les objections formulées contre la possibilité de la conscience artificielle, i.e. des objections qui avancent qu'il est impossible, autant en théorie qu'en pratique, qu'une machine ait une conscience phénoménale. Je proposerai une réponse à chacune de ces objections, dans le but de réitérer et de maintenir ma thèse, selon laquelle la conscience artificielle est possible.

3.1 La chambre chinoise de John Searle

Dans un article paru en 1980 intitulé *Minds, Brains and Programs*[117], John Searle argumente que le fait d'implémenter un programme n'est pas suffisant pour produire un esprit, une conscience, i.e. que la conscience n'est pas *computable*. Pour ce faire, il détaille une expérience de pensée qui est maintenant bien connue : la chambre chinoise.

Cette expérience de pensée va comme suit : imaginons une pièce dans laquelle se trouve une personne qui ne comprend pas un seul mot du chinois. Dans la pièce se trouve également un livre d'instructions qui lui permet de manipuler les symboles chinois (par exemple : si vous recevez *x*, répondez avec *y*). Ce livre ne lui permet toutefois pas de comprendre ce que ces symboles signifient; il ne fait que lui indiquer comment manipuler ces symboles. Des personnes se trouvant à l'extérieur de la pièce lui envoient sur des bouts de papier des questions écrites avec des symboles chinois, auxquelles il doit répondre. Il repère donc les symboles qu'il a reçus dans le livre d'instructions, écrit sur un autre bout de papier les symboles que le livre d'instructions lui indiquent et l'envoie aux personnes qui se trouvent à l'extérieur de la pièce. Cette procédure lui permet de répondre correctement aux questions en chinois, sans toutefois

[117] John Searle, « Minds, Brains and Programs », *Behavioral and Brain Sciences* 3 (3) (1980): 417-457.

comprendre cette langue. À ce moment-ci, les personnes qui se trouvent à l'extérieur de la chambre chinoise sont ainsi convaincues que la personne comprend le chinois. Or, ce dernier ne fait que manipuler des symboles sans toutefois avoir aucune compréhension des questions qu'il reçoit, ni de ce qu'il répond en retour.[118]

Searle utilise cette expérience de pensée pour illustrer le fonctionnement d'un ordinateur, mais surtout dans le but de démontrer que, bien que la personne à l'intérieur de la chambre chinoise semble comprendre le chinois, car elle répond exactement comme le ferait quelqu'un qui comprend véritablement le chinois, il n'en est rien : la personne ne comprend pas ce que signifie les symboles qu'elle reçoit, ni les symboles qu'elle renvoie. L'argument de Searle est que le fait d'exécuter un programme n'est pas suffisant pour « comprendre » le chinois. Avec cette expérience de pensée, Searle veut démontrer l'absurdité d'attribuer quelconque forme d'intentionnalité ou de conscience à un système qui ne ferait que manipuler des symboles. En effet, on considère que la personne dans la chambre chinoise exécute un programme algorithmique, précisément comme le font les ordinateurs; ainsi, à l'instar de la personne dans la chambre chinoise, nous ne pourrions pas dire qu'un ordinateur, ne faisant qu'exécuter un programme, « comprend » réellement ce qu'il fait, ce que les symboles veulent dire et ce que le résultat de la manipulation de ces symboles signifie.

De manière plus formelle, l'argument de Searle prend donc cette forme :

1. Si l'IA forte est vraie, alors il y a un programme pour le chinois tel que, si n'importe quel système donné exécute ce programme, alors ce programme comprendra le chinois.
2. Je pourrais exécuter un programme pour le chinois sans toutefois comprendre le chinois.
3. Donc, l'IA forte doit être fausse

[118] John Searle, « Minds, Brains and Programs », *Behavioral and Brain Sciences* 3 (3) (1980): 417.

Cette expérience de pensée s'attaque spécifiquement à l'intentionnalité, mais, comme l'avance Chalmers, le problème de la conscience est à la base de l'argument formulé par Searle.[119] Si l'argument de la chambre chinoise de Searle s'avère être probant, cela voudra dire par le fait même que le système est dépourvu de certains états conscients, plus précisément de l'expérience consciente de comprendre le chinois. Ainsi, si nous appliquons l'expérience de pensée de la chambre chinoise à l'expérience consciente du rouge, nous en conclurions que la personne qui ne fait qu'exécuter un programme algorithmique ne fait pas réellement l'expérience de la couleur rouge. De manière analogue, Searle en conclurait donc que n'importe quel programme prétendant pouvoir produire une expérience de rouge est voué à l'échec; en d'autres mots, l'implémentation d'un programme n'est pas suffisante pour produire une conscience.[120]

3.1.1 Réponse à l'argument de la chambre chinoise

Plusieurs objections ont été formulées contre l'expérience de pensée de la chambre chinoise ainsi que sur l'argument plus général de Searle contre le computationnalisme. Searle lui-même en a proposées quelques-unes dans son article original.[121] Je trouve que les objections les plus convaincantes sont celles qui, d'une façon ou d'une autre, avancent que, bien que la personne se trouvant dans la chambre chinoise ne comprenne pas le chinois, le système en entier peut le comprendre et peut donc être conscient.[122] Un des problèmes avec l'expérience de pensée de la chambre chinoise est qu'on examine et analyse qu'une seule des parties (la personne qui manipule les symboles) pour déterminer la conscience de tout le système. Cette objection est connue sous le nom de « la réponse des systèmes » (*systems reply*). La

[119] David Chalmers, *The Conscious Mind: In Search of Fundamental theory* (New York: Oxford University Press, 1996), 322-327.
[120] John Searle, « Minds, Brains and Programs », *Behavioral and Brain Sciences* 3 (3) (1980): 419.
[121] Ibid., 417-457.
[122] Lorsqu'on fait référence au système « en entier », on fait référence au système constitué de la personne, de la chambre chinoise du livre d'instructions, etc.

réponse des systèmes a été défendue et mise de l'avant par plusieurs penseurs au fil des années.[123][124][125][126]

Pour nous aider à comprendre ce problème, nous pouvons retourner à l'expérience de pensée de la nation chinoise proposée par Block : bien que les parties qui forment le système (soit chacun.e des chinois.es) ne soient pas conscientes (du moins, pas conscientes de la même façon que la nation chinoise pourrait l'être), cela n'empêche pas que le système en entier puisse l'être. Dans le même ordre d'idées, si on observe et examine qu'une partie du cerveau humain (ou même un seul neurone), il nous semble raisonnable de penser que cette partie ne soit consciente et qu'il n'y ait pas d'effet que cela fait que d'être cette partie du cerveau. Pourtant, lorsque toutes ces parties (et tous les neurones) sont combinées de manière adéquate, le système qui en résulte est tout de même conscient.

La réponse des systèmes s'apparente à la réponse que j'avais proposée à l'expérience de pensée de la nation chinoise dans le chapitre 2. En effet, il me semble tout à fait sensé de penser qu'un neurone, ou un petit ensemble de neurones, pris individuellement, ne soit pas conscient. De manière analogue, il est plausible de penser qu'une partie du système qu'est la chambre chinoise ne le soit pas non plus, soit la personne qui se trouve à l'intérieur de la chambre.[127] Ceci étant dit, le fait qu'une partie d'un système ne soit pas conscient ne nous permet pas d'en déduire que le système

[123] Jack Copeland, « The Chinese Room from a Logical Point of View », dans *View Into the Chinese Room : New Essays on Searle and Artificial* Intelligence, ed. John M. Preston et John Mark Bishop (Londres : Oxford University Press, 2003), 109-122.
[124] Douglas Hofstadter, « Reflections on Searle », dans *The Mind's I*, ed. Douglas Hofstadter Daniel Dennett (New York: Basic Books, 1981), 373-382.
[125] John Haugeland, « Syntax, Semantics, Physics », dans *View Into the Chinese Room : New Essays on Searle and Artificial* Intelligence, ed. John M. Preston et John Mark Bishop (Londres : Oxford University Press, 2003), 379-392.
[126] Georges Rey, « What's Really Going on in Searle's 'Chinese Room' », *Philosophical Studies* 50 (1986): 169-85.
[127] Dans ce cas-ci, plutôt que de dire que la personne n'est pas consciente, nous pourrions dire que la personne ne comprend pas le chinois.

dans son ensemble ne l'est pas. Pour revenir au système qu'est le cerveau : malgré le fait que les neurones pris individuellement ne soient visiblement pas conscients, cela ne signifie pas pour autant que le système, le cerveau, n'est pas conscient. Bien au contraire, une conscience semble émerger du cerveau malgré le fait que ses parties, prises individuellement, ne nous paraissent pas comme étant conscientes. En fait, la personne dans la chambre joue sensiblement le même rôle causal qu'un neurone ou un système de neurone.

Une autre réponse à l'argument de Searle que je trouve convaincante est celle de l'esprit virtuel (*virtual mind reply*). Les défenseurs de cette réponse avancent que ce qui importe dans le cas de la chambre chinoise n'est pas si l'agent qui se trouve dans la chambre comprenne ou non le chinois, mais bien s'il y a une sorte de compréhension qui se produit.[128] En effet, selon eux, il serait possible que le fait de réaliser un programme algorithmique (comme le fait la personne à l'intérieur de la chambre) produise un agent qui, lui, comprend le chinois. Cet agent serait non seulement distinct de la personne qui se trouve dans la chambre (trivialement, puisque cette dernière ne comprend pas le chinois), mais aussi distinct du système en entier. C'est ici que la réponse de l'esprit virtuel se distingue de la réponse des systèmes. La réponse de l'esprit virtuel a été défendue par plusieurs.[129][130][131]

C'est cette réponse que propose David Cole dans son article « Artificial Intelligence and Personal Identity »[132], dans lequel il argumente que les souvenirs et la personnalité de la personne qui comprendra le chinois seront distincts de ceux de la

[128] David Cole, « Artificial Intelligence and Personal Identity », *Synthese* 88 (1991): 401.
[129] Aaron Sloman et Monica Croucher, « How to turn an information processor into an understanding », *Brain and Behavioral Sciences* 3 (1980): 447-8.
[130] Ned Block, « Searle's Arguments Against Cognitive Science », dans *View Into the Chinese Room : New Essays on Searle and Artificial* Intelligence, ed. John M. Preston et John Mark Bishop (Londres : Oxford University Press, 2003), 70-79.
[131] Patrick Hayes, Stevan Harnad, Donald Perlis et Ned Block, « Virtual Symposium on Virtual Mind », *Minds and Machines*, 2(3) (1992): 217-238.
[132] David Cole, « Artificial Intelligence and Personal Identity », *Synthese* 88 (1991): 399-417

personne qui se trouve dans la chambre chinoise.[133] Selon lui, il faudrait donc en conclure que la personne qui se trouve dans la chambre n'est pas la même personne que celle qui comprend le chinois. :

> There may well be a mind realized by Searle's activity[134], a virtual person. But the same mind could have been realized by the activity of someone other than Searle - Searle could even resign his job in the Room and be replaced by another - while the Chinese conversation continues. This is additional evidence that the Chinese understanding person is not Searle. Searle is not essential to the existence of the Chinese understanding person.[135]

Bref, contrairement à Searle qui avance qu'il n'y a pas de compréhension du chinois parce que la personne dans la chambre ne comprend pas le chinois, Cole argumente que, le fait que cette personne ne comprenne pas véritablement le chinois ne pose pas réellement problème puisque ce qui *fait la compréhension* (*does the understanding*) du chinois est un agent qui est bien distinct de la personne dans la chambre et même du système qu'est la chambre.

3.2 L'argument de l'intuition de Hubert Dreyfus

En 1972, le philosophe américain Hubert Dreyfus publie un livre intitulé « *What Computers Can't Do* »[136], dans lequel il présente, comme le titre de son livre l'indique, son évaluation pessimiste quant au développement de l'intelligence artificielle ainsi que les limites potentielles de celle-ci. Ces idées ont notamment été développées et retravaillées dans un livre paru vingt ans plus tard, « *What Computers Still Can't Do* ».[137] Une des idées principales présentées par Dreyfus est que l'intelligence et l'expertise humaine résultent notamment de processus inconscients qui ne peuvent pas être capturés par un ordinateur, i.e. par un système de règles formelles.

---

[133] David Cole, « Artificial Intelligence and Personal Identity », *Synthese* 88 (1991): 406.
[134] Cole fait référence à Searle comme étant la personne se trouvant dans la chambre chinoise.
[135] Ibid., 406.
[136] Hubert Dreyfus, *What Computers Can't Do* (New York: MIT Press, 1972).
[137] Hubert Dreyfus, *What Computers Still Can't Do* (New York: MIT Press, 1992).

Cette idée que l'intuition humaine ne puisse pas être reproduite par un ordinateur est au cœur de « Mind over Machine »[138] dans lequel il argumente que l'esprit humain ne pourra jamais reproduit par un programme algorithmique.

Dreyfus fait la différence entre « savoir-que » (*knowing-that*) et « savoir-faire» (*knowing-how*).[139] Le « savoir-que » correspond à notre habilité à consciemment résoudre des problèmes « étape par étape ». Nous faisons appel à notre « savoir-que » lorsque nous faisons face à des problèmes qui requiert que nous nous arrêtions, prenions un pas de recul et que nous évaluions les possibilités et idées une après l'autre. Cette manipulation consciente de « symboles » est analogue au fonctionnement d'un ordinateur, qui manipule des symboles à l'aide de règles d'inférence. C'est d'ailleurs pour cela que Dreyfus avance que le « savoir-que » peut être capturé et reproduit par un programme algorithmique.

De l'autre côté, le « savoir-faire » correspond à la façon dont nous traitons les choses *dans la vie de tous les jours*. En effet, il s'agit de notre capacité à agir en n'utilisant pas de manipulation de symboles comme c'est le cas pour le « savoir-que ». Par exemple, lorsque nous reconnaissons un visage familier, nous n'avons pas à faire faire appel à un raisonnement basé sur des règles d'inférence pour reconnaître la personne devant nous. Il ne s'agit pas d'un calcul que nous devons faire; nous reconnaissons la personne devant nous, naturellement. En fait, nous n'avons même pas à « réfléchir »; nous sautons directement à la conclusion, sans trop nous poser de questions. C'est ce que Dreyfus appelle « l'essence de l'expertise humaine » : certains processus humains sont basés sur une sorte d'intuition plutôt que sur une forme de raisonnement. Cette expertise humaine est basée dans toutes nos intuitions et attitudes

[138] Hubert Dreyfus et Stuart Dreyfus, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* (New York: The Free Press, 1986).
[139] Cette distinction que fait Dreyfus entre « savoir-que » et « savoir-faire» prend racine dans la distinction que soulève Heidegger entre « present-at-hand » et « ready-to-hand ».

inconscientes. C'est ce que Dreyfus appelle le « contexte ».[140] Ce dernier n'est pas stocké dans nos cerveaux de manière symbolique, mais plutôt de manière intuitive.

Puisque ce « savoir-faire » ne correspond pas à quelconque manipulation consciente de symboles selon des règles d'inférence, Dreyfus argumente qu'il ne peut pas être capturé ou reproduit par un ordinateur, notamment en raison du fait qu'un programme algorithmique n'est qu'un système de manipulation de symboles. Ainsi, il serait possible qu'un programme algorithmique reproduise le « savoir-que » de l'humain, mais serait incapable d'en reproduire le « savoir-faire ». C'est ainsi que Dreyfus s'oppose à la possibilité de la théorie computationnelle de l'esprit; tous les processus inconscients qui jouent un rôle important dans le fonctionnement du cerveau et du comportement humain ne peuvent être reproduits par quelconque programme algorithmique. Dreyfus avance notamment que les ordinateurs, qui n'ont pas de corps, pas d'enfance, ni de pratiques culturelles ne peuvent pas acquérir d'intelligence. Une des raisons qu'avancent Dreyfus pour soutenir cette position est que le raisonnement humain est en partie implicite/tacite, ce qui fait en sorte qu'il ne puisse pas être articulé et incorporé dans un programme algorithmique.[141]

3.2.1 Réponse à l'argument de l'intuition

Malgré la récente domination des programmes d'intelligence artificielle contre les humains aux échecs[142], certains étaient plutôt pessimistes quant aux développements et aux visées de l'IA.[143] En effet, plusieurs pensaient que l'IA n'allait jamais surpasser l'humain au jeu de Go[144], puisque, selon certaines personnes, ce jeu

[140] Ce concept fait notamment référence au concept de *Dasein* de Heidegger.
[141] Ragnar Fjelland, « Why general artificial intelligence will not be realized », *Humanit Soc Sci Commun* 7 (10) (2020) : 1-9.
[142] Karsten Müller et Jonathan Schaeffer, *Man vs Machine : Challenging Human Supremacy at Chess* (Connecticut : Russell Enterprises Inc., 2018).
[143] Jerry Kaplan, *Artificial Intelligence: What Everyone Needs to Know* (Oxford University Press, 2016).
[144] Matthew Hutson, « This computer program can beat humans at Go with no human instruction », *Science*, publié le 18 octobre 2017, https://www.science.org/content/article/computer-program-can-beat-humans-go-no-human-instruction.

requiert une certaine forme d'intuition humaine.[145] En effet, certains joueurs avancent que pour être bon au jeu de Go, il faut un mélange d'intuition et de raisonnement.[146] Certains champions vont même jusqu'à dire que, dans certains cas, ils jouent un certain coup, mais ne peuvent pas expliquer pourquoi ils ont joué ce coup et pas un autre; en d'autres mots, c'est leur intuition qui leur a indiqué le coup à jouer.[147]

Il est facile de faire un lien entre ces capacités intuitives des joueurs de Go et le concept de « savoir-faire » décrit par Dreyfus : les joueurs ne jouent pas nécessairement en fonction d'un raisonnement conscient, mais plutôt parce qu'il s'agit de « la chose à faire »; une sorte de conclusion intuitive et apparemment évidente. Le fait que le jeu de Go requière une sorte d'intuition inexplicable est ce qui poussait certaines personnes à affirmer qu'aucun ordinateur ne pourrait battre un champion de Go, puisque l'intuition humaine ne serait pas *computable*, comme le pense Dreyfus.

Or, en 2016 le programme AlphaGo, développé par Google DeepMind a réalisé l'impossible et a battu Lee Sedol, qui était alors le meilleur joueur de Go du monde et qui avait prédit qu'il dominerait AlphaGo.[148] Ce résultat a notamment été comparé à la défaite de Garry Kasparov face à Deep Blue en 1996. AlphaGo avait battu Sedol 4 parties contre 1 et a depuis été grandement amélioré; AlphaGo Zero (le descendant de AlphaGo) a battu le programme qui avait battu Sedol, 100 parties contre 0.[149] Ces résultats écrasants sont sans équivoque : l'ordinateur a surpassé l'humain au jeu de Go, ce que plusieurs pensaient être impossible.

---

[145] *AlphaGo*, réalisé par Greg Kohs (2017; É-U : Moxie Pictures), Netflix.
[146] George Johnson, « To Test a Powerful Computer, Play an Ancient Game », *The New York Times*, publié le 29 juillet 1997, https://www.nytimes.com/1997/07/29/science/to-test-a-powerful-computer-play-an-ancient-game.html.
[147] *AlphaGo*, réalisé par Greg Kohs (2017; É-U : Moxie Pictures), Netflix.
[148] Matt MacFarland, « Google AI defeats world's Go champion in gripping 'man vs machine' film », *CNN Business*, publié le 29 septembre 2017, https://money.cnn.com/2017/09/29/technology/future/alphago-movie/index.html.
[149] Demis Hassabis et David Siver, « AlphaGo Zero : Learning from Scratch », *DeepMind*, publié le 18 octobre 2017, https://www.deepmind.com/blog/alphago-zero-starting-from-scratch.

Les conclusions que nous pouvons tirer des avancées de AlphaGo sont soit que 1) AlphaGo est en mesure de capturer et de reproduire l'intuition humaine que requiert le jeu de Go ou encore que 2) le jeu de Go ne requière pas d'intuition humaine, et donc que, bien qu'AlphaGo soit supérieur à l'humain au jeu de Go, cela ne signifierait pas pour autant qu'il est en mesure de capturer et de reproduire l'intuition humaine. Je considère que la première option est beaucoup plus probable et plausible. En effet, j'estime que l'argument de Dreyfus ne pose pas de réellement de problème pour la possibilité de la conscience artificielle, puisque, comme cela semble être le cas pour AlphaGo, un programme algorithmique, bien qu'il ne s'agisse que d'un système de manipulation de symboles, peut reproduire le « savoir-faire » de l'humain. De plus, de nombreux algorithmiques ont la capacité de reconnaître des visages et des expressions faciales, qui est une capacité que Dreyfus avait associée au « savoir-faire », et qui ne serait pas *reproduisible* par un programme algorithmique.[150][151]

3.3 L'argument de la simulation de Christof Koch

Dans « The Feeling of Life Itself »[152], Christof Koch se montre critique de la possibilité de télécharger un esprit (*mind uploading*) sur un ordinateur: « Two systems can be functionally equivalent, they can compute the same input-output function, but they don't share the same intrinsic cause-effect form ».[153] Pour appuyer cette idée, il illustre un ordinateur qui n'existe pas intrinsèquement, alors que le circuit qui est

[150] Paramjit Kaur, Kewal Krishan, Suresh K. Sharma et Tanuj Kanchan, « Facial-recognition algorithms : A literature review », *Medicine, Science and the Law* 60 (2) (2020) : 131-139.
[151] Cette argument de Dreyfus est lié à celui de chambre chinoise de Searle parce qu'on pourrait se demander si l'algorithmique reconnaît « véritablement » les visages ou s'il ne fait que manipuler des symboles *inconsciemment* sans toutefois reconnaître les visages de manière intuitive et immédiate (comme cela pourrait sembler être le cas pour les humains), de la même façon qu'on se demandait si la personne placée dans la chambre chinoise comprenait « véritablement » le chinois ou si elle ne faisait que manipuler des symboles de manière à nous laisser croire qu'elle comprenait le chinois.
[152] Christof Koch, *The Feeling of Life Itself: Why Consciousness is Widespread but Can't Be Computed*, (Massachusetts: MIT Press, 2019).
[153] Ibid., 148.

simulé existe intrinsèquement.[154][155] Ce qui pousse Koch à avancer que bien que les deux systèmes *font* la même chose, seulement un des deux existe pour lui-même (*is for itself*). Bref, c'est ce qui pousse Koch à avancer qu'une simulation peut répliquer complètement n'importe quel circuit, sans toutefois avoir d'expériences conscientes, et ce, peu importe ce que l'ordinateur est programmé à faire.[156]

Selon Koch, la conscience n'est pas un « algorithme astucieux ».[157] La conscience d'un système est basée sur le pouvoir causal que ce système a sur lui-même, et non pas sur une forme de computation. Le pouvoir causal dont parle Koch fait référence à la capacité d'un système à s'influencer lui-même et à influencer les autres. Le problème du projet de la conscience artificielle est, selon Koch, que ce pouvoir causal ne peut être simulé.

Pour illustrer cette idée, Koch donne l'exemple des physiciens qui font des simulations informatiques de trous noirs. Ces trous noirs, avec leur masse énorme, exercent une force gravitationnelle si grande que même la lumière ne peut en échapper. Les physiciens sont en mesure de simuler le comportement des trous noirs, notamment à l'aide des équations de la théorie de la relativité générale d'Einstein. C'est à ce moment que Koch nous demande : si les physiciens sont en mesure de reproduire si fidèlement les trous noirs et leur immense force gravitationnelle, pourquoi ne sont-ils pas aspirés dans leurs ordinateurs par les trous noirs qu'ils simulent? Parce que la gravité n'est pas une computation! Bien que nous soyons en mesure de simuler des trous noirs, cela ne fait pas en sorte que, magiquement, ces simulations acquièrent le

[154] Plus précisément, il donne l'exemple d'un ordinateur qui est fonctionnellement équivalent à un circuit triadique. Selon la théorie de l'information intégrée (*Integrated information theory*), l'ordinateur n'est pas conscient puisque qu'il n'intègre pas d'information, malgré le fait qu'il simule un circuit dont l'information intégrée est non-nulle.
[155] Christof Koch, *The Feeling of Life Itself: Why Consciousness is Widespread but Can't Be Computed*, (Massachusetts: MIT Press, 2019), 148.
[156] Ibid., 148.
[157] Ibid., 148.

pouvoir causal des systèmes qu'elles reproduisent. Bref, selon Koch, une simulation d'un trou noir n'est pas un trou noir, et ce, pour les mêmes raisons qu'un ordinateur simulant un ouragan n'est pas mouillé; le programme peut être fonctionnellement identique à certains aspects de la réalité (comme c'est le cas de la simulation d'un trou noir ou d'une tempête), mais n'a tout de même pas les mêmes pouvoir causaux que les choses réelles qu'elle imite.[158] Selon Koch, ce constat devrait également s'appliquer à la conscience : une simulation informatique de la conscience ne serait qu'une simulation, ce que Koch appelle de la « fausse conscience », puisqu'elle n'aurait pas le pouvoir causal de la « vraie conscience ».[159]

C'est donc dire que, si nous arrivions à imiter un cerveau humain en entier (comme dans l'expérience de pensée décrite dans le chapitre 2), incluant tous les événements synaptiques et neuronaux qui y ont lieu, cette simulation serait tellement fidèle au fonctionnement du cerveau « réel » que le comportement de celle-ci serait *indifférenciable* de celui du cerveau qu'il pourrait même passer le test de Turing de la conscience (voir chapitre 1).

Ceci étant dit, selon Koch, tant et aussi longtemps que l'architecture du programme qui simule le cerveau humain ressemblera à celle d'une machine de von Neumann[160], le programme « ne verra pas », « n'entendra pas »; bref, il « n'aura pas d'expérience »[161], puisque le pouvoir causal, soit l'habileté d'influencer les autres et soi-même, ne peut pas être simulé. Il n'y aurait donc pas d'effet que cela fait que d'être une simulation. Ce programme ne serait qu'un « astucieux algorithme » qui permet de reproduire et d'imiter le fonctionnement et le comportement d'un cerveau humain

[158] Christof Koch, *The Feeling of Life Itself: Why Consciousness is Widespread but Can't Be Computed*, (Massachusetts: MIT Press, 2019), 149.
[159] Ibid., 150.
[160] La machine de Von Neumann est composée de 4 parties distinctes : 1- l'unité arithmétique et logique, 2- l'unité de contrôle, 3- la mémoire et 4- les dispositifs entrée-sortie (*input-output*).
[161] Ibid, 150.

entier, mais cela ne ferait tout de même pas en sorte qu'une conscience phénoménale puisse émerger d'un tel programme. Ceci étant dit, Koch admet tout de même que si un matériel (*hardware*) était spécifiquement basé sur celui du cerveau (*neuromorphic electronic hardware*), alors celui si pourrait amasser suffisamment de pouvoir intrinsèque pour qu'il y ait un effet que cela fait que d'être celui-ci[162], i.e. pour qu'il soit conscient.

3.3.1 Réponse à l'argument de la simulation de Christof Koch

J'estime qu'une des failles de l'argument de Koch est qu'il semble assumer que tout ce qui résulte d'un système algorithmique ne peut être, au mieux, qu'une simulation de ce que le programme tente de reproduire. Il est vrai que la simulation d'un trou noir n'aspirerait pas les scientifiques ou encore que l'ordinateur qui implémente la simulation d'un ouragan ne sera pas mouillé; cependant, il est important de se demander si c'est bel et bien le cas que tout programme computationnel est voué à n'être qu'une simulation.

Revenons à l'exemple d'AlphaGo. Ce programme reproduit le comportement d'un joueur de Go à l'aide d'un programme computationnel, mais nous ne pourrions pas dire qu'il ne s'agit que d'une simulation d'un joueur de Go : AlphaGo joue véritablement à Go. C'est aussi le cas pour les programmes de reconnaissance d'images. Ces programmes ne sont pas que des simulations de personnes qui seraient en mesure de reconnaître des images; ces programmes reconnaissent bel et bien des images.

L'argument de Koch se base sur l'exemple des physiciens qui simulent le comportement d'un trou et qui ne sont évidemment pas aspirés par leur ordinateur. Ce

[162] Christof Koch, *The Feeling of Life Itself: Why Consciousness is Widespread but Can't Be Computed*, (Massachusetts: MIT Press, 2019), 150.

que Koch omet de dire est que ce qui résulte du programme computationnel des physiciens n'est qu'une simulation d'un trou noir, précisément parce que c'est ce que les physiciens cherchaient à faire : ils ne cherchaient pas à créer un trou noir, mais bien à en simuler le comportement. Il n'est donc pas surprenant que le résultat qui en découle soit une simulation d'un trou noir, et non pas un trou noir en soi.[163]

Ceci étant dit, à l'instar de Koch, je demeure sceptique quant à la possibilité de créer un trou noir à partir d'un programme computationnel. Il est donc envisageable que certains phénomènes, comme les trous noirs, ne puissent être, qu'au mieux, simulés. Maintenant, la question que nous devons nous poser est la suivante : à quelle catégorie la conscience phénoménale appartient-elle? À celle des phénomènes qui peuvent être synthétisés à l'aide d'un programme computationnel, comme c'est le cas des habiletés d'AlphaGo, ou encore, à celle des phénomènes qui ne peuvent, au mieux, qu'être simulés? L'idée qui me fait pencher vers la première option est le fait que les phénomènes que sont les trous noirs et les ouragans requièrent une composante physique; par exemple, un ouragan sans particule d'eau sera, au mieux, voué à n'être qu'une simulation d'un ouragan et pas un « vrai » ouragan. Il me semble que ce ne soit pas tout à fait le cas lorsqu'on parle de la conscience. Il me semble que celle-ci ne soit pas, en soi, un phénomène purement physique (comme le sont les ouragans et les trous noirs), mais plutôt un phénomène qui pourrait émerger d'un système computationnel, comme c'est le cas des habiletés du programme AlphaGo : bien qu'il n'ait pas de composante « physique », AlphaGo est tout de même en mesure de jouer au jeu de Go.

Ceci étant dit, il est possible que la conscience nécessite un substrat non-virtuel, et donc, qu'un simple programme computationnel ne serait pas suffisant pour en

[163] Ceci étant dit, nous pourrions aussi mentionner que, malgré le fait que la simulation du trou noir n'ait pas exactement les mêmes pouvoirs causaux que le *vrai* trou noir, cette simulation possède tout de même certains pouvoirs de causalité : les scientifiques peuvent l'analyser, et elle a le pouvoir d'influencer leurs théories et leurs prédictions.

produire une. Cependant, cela ne pose pas un très grand problème quant à la possibilité de la conscience artificielle puisqu'il serait possible d'implanter ce programme computationnel dans un substrat physique; possiblement par l'intermédiaire de puces de silicone, comme je l'ai évoqué dans le chapitre 2. C'est donc dire, que l'argument de Koch pourrait nous faire douter de la capacité à un programme computationnel de produire une conscience, mais cela ne nous empêche toutefois pas de penser qu'un système physique de puces de silicone puisse tout de même être en mesure de faire émerger une conscience.[164][165]

3.4 L'argument de la théorie de l'animal-machine d'Anil Seth

Dans son livre « Being You »[166], Anil Seth propose une théorie appelée « théorie de l'animal-machine » (*Beast Machine Theory*). Selon celle-ci, « la conscience chez les humains et les animaux est apparue au cours de l'évolution et émerge en chacun de nous au cours du développement et opère d'instant en instant de manière intimement liée à notre statut de systèmes vivants ».[167] Bref, cette théorie fonde les expériences du monde et de soi dans une volonté biologique de rester en vie.[168]

Pour illustrer ses doutes quant à la possibilité de la conscience artificielle, Seth nous demande d'imaginer un robot composé d'un cerveau de silicone et d'un corps semblable à celui de l'humain, muni de plusieurs types de senseurs et d'effecteurs :

> The signals flowing through its circuits implement a generative model of its environment, and of its own body. It is constantly using this model to make Bayesian best guesses about the causes of its

[164] Il est intéressant de mentionner qu'il existe déjà du matériel neuromorphique ainsi que des algorithmes spécifiques pour celui-ci, appelés réseaux neuronaux à picots (*Spiking Neural Networks*). Ceux-ci ont la particularité d'imiter fidèlement les réseaux de neurones naturels.
[165] Michael Pfeiffer et Pfeil Thomas, « Deep Learning With Spiking Neurons : Opportunities and Challenges », *Frontiers in Neuroscience* 12 (2018).
[166] Anil Seth, *Being You : A New Science of Consciousness* (New York: Dutton, 2021).
[167] Ibid., 262.
[168] Ibid., 262.

> sensory inputs. The synthetic controlled (and controlling) hallucinations are geared, by design, toward keeping the robot in an optimal functional state — to keep it, by its own lights, 'alive'.[169]

Seth nous demande par la suite si un tel robot (soit un robot que nous pourrions qualifier d'isomorphe fonctionnel d'un être humain normal), serait conscient. En répondant à cette question, Seth se montre prudent, mais tout de même pessimiste quant à une telle possibilité. En effet, selon la théorie de l'animal-machine, la « matérialité de la vie » est un aspect important de toutes les manifestations de conscience. Il continue :

> Self-maintenance for living systems goes all the way down, even down to the level of individual cells. Every cell in your body — in *any* body — is continually regenerating the conditions necessary for its own integrity over time The same cannot be said for any current or near-future computer, and would not be true even for a silicon beast machine of the sort I just described.[170]

3.4.1 Réponse à l'argument de l'animal-machine

Je comprends l'argument de Seth selon lequel c'est la vie et le maintien de celle-ci qui est au cœur de la conscience, plutôt que le traitement d'informations. Cependant, je trouve que Seth n'explique pas clairement pourquoi des programmes computationnels ne seraient pas en mesure de reproduire ces aspects de la vie humaine qui sont caractéristiques du substrat biologique qui composent les humains et les autres animaux. Il me semble plutôt que, dans un système de silicone, qui reproduit non seulement le cerveau, mais également l'ensemble du corps humain, les « cellules » artificielles qui le composent auront tout autant le désir et la volonté de survivre et de se maintenir en vie, comme c'est le cas dans un corps biologique. De plus, on peut aussi se demander si un système composé de cellules artificielles qui reproduiraient parfaitement, au niveau des structures cellulaires (ou même au niveau atomique),

[169] Anil Seth, *Being You : A New Science of Consciousness* (New York: Dutton, 2021), 262.
[170] Ibid., 263.

correspondrait aux critères établis par Seth pour qu'un système puisse être conscient; soit la nécessité que les mécanismes d'*auto-entretien* (*self*-maintenance) soient présents au niveau cellulaire.[171] Je crois que c'est d'ailleurs pour cette raison qu'Anil Seth a admis lui-même que si des organoïdes devenaient suffisamment complexes, leur activité pourrait s'apparenter à celle des humains conscients.[172]

De plus, il semble y avoir un autre problème apparent avec l'argument de Seth. Il caractérise la fonctionnalité de la vie d'une telle manière qu'on pourrait se demander : si un système composé de silicone correspondait à ces critères *fonctionnalistes* établis par Seth, alors pourquoi serait-il nécessaire d'insister sur le fait que les parties qui composent le système correspondent également à ces critères d'*auto-entretien*. Si c'était le cas, on pourrait penser qu'on se retrouverait alors avec un système dont les parties seraient elles-mêmes conscientes; ce qui ne semble vraisemblablement pas être une condition nécessaire pour que le système soit conscient. Nous n'avons qu'à penser à notre propre cerveau pour nous en apercevoir.

3.5 L'argument gödelien de Lucas-Penrose

La prochaine objection que j'analyserai est l'argument gödelien formulé tour à tour par John Randolph Lucas et Roger Penrose. Tous deux avancent que la véracité des théorèmes d'incomplétude de Gödel implique que l'esprit humain ne peut être une machine, i.e. que l'esprit possède des capacités qui ne pourront jamais être reproduites par quelque machine que ce soit. Ainsi, si leur argument s'avère probant, la théorie computationnelle de l'esprit devrait être repensée et le projet de la conscience artificielle, abandonné.[173] Bien que cette objection ne s'attarde pas directement à la

[171] Anil Seth, *Being You : A New Science of Consciousness* (New York: Dutton, 2021), 262.
[172] Benjamin Thompson et Noah Baker, « Lab-grown brains and the debate over consciousness », *Nature*, publié le 28 octobre 2020. https://www.nature.com/articles/d41586-020-03033-6#:~:text=Anil%20Seth%20works%20on%20cognitive,to%20that%20of%20conscious%20humans.
[173] Cette discussion sur l'argument gödelien de Lucas-Penrose contre la théorie computationnelle de l'esprit est en partie tirée/inspirée d'un travail de session que j'ai réalisé dans le cadre du séminaire PHI

possibilité de la conscience artificielle, elle demeure tout de même intéressante puisqu'elle aborde les limites de l'intelligence artificielle quant à la possibilité de synthétiser artificiellement l'esprit humain.[174]

3.5.1 Les théorèmes d'incomplétude de Gödel

Avant de se pencher sur les arguments formulés par Lucas et Penrose, il est important de comprendre de quoi en retourne les théorèmes d'incomplétude de Gödel[175], sur lesquels sont basés leurs arguments. Avec son premier théorème d'incomplétude, Gödel démontre que pour n'importe quel système axiomatique formel cohérent *F*, dans lequel nous pouvons démontrer les bases de l'arithmétique, est nécessairement incomplet, i.e. qu'il sera toujours possible de construire un énoncé *G*, qui est vrai, mais qui ne peut pas être démontré ni réfuté dans ce système *F*. Plus spécifiquement, dans un tel système, il est toujours possible de produire un énoncé *G* tel que *G* : L'énoncé *G* n'est pas démontrable dans *F*. La particularité de cet énoncé, que l'on appelle « énoncé gödelien », est qu'il fait référence à lui-même, i.e. que cet énoncé *G* dit que l'énoncé *G* n'est pas démontrable dans *F*. La possibilité d'un tel énoncé est problématique puisque si l'énoncé est bel et bien démontrable dans *F* (et que *F* est un système cohérent), i.e. qu'en combinant les axiomes de *F* ainsi que ses règles d'inférence, il existe une preuve pour démontrer *G* dans le système, alors *G* serait vrai et donc, *G* ne serait donc pas démontrable dans *F* (puisque c'est ce que *G* représente); contradiction! Inversement, s'il n'existe pas de preuve dans *F* pour démontrer *G*, alors, trivialement, *G* n'est pas démontrable dans *F*. C'est donc dire que si le système *F* est cohérent, alors *G* n'est pas démontrable dans *F*; le système est donc incomplet; puisqu'il existe au moins un énoncé (dans ce cas-ci, *G*) qui est vrai, mais

6340 : Logique et philosophie contemporaine à l'hiver 2021 (Prof. Jean-Pierre Marquis) qui portait sur les théorèmes d'incomplétude de Gödel.

[174] Il semble tout naturel de supposer que si ces capacités humaines s'avéraient être correctes, elles devraient, d'une manière ou d'une autre, être impliquées dans la conscience.

[175] Kurt Gödel, *On Formally Undecidable Propositions of Principia Mathematica and Related Systems* (New York: Dover Publications, 1992 [1931]).

qui ne peut être démontré dans *F*. Évidemment, l'énoncé *G* est vrai puisqu'il est vrai que *G* ne peut être démontré ou réfuté dans *F*.[176]

La grande force de ce théorème est qu'il démontre non seulement que *F* est incomplet, mais également que *F* est *incomplétable.* En effet, nous pourrions penser qu'il suffit d'ajouter l'énoncé gödelien *G* au système *F* pour le compléter et pour que *G* soit, par le fait même, démontrable dans *F*. Par contre, le premier théorème de Gödel démontre qu'il serait tout de même possible de construire un autre énoncé gödelien (*G*') dans le système (*F*+*G*) qui serait vrai, mais qui ne pourrait toutefois pas être démontré dans *F*. Ainsi, n'importe quel système formel cohérent, assez puissant pour démontrer les bases de l'arithmétique est voué à être incomplet.[177] C'est sur ceci que se baseront Lucas et Penrose pour avancer l'impossibilité de la computabilité de l'esprit.[178]

Le second théorème d'incomplétude de Gödel avance que si *F* est un système formel cohérent dans lequel nous pouvons démontrer les bases de l'arithmétique, alors *F* ne peut pas démontrer la cohérence de *F*. C'est donc dire que le système *F* ne peut pas démontrer sa propre cohérence![179] Ce second théorème sera surtout important lorsque nous analyserons des réfutations possibles aux arguments de Lucas et de Penrose.

---

[176] Peter Smith, *An Introduction to Gödel's Theorems*, 2ème édition (Publié indépendamment, 2020), 161-166.
[177] Ernest Nagel et James R. Newman, *Godel's Proof*, édition revisitée (New York : New York University Press, 2008), 103-104.
[178] Torkel Franzén, *Gödel's Theorem: An Incomplete Guide to its Use and Abuse* (Massachussetts : A K Peters, 2005), 115-126.
[179] Peter Smith, *An Introduction to Gödel's Theorems*, 2ème édition (Publié indépendamment, 2020), 233-238.

3.5.2 L'argument gödelien de J.R. Lucas

En 1961, Lucas publie un article intitulé *Minds, Machines and Gödel*[180], qui allait constituer la base des discussions à venir concernant les conséquences des théorèmes de Gödel sur la théorie computationnelle de l'esprit et sur la possibilité de la conscience artificielle. Dans cet article, Lucas tente de démontrer que l'esprit humain sera toujours « meilleur » que la machine, i.e. que l'esprit humain a des capacités qui ne peuvent pas être reproduites par quelque machine que ce soit. Une de ces capacités est celle de pouvoir attribuer une valeur de vérité à un énoncé gödelien.

L'argument de Lucas va comme suit : imaginons une machine produite à partir des bases de l'arithmétique. Puisque cette machine est conçue à partir d'un nombre fini d'axiomes initiaux ainsi que d'un nombre fini d'opérations sur ces axiomes, il serait possible de les représenter sur une feuille de papier. En d'autres mots, si nous bénéficiions d'assez de temps (et de patience), il nous serait possible de représenter toutes les opérations effectuées par cette machine sur une feuille de papier. Lucas utilise cet argument dans le but de démontrer que le comportement de la machine est analogue à une système formel *S*, i.e. le système que nous produisons sur une feuille de papier. Ainsi, toutes les vérités qui peuvent être démontrées par la machine sont exactement les mêmes que celles qui peuvent être démontrées par le système formel *S* correspondant.

Lucas applique donc le premier théorème de Gödel à ce système formel *S* : il est possible de construire un énoncé gödelien *G*, tel que *G* n'est pas démontrable ni réfutable par *S*. En d'autres mots, le système formel *S* ne sera pas en mesure de produire *G* comme étant une vérité arithmétique. Lucas argumente cependant qu'il est possible pour l'esprit humain de regarder la feuille de papier (le système formel *S*) et de voir que l'énoncé *G* est vrai. Ainsi, selon Lucas, il y a donc au moins une chose que l'esprit

[180] John Randolph Lucas, « Minds, Machines and Gödel », *Philosophy* 36 (1961): 112-127.

humain peut faire, mais dont la machine est incapable, soit d'attribuer une valeur de vérité à l'énoncé gödelien *G*. La force de l'argument de Lucas, à l'instar de celle du théorème de Gödel, est qu'il démontre non seulement que le système formel analogue à la machine est incomplet, mais également qu'il est *incomplétable*; i.e. que même si nous ajoutions *G* aux axiomes initiaux de *S*, ce qui résulterait en un système plus complet (*S*+*G*), il nous serait tout de même possible de construire un nouvel énoncé *G'* que la machine ne pourrait pas démontrer ni réfuter, mais que l'esprit humain pourrait voir comme étant vrai. Ainsi, Lucas avance que l'esprit humain serait, en ce sens, toujours « supérieur » à la machine. Il en conclut donc que, puisque la machine ne peut pas reproduire toutes les capacités de l'esprit humain, la théorie computationnelle de l'esprit doit conséquemment être rejetée.[181]

3.5.3 L'argument gödelien de Roger Penrose

Roger Penrose va, quelque sorte, formuler à son tour un argument gödelien contre la théorie computationnelle de l'esprit similaire en plusieurs points à celui de Lucas, dans un premier temps dans « The Emperor's New Mind »[182], puis dans « Shadows of the Mind ».[183][184] Tout comme Lucas, Penrose tente de démontrer que l'esprit humain n'est pas *computable*.

L'argument de Penrose va comme suit : supposons que mes capacités de raisonnements soient capturées par un système formel *F* et que ce système soit adéquat (sound). Il serait donc possible de construire un énoncé gödelien *G* à partir de *F*. L'esprit humain serait en mesure de déterminer la valeur de vérité d'un tel énoncé; i.e. qu'il serait capable d'affirmer que cet énoncé est vrai. Inversement, comme le démontre

[181] John Randolph Lucas, « Minds, Machines and Gödel », *Philosophy* 36 (1961): 116.
[182] Roger Penrose, *The Emperor's New Mind* (Oxford: Oxford University Press, 1989)
[183] Roger Penrose, *Shadows of the Mind*, (Oxford: Oxford University Press, 1994)
[184] Puisque « Shadows of the Mind » est, en quelque sorte, une version améliorée de son argument présenté dans « The Emperor's New Mind », je me concentrai surtout sur la formulation de son argument qui se trouve dans « Shadows of the Mind ».

le théorème de Gödel, le système *F* lui-même ne serait pas en mesure de démontrer ni de réfuter l'énoncé *G*. Par transitivité, le système formel *F* devrait également être en mesure de déterminer la valeur de vérité de l'énoncé gödelien, puisque *F* capture les capacités de raisonnements de l'esprit humain. Nous arrivons donc à une contradiction, puisque le système formel *F* est à la fois capable et incapable de déterminer la valeur de vérité d'un tel énoncé. Penrose en conclut donc que nous devons rejeter l'hypothèse initiale qui était qu'un système formel *F* puisse capturer les capacités de raisonnement de l'esprit humain. Ainsi, selon Penrose, l'esprit humain ne serait pas computable, puisqu'aucune machine ne serait en mesure d'en capturer les capacités de raisonnement; il faudrait donc, par le fait même, abandonner la théorie computationnelle de l'esprit.[185]

3.5.4 Réponses à l'argument gödelien de Lucas-Penrose

Une objection évidente qui pourrait être soulevée serait de dire que le système formel qu'est l'esprit humain n'est pas cohérent. Si c'était le cas que l'esprit humain n'est pas un système cohérent, alors le premier théorème d'incomplétude de s'appliquerait donc pas. Marvin Minsky argument qu'il semble faux d'affirmer que l'esprit humain est nécessairement cohérent, notamment car il est possible pour celui-ci de croire que certaines idées sont vraies.[186] C'est donc dire que l'esprit humain est en mesure de déterminer la valeur de vérité de l'énoncé de Gödel non pas parce qu'il est, en quelque sorte, « supérieur » à la machine (ou encore, qu'il n'est tout simplement pas une machine), mais bien parce qu'il ne s'agit pas d'un système cohérent; ce qui est une condition nécessaire pour que le théorème d'incomplétude s'applique. C'est donc pourquoi l'esprit humain serait en mesure de déterminer la valeur d'un énoncé gödelien *G*, alors qu'une machine (un système cohérent) en serait incapable. Plusieurs versions

[185] Ceci correspond au résumé de l'argument gödelien formulé par David Chalmers dans son article « Minds, Machines, and Mathematics » paru en 1995.
[186] Marvin Minsky « Conscious machines », dans *Machinery of Consciousness* (National Research Council of Canada, *75th Anniversary Symposium on Science in Society*, 1991).

de cet argument ont été proposées, notamment par Graham Priest[187] et J.R. Lucas[188] lui-même.

D'autres objections s'attardent à la condition de cohérence qui est nécessaire dans les théorèmes d'incomplétude de Gödel.[189][190][191][192] Ces arguments consistent à dire qu'il ne nous ait pas possible déterminer si notre propre esprit est cohérent. La conclusion de ces arguments est quelque peu moins forte que celle de Priest, puisqu'au lieu d'affirmer que l'esprit humain n'est pas cohérent, ceux-ci se contentent d'affirmer que nous ne sommes pas en mesure de déterminer si notre esprit est cohérent. Ces arguments se basent notamment sur le second théorème d'incomplétude de Gödel puisque ce dernier démontre que si un système est cohérent, alors il ne peut pas démontrer sa propre cohérence. C'est donc dire que si notre esprit humain est cohérent, alors il ne peut pas démontrer sa propre cohérence, et s'il est incohérent, alors, trivialement, il n'est pas cohérent. Puisque nous ne sommes pas en mesure de déterminer si notre esprit est bel et bien cohérent, alors nous ne pouvons pas être certains que le premier théorème de Gödel s'applique à notre esprit, et donc, nous ne pouvons pas accepter la conclusion de l'argument gödelien de Lucas-Penrose avec certitude.

---

[187] Graham Priest, « Inconsistent Arithmetic: Issues Technical and Philosophical », dans *Trends in Logic: 50 Years of Studia Logica,* éd. Vincent F. Hendricks and Jacek Malinowski (Dordrecht: Kluwer Academic Publishers, 2003), 277-299.
[188] John Randolph Lucas, « Minds, Machines and Gödel », *Philosophy* 36 (1961): 112-127.
[189] Hartley Rogers, *Theory of Recursive Functions and Effective Computability* (Massachussetts : MIT Press, 1957).
[190] Hilary Putnam, « Minds and Machines», dans *Dimensions of Minds*, éd. Sidney Hook (New York : New York University Press, 1960), 138-164.
[191] Anthony Hutton, « This Gödel is Killing Me », *Philosophia* 3 (1976): 135-44.
[192] Geoffrey LaForte, Patrick Hayes et Kenneth Ford, « *Why Gödel's Theorem Cannot Refute Computationalism* », *Artificial Intelligence*, 104 (1998): 265–286

Dans le même ordre d'idées, David Chalmers[193] et Daryl McCullough[194] argumentent qu'il n'est pas aussi évident que Penrose peut le prétendre que (1) l'esprit humain est cohérent et (2) qu'il est possible pour l'esprit humain de démontrer sa propre cohérence.[195]

En 1967, Paul Benacerraf publie un article intitulé « God, the Devil, and Gödel»[196], dans lequel il offre plusieurs objections à l'argument de Lucas. L'objection pour laquelle il est certainement le plus connu en est une qui stipule que l'esprit humain est trop complexe pour qu'il soit possible de construire un énoncé de Gödel. En effet, pour construire un énoncé de Gödel pour un système formel donné, il faut avoir une très bonne compréhension de ce système. Or, il semble que l'esprit humain soit un système tellement complexe qu'il ne serait pas possible de le comprendre assez bien pour être en mesure de construire un énoncé de Gödel et de déterminer la valeur de vérité de cet énoncé. Bref, l'esprit humain serait une machine infiniment plus complexe que les systèmes pour lesquels nous sommes en mesure de construire un énoncé de Gödel, mais il serait néanmoins une machine.

Cet argument de la complexité pourrait également s'appliquer à notre capacité de savoir si le système *F* est cohérent, pour que le théorème de Gödel puisse s'appliquer. En effet, pour que l'énoncé *G*, il faut que le système *F* soit cohérent, mais comment est-il possible pour nous de déterminer qu'un tel système est cohérent; cette tâche s'avère d'autant plus difficile (voire impossible) lorsque que le système en question est d'une extrême complexité, comme ce serait le cas si ce système encodait

[193] David Chalmers, « Minds, Machines, and Mathematics », *Psyche* 2 (1995): 11-20.
[194] Daryl McCullough, « Can Humans Escape Gödel? », *Psyche* 2 (1996): 57-65.
[195] Je considère cet argument comme étant très convaincant, puisqu'il se base sur le second théorème d'incomplétude de Gödel qui est, en quelque sorte, dérivé de son premier théorème; si nous voulons proposer un argument qui se base sur la premier théorème (comme c'est le cas de Lucas et de Penrose), nous n'avons pas le choix de prendre en considération le second théorème, qui semble miner la force et la validité de tout argument qui prend racine dans le premier.
[196] Paul Benacerraf, « God, the Devil, and Gödel », *Monist* 51 (1967): 9-32.

toute notre identité et notre vie cognitive. Il peut être assez facile de déterminer (même *intuitivement*) qu'un certain système donné peu complexe est cohérent; par exemple, si nous regardons une liste d'axiomes de Peano. Par contre, cela s'avère particulièrement difficile à faire dans le cas d'une extrême complexité comme celui dont il est question dans l'argument gödelien contre la théorie computationnelle de l'esprit. C'est cet argument que Haim Gaifman propose et développe dans son article « What Gödel's Incompleteness Does and Does Not Show ».[197][198]

De son côté, Charles Henry Whiteley répond à Lucas en avançant que l'esprit humain est limité d'une façon similaire à la machine; l'esprit humain et la machine ne seraient donc pas aussi différents que peut le prétendre Lucas.[199] Cette objection a, en quelque sorte, été reprise par Douglas Hofstadter.[200] L'argument est qu'il est possible de construire un énoncé de Whiteley, — par exemple, « Lucas ne peut pas exprimer cet énoncé tout en étant cohérent » — qui pose le même genre de problème à l'esprit humain que l'énoncé de Gödel à la machine. Je considère que cette objection est moins convaincante et/ou pertinente que les autres, puisqu'elle ne s'attaque pas directement au problème posé par Lucas, i.e. qu'il y a des choses que l'esprit humain peut faire, que la machine ne peut pas faire, et se contente d'affirmer que les deux sont limités de

[197] Haim Gaifman, « What Gödel's Incompleteness Result Does and Does Not Show », *The Journal of Philosophy* 97, (8) (2000): 462-70.
[198] Ces contre-arguments respectifs de Benacerraf et de Gaifman sont intéressants mais je ne crois pas qu'ils posent un grand problème à l'argument gödelien de Lucas-Penrose. En effet, je considère que le fait que l'esprit soit complexe ne fait pas en sorte que le premier théorème d'incomplétude ne s'applique pas. J'estime que même si nous ne sommes pas, en pratique, en mesure de construire un énoncé gödelien en raison de la complexité du système qu'est l'esprit humain, il me paraît tout de même sensé de penser qu'en principe, i.e. si nous avions assez de temps et de patience, il nous serait possible de construire cet énoncé. En d'autres mots, j'avance que ce n'est pas parce que le système est complexe que cela fait en sorte qu'il passe au-dessus du théorème de Gödel et de ses implications. C'est donc dire que, pour n'importe quel système, aussi complexe soit-il, le théorème de Gödel devrait tout de même s'appliquer.
[199] Charles Whiteley, « Minds, Machines and Gödel: A Reply to Mr. Lucas ». *Philosophy* 37 (1962): 61-62.
[200] Daniel Dennett et Douglas Hofstadter, *The Mind's I: Fantasies and Reflections on Self and Soul* (New York : Basic Books, 1981).

manière similaire; il ne me semble pas évident que, puisque les deux sont limités de manière similaires, l'esprit humain soit nécessairement une machine.

Pour toutes les raisons explorées ci-dessus, j'estime que l'argument gödelien de Lucas-Penrose n'est pas convaincant. Évidemment, il existe de nombreuses autres objections à cet argument ainsi que de nombreuses réponses à celles-ci; j'ai choisi ces arguments dans le but de circonscrire la grande majorité des objections qui ont été avancées et/ou qui pourraient être avancées contre leur argument, et ainsi, avoir une vue d'ensemble du débat entourant l'argument gödelien de Lucas-Penrose contre la théorie computationnelle de l'esprit.

Ceci étant dit, comme je l'ai mentionné, l'argument gödelien de Lucas-Penrose avance que la machine possède certaines limites que l'esprit humain ne possède pas, élevant ainsi à un niveau « supérieur » ce dernier. En effet, ils avancent que la machine possède des limites que ne possède pas l'esprit. Par contre, ce n'est pas tout à fait clair que ces limites se situent au niveau de la conscience. Il me semble plutôt les limites évoquées se situent au niveau du « raisonnement » ou encore de la pensée, plutôt que de la conscience. C'est donc dire que, selon l'argument de Lucas-Penrose, l'esprit humain pourrait ne pas être reproductible par une machine, mais cela n'implique pas pour autant que la conscience ne pourrait pas l'être : il me semble tout à fait plausible de penser qu'il puisse être possible qu'une machine ne puisse pas être capable de voir la valeur de vérité d'un certain énoncé gödelien tout en étant conscient. De plus, si nous posons comme condition nécessaire à la conscience le fait de pouvoir évaluer la valeur de vérité d'un énoncé gödelien, nous excluons par le fait même plusieurs individus et espèces qui, bien que conscients, ne sont pas en mesure de faire une telle chose.

CONCLUSION

Pour résumer, après avoir caractérisé la conscience et souligné quelques problèmes qui découlent directement de la particularité de la conscience, j'ai défendu, dans le deuxième chapitre, la thèse selon laquelle la conscience artificielle est possible. Pour défendre cette position, je me suis, entre autres, basé sur le principe d'invariance organisationnelle. Selon celui-ci, si deux systèmes ont la même organisation fonctionnelle, alors ces deux systèmes auront la même expérience consciente. C'est donc dire que si un système artificiel (par exemple, un ordinateur) avait la même organisation fonctionnelle qu'un système conscient (par exemple, le système biologique qu'est le cerveau humain), alors il faudrait en conclure que celui est tout aussi conscient. Après avoir évalué quelques objections spécifiques au principe d'invariance organisationnelle, je me suis tourné, dans le chapitre 3 vers des objections plus générales quant à la possibilité de la conscience artificielle, avant de répondre tour à tour à celles-ci, dans le but de maintenir ma thèse initiale aussi intacte que possible.

Maintenant, si la conscience artificielle est véritablement possible, nous devrons faire face à plusieurs dilemmes éthiques et questionnements moraux en lien avec notre utilisation de ces technologies qui pourraient être conscientes. J'estime qu'il serait judicieux de se poser ces questions avant qu'un ordinateur, un robot ou une machine ne développe une conscience, à moins qu'il ne soit déjà trop tard et que certaines technologies ne soient déjà conscientes…

BIBLIOGRAPHIE

INTRODUCTION

Dainton, Barry, et Tim Bayne. « Consciousness as a guide to personal persistence », *Australasian Journal of Philosophy* 83 (4) (2005): 549-571.

Gibert, Martin, et Dominic Martin. « In search of the moral status of AI: why sentience is a strong argument », *AI and Society* 37 (2022) : 319-330.

Himma, Kenneth Einar. « Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? », *Ethics and Information Technology* 11 (2009): 19-29.

Levy, Neil. *Consciousness and Moral Responsibility*. Oxford: Oxford University Press, 2014.

Véliz, Carissa. « Moral zombies: why algorithms are not moral agents », *AI and Society* 36 (2021) : 487-497.

Wallach, Wendell, et Colin Allen, *Moral Machines : Teaching Robots Right from Wrong.* Oxford : Oxford University Press, 2010.

CHAPITRE I

Block, Ned. « Psychologism and behaviorism », *Philosophical Review* 90 (1) (1981): 5-43.

Block, Ned. « Some concepts of consciousness », dans *Philosophy of Mind: Classical and Contemporary Readings*, édité par David Chalmers, 206-219. Oxford, Oxford University Press, 2002.

Block, Ned. « Troubles with functionalism », *Minnesota Studies in the Philosophy of Science* 9 (1978): 261-325.

Chalmers, David. *The Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford University Press, 1996.

Chalmers, David. « The hard problem of consciousness », dans *The Blackwell Companion to Consciousness*, édité par Max Velmans et Susan Schneider. New York :Wiley-Blackwell, 2017.

Elamrani, Aida, et Roman Yampolskly. « Reviewing Tests for Machine Consciousness », *Journal of Consciousness Studies* 26 (5-6) (2019): 35-64.

Frankish, Keith. « Illusionism as a Theory of Consciousness », *Journal of Consciousness Studies* 23 (2016): 11-39.

Goff, Philip. « The Case for Panpsychism ». *Philosophy Now* 121 (2017): 6-8.

Harnard, Steven. « Can a Machine Be Conscious? How? », *Journal of Consciousness Studies* 10 (2003) : 67-75.

Jansen, Franz Klaus. « The Hard Problem of Consciousness from a Bio-Psychological Perspective », *Philosophy Study* 7 (11) (2017): 579-594.

Kirk, Robert. *Zombies and Consciousness*. Oxford : Oxford University Press, 2005.

Levine, Joseph. « Materialism and Qualia: The Explanatory Gap », *Pacific Philosophical Quarterly* 64 (1983): 354-61.

Nagel, Thomas. « What is it Like to be a Bat ? », *Philosophical Review* 83 (1974) : 435-50.

Nani, Andrea, Jordi Manuello, Lorenzo Mancuso, Donato Liloia, Tommaso Costa et Franco Cauda. « The Neural Correlates of Consciousness and Attention: Two Sister Processes of the Brain », *Frontiers in Neuroscience* 13 (2019), https://doi.org/10.3389/fnins.2019.01169

Pinker, Steven. « The mystery of consciousness », *Time* 169 (2007) : 58-62, 65.

Reggia, James. « The rise of machine consciousness: Studying consciousness with computational models », *Neural Networks* 44 (2013) : 115.

Reynolds, Jack. « Problems of other minds: Solutions and dissolutions in analytic and continental philosophy », *Philosophy Compass* 5 (4) (2010): 326-335.

Schneider, Susan. *Artificial You : AI and the Future of your* Mind. Princeton et Oxford : Princeton University Press, 2019

Searle, John. *Mind: A Brief Introduction.* Oxford : Oxford University Press, 2004.

Searle, John. « Minds, brains, and programs », *Behavioral and Brain Sciences* 3 (3) (1980): 417-57.

Stalnaker, Robert. « What is it like to be a zombie? », dans *Conceivability and Possibility,* édité par Tamar Szabo Gendler et John Hawthorne, 385-400. Oxford : Oxford University Press, 2002.

Stins, John. « Establishing consciousness in non-communicative patients: A modern-day version of the Turing test », *Consciousness and Cognition* 18 (1) (2009): 187-192.

Turing, Alan. « Computing Machinery and Intelligence », *Mind* 59 (1950): 434-60.

Udell, David Billy, et Eric Schwitzgebel. « Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed », *Journal of Consciousness Studies* 28 (5-6) (2021): 121-144.

CHAPITRE II

Beniaguev, David, et Idan Segev, Michael London. « Single cortical neurons as deep artificial neural networks », *Neuron* 109 (17) (2021), 2727-2739.

Block, Ned. « How to Find the Neural Correlate of Consciousness », *Royal Institute of Philosophy Supplement* 43 (1998): 23-34.

Block, Ned. « Sexism, ageism, racism, and the nature of consciousness », *Philosophical Topics* 26 (1-2) (1999) : 39-70.

Block, Ned. « Troubles with functionalism », *Minnesota Studies in the Philosophy of Science* 9 (1978): 261-325.

Brunet, Tim, et Marta Halina. « Minds, Machines, and Molecules », *Philosophical Topics* 48 (1) (2020): 221-42.

Casati, Roberto. « What is Wrong in Inverting Spectra ? », *Theoria* 10 (1990) : 183-186.

Chalmers, David. « *Absent qualia, fading qualia, dancing qualia* », dans *Conscious Experience, édité par* Thomas Metzinger, 309-328. Ferdinand Schoningh, 1995.

Chalmers, David, *The Conscious Mind: In Search of Fundamental theory*. New York: Oxford University Press : 1996.

Churchland, Paul, et Patricia Churchland. « Functionalism, qualia and intentionality », *Philosophical Topics* 12 (1) (1981) : 121-145.

Crick, Francis, et Christof Koch. « Towards a neurobiological theory of consciousness », *Semin. Neurosci.* 2 (1990) : 263–275.

Duncan, Stewart. « Leibniz's Mill Arguments Against Materialism », *Philosophical Quarterly* 62 (247) (2012): 250-72.

Elugardo, Reinaldo. « Functionalism and the Absent Qualia Argument », *Canadian Journal of Philosophy* 13 (2) (1983):161-179.

Godfrey-Smith, Peter. « Mind, Matter, and Metabolism », *Journal of Philosophy* 113 (10) (2016): 481-506.

Hardin, Clyde. « Qualia and materialism : Closing the explanatory gap », *Philosophy and Phenomenological Research* 48 (2) (1987) : 281-298.

Horgan, Terence. « Functionalism, Qualia, and the Inverted Spectrum », *Philosophy and Phenomenological Research* 44 (1984) : 453-470.

Koch, Christof, et Marcello Massimini, Melanie Boly, *et al.* « Neural correlates of consciousness: progress and problems », *Nat Rev Neurosci* 17 (2016): 307-321.

Lycan, William, Consciousness. Cambridge, Massachusetts : MIT Press, 1987.

Meyer, Uwe. « Do Pseudonormal Persons Have Inverted Qualia? Scientific Hypotheses and Philosophical Interpretations », *Facta Philosophica*, 2 (2000): 309-325.

Nida-Rümelin, Martine. « Pseudonormal Vision : An Actual Case of Qualia Inversion? », *Philosophical Studies* 82 : 145-157.

Pylyshyn, Zenon. « The 'causal power' of machines », *Behavioral and Brain Sciences* 3 (3) (1980): 442-444.

Sayan, Erdinç. « A Closer Look at the Chinese Nation Argument », *Philosophy Research Archives* 13 (1987): 129-136.

Schwitzgebel, Eric. « If materialism is true, the United States is probably conscious », *Philosophical Studies* 172 (7) (2015): 1697-1721.

Shoemaker, Sydney. « Functionalism and qualia », *Philosophical Studies* 27 (1975): 291-315.

Tye, Michael. « Absent Qualia and the Mind-Body Problem », *Philosophical Review* 115 (2) (2006) : 139-168.

Tye, Michael. « Qualia, Content, and the Inverted Spectrum », *Noûs* 28 (2) (1994): 159-83.

CHAPITRE III

Benacerraf, Paul. « God, the Devil, and Gödel », *Monist* 51 (1967): 9-32.

Block, Ned. « Searle's Arguments Against Cognitive Science », dans *View Into the Chinese Room : New Essays on Searle and Artificial* Intelligence, edité par John M. Preston et John Mark Bishop, 70-79. Londres : Oxford University Press, 2003.

Chalmers, David. « Minds, Machines, And Mathematics : A Review of Shadows of the Mind by Roger Penrose », *PSYCHE: An Interdisciplinary Journal of Research On Consciousness* 2 (1995):11-20.

Chalmers, David. *The Conscious Mind: In Search of Fundamental theory*. New York: Oxford University Press, 1996.

Cole, David. « Artificial Intelligence and Personal Identity », *Synthese* 88 (1991): 399-417.

Copeland, Jack. « The Chinese Room from a Logical Point of View », dans *View Into the Chinese Room : New Essays on Searle and Artificial* Intelligence, edité par John M. Preston et John Mark Bishop, 109-122. Londres : Oxford University Press, 2003.

Dennett, Daniel, et Douglas Hofstadter. *The Mind's I: Fantasies and Reflections on Self and Soul*. New York : Basic Books, 1981.

Dreyfus, Hubert, et Stuart Dreyfus. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* New York: The Free Press, 1986.

Dreyfus, Hubert. *What Computers Can't Do*. New York: MIT Press, 1972.

Dreyfus, Hubert. *What Computers Still Can't Do*. New York: MIT Press, 1992.

Fjelland, Ragnar. « Why general artificial intelligence will not be realized », *Humanit Soc Sci Commun* 7 (10) (2020) : 1-9.

Franzén, Torkel. *Gödel's Theorem: An Incomplete Guide to its Use and Abuse*. Massachussetts : A K Peters, 2005.

Gaifman, Haim. « What Gödel's Incompleteness Result Does and Does Not Show », *The Journal of Philosophy* 97, (8) (2000): 462-70.

Gödel, Kurt. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. New York: Dover Publications, 1992 [1931].

Hassabis, Demis, et David Siver. « AlphaGo Zero : Learning from Scratch », *DeepMind*, publié le 18 octobre 2017, https://www.deepmind.com/blog/alphago-zero-starting-from-scratch.

Haugeland, John. « Syntax, Semantics, Physics », dans *View Into the Chinese Room : New Essays on Searle and Artificial* Intelligence, edité par John M. Preston et John Mark Bishop, 379-392. Londres : Oxford University Press, 2003.

Hayes, Patrick, Stevan Harnad, Donald Perlis et Ned Block, « Virtual Symposium on Virtual Mind », *Minds and Machines*, 2(3) (1992): 217-238.

Hofstadter, Douglas. « Reflections on Searle », dans *The Mind's I*, edité par Douglas Hofstadter Daniel Dennett, 373-382. New York: Basic Books, 1981.

Hutson, Matthew. « This computer program can beat humans at Go with no human instruction », *Science*, publié le 18 octobre 2017, https://www.science.org/content/article/computer-program-can-beat-humans-go-no-human-instruction.

Hutton, Anthony. « This Gödel is Killing Me », *Philosophia* 3 (1976): 135-44.

Johnson, George. « To Test a Powerful Computer, Play an Ancient Game », *The New York Times*, publié le 29 juillet 1997, https://www.nytimes.com/1997/07/29/science/to-test-a-powerful-computer-play-an-ancient-game.html.

Kaplan, Jerry. *Artificial Intelligence: What Everyone Needs to Know*. Oxford University Press, 2016.

Kaur, Paramjit, Kewal Krishan, Suresh K. Sharma et Tanuj Kanchan, « Facial-recognition algorithms : A literature review », *Medicine, Science and the Law* 60 (2) (2020) : 131-139.

Koch, Christof. *The Feeling of Life Itself: Why Consciousness is Widespread but Can't Be Computed*. Massachusetts: MIT Press, 2019.

Kohs, Greg, réal. *AlphaGo*. É-U : Movie Pictures, 2017. Netflix

LaForte, Geoffrey, Patrick Hayes et Kenneth Ford, « *Why Gödel's Theorem Cannot Refute Computationalism* », *Artificial Intelligence*, 104 (1998): 265–286.

Lucas, John Randolph. « Minds, Machines and Gödel », *Philosophy* 36 (1961): 112-127.

MacFarland, Matt. « Google AI defeats world's Go champion in gripping 'man vs machine' film », *CNN Business*, publié le 29 septembre 2017, https://money.cnn.com/2017/09/29/technology/future/alphago-movie/index.html.

McCullough, Daryl. « Can Humans Escape Gödel? », *Psyche* 2 (1996): 57-65.

Minsky, Marvin. « Conscious machines », dans *Machinery of Consciousness*. National Research Council of Canada, *75th Anniversary Symposium on Science in Society*, 1991.

Müller, Karsten, et Jonathan Schaeffer. *Man vs Machine : Challenging Human Supremacy at Chess*. Connecticut : Russell Enterprises Inc., 2018.

Nagel, Ernest, et James R. Newman. *Godel's Proof*, édition revisitée. New York : New York University Press, 2008.

Penrose, Roger. *Shadows of the Mind*. Oxford: Oxford University Press, 1994.

Penrose, Roger. *The Emperor's New Mind*. Oxford: Oxford University Press, 1989.

Pfeiffer, Michael, et Pfeil Thomas, « Deep Learning With Spiking Neurons : Opportunities and Challenges », *Frontiers in Neuroscience* 12 (2018).

Priest, Graham. « Inconsistent Arithmetic: Issues Technical and Philosophical », dans *Trends in Logic: 50 Years of Studia Logica,* édité par Vincent F. Hendricks and Jacek Malinowski, 277-299. Dordrecht: Kluwer Academic Publishers, 2003.

Putnam, Hilary. « Minds and Machines», dans *Dimensions of Minds*, édité par Sidney Hook, 138-164. New York : New York University Press, 1960.

Rey, Georges. « What's Really Going on in Searle's 'Chinese Room' », *Philosophical Studies* 50 (1986): 169-85.

Rogers, Hartley. *Theory of Recursive Functions and Effective Computability*. Massachussetts: MIT Press, 1957.

Searle, John. « Minds, Brains and Programs », *Behavioral and Brain Sciences* 3 (3) (1980): 417-457.

Seth, Anil. *Being You : A New Science of* Consciousness. New York: Dutton, 2021.

Sloman, Aaron, et Monica Croucher. « How to turn an information processor into an understanding », *Brain and Behavioral Sciences* 3 (1980): 447-8.

Smith, Peter. *An Introduction to Gödel's Theorems*, 2ème édition. Publié indépendamment, 2020.

Thompson, Benjamin, et Noah Baker. « Lab-grown brains and the debate over consciousness », *Nature*, publié le 28 octobre 2020. https://www.nature.com/articles/d41586-020-03033-6#:~:text=Anil%20Seth%20works%20on%20cognitive,to%20that%20of%20conscious%20humans.

Whiteley, Charles. « Minds, Machines and Gödel: A Reply to Mr. Lucas », *Philosophy* 37 (1962): 61-62.