Année universitaire 2023-2024

Les IA rêveront-elles un jour de moutons électriques ?

Putnam, McDowell et le projet d'une intelligence artificielle

Présenté par Bastien Goumy

Sous la direction de M. Manuel Rebuschi

UNIVERSITE DE LORRAINE

Année universitaire 2023-2024

Les IA rêveront-elles un jour de moutons électriques ?

Putnam, McDowell et le projet d'une intelligence artificielle

Présenté par Bastien Goumy
Sous la direction de M. Manuel Rebuschi

Mémoire présenté le 04/07/2024 devant un jury composé de
Baptiste Mélès, CR CNRS
Anna C. Zielinska, MCF UL
Manuel Rebuschi, MCF HDR UL

[Le perceptron] est l'embryon d'un ordinateur électronique qui devrait être capable de marcher, de parler, de voir, d'écrire, de se reproduire et d'être conscient de son existence.

The New York Times à propos des travaux de Frank Rosenblatt, 1957

L'esprit humain n'est pas, à l'instar de ChatGPT et de ses semblables, un moteur statistique encombrant pour la recherche de modèles, se nourrissant de centaines de téraoctets de données et extrapolant la réponse conversationnelle la plus probable ou la réponse la plus vraisemblable à une question scientifique. Au contraire, l'esprit humain est un système étonnamment efficace et même élégant qui fonctionne avec de petites quantités d'informations.

Noam Chomsky, 2023

Remerciements

Je tiens à remercier mon directeur de mémoire, M. Manuel Rebuschi pour sa patience et l'aide qu'il m'a apportée dans mes réflexions et la rédaction.

Je tiens aussi à remercier ma famille et mes amis qui ont eu la patience de supporter mes élucubrations philosophiques.

Je remercie le programme ORION pour sa contribution au financement de mon stage de recherche. Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du Programme d'Investissements d'Avenir portant la référence ANR-20-SFRI-0009.

Table des matières

Introduction

Recréer un esprit artificiel grâce à la technologie et aux ordinateurs. Au-delà d'un scénario de science-fiction devenu plutôt commun durant ces dernières décennies, c'est aussi et surtout une ambition bien réelle. L'essor de l'informatique durant le XX^{e} siècle a, en effet, laissé entrevoir la possibilité de reproduire, ou de créer, un esprit humain. Cette ambition fut et est encore très débattue. Il y a un questionnement moral, bien sûr. Nous pouvons nous demander par exemple si nous avons vraiment besoin de robots ou d'ordinateurs doués de conscience ou si cela n'entraînerait pas des conséquences néfastes sur la société. Mais, il y a également des questionnements plus conceptuels et pratiques concernant la faisabilité d'un tel projet. En effet, la question de savoir si l'esprit peut être reproduit se pose d'emblée. Il faut donc se demander quelle est la nature de l'esprit et s'il est possible d'obtenir une description précise de tous ses mécanismes et processus afin que ces derniers puissent être reproductibles par les ordinateurs. Ces questionnements et débats se sont structurés autour d'un programme de recherche, apparu au début de la seconde moitié du XX^{ème} siècle et qui existe encore aujourd'hui.

L'objectif prométhéen de l'Intelligence Artificielle de reproduire un esprit humain était cependant quelque peu tombé en désuétude, du moins pour le grand public. Nous étions encore loin des ordinateurs conscients et il semblait y avoir peu de progrès réalisés. Le projet ambitieux de l'Intelligence Artificielle semblait irréalisable. Or, depuis les années 2010, un regain d'intérêt a eu lieu. En effet, l'IA de type *deep learning* réalisa de nouveaux exploits et se mit à faire des progrès fulgurants. Aujourd'hui encore, nous sommes témoins d'IA de plus en plus puissantes, performantes et réalisant des tâches de plus en plus complexes. L'IA se trouve désormais à nos côtés au quotidien et il semble que cette tendance s'accentuera avec le temps. On retrouve l'IA dans nos ordinateurs, dans nos téléphones, nos appareils électroménagers et même nos voitures. Alors, naturellement, ces progrès fulgurants reposent la question de la création d'un esprit artificiel. Est-ce que, désormais, un tel projet est réalisable ?

C'est ce que nous allons nous demander dans cet écrit et, pour mener à bien notre réflexion, il faut se questionner sur la nature de l'esprit humain afin de savoir comment devrait être l'IA parfaite.

Un modèle de l'esprit était sous-entendu dans le programme de recherche en IA, du moins dans ses débuts. En effet, la recherche en Intelligence Artificielle a vu le jour à une époque où le

naturalisme et la volonté de tout expliquer par le biais des sciences s'est grandement développée et répandue. Le programme de recherche est donc apparu conjointement à la pensée d'Hilary Putnam qui a mis en place en 1967, dans son article « La nature des états mentaux »[1], une théorie du fonctionnement de l'esprit qui se nomme le fonctionnalisme computationnel (ou computationnalisme). Selon cette théorie, l'esprit serait analogue à un ordinateur. Pour Putnam, l'esprit est au cerveau ce que le *software*, le logiciel d'un ordinateur, est au *hardware*, le matériel qui constitue l'ordinateur. Cette théorie fut très importante pour le programme de recherche en Intelligence Artificielle car elle supposait que l'esprit humain était une machine de Turing, c'est-à-dire une machine bien spécifique capable d'exécuter certaines tâches, en suivant un certain nombre d'instructions, des algorithmes, qui permettent de traiter des informations en entrée pour en fournir en sortie, tout comme un ordinateur. Les ordinateurs sont, en effet, également des machines de Turing. Ainsi, si les humains ne sont que des machines recevant certaines données en entrée et qui, en fonction de leur état actuel, créent des réactions comportementales et fournissent des informations en sortie tout en basculant dans un nouvel état, alors un ordinateur devrait pouvoir reproduire ce genre de machine. Ce présupposé fut fondamental pour le lancement du programme de recherche. Le cerveau humain et les machines exécutant des algorithmes étaient considérés, à l'époque, comme ayant un fonctionnement proche, voire identique. C'est donc une conception de l'esprit qui mérite toute notre attention.

Mais ce n'est pas la seule qui sera étudiée. En effet, nous n'allons pas considérer le computationnalisme comme étant le réel fonctionnement de l'esprit humain car, il s'avère qu'au fil des années, la pensée de Putnam a changé. Elle a changé à tel point qu'il finira même par rejeter sa propre théorie (donc le computationnalisme) et la conception réaliste dans laquelle elle s'inscrit. Pour Putnam, le fonctionnalisme est faux notamment parce que notre manière de concevoir notre rapport à la réalité est fausse. Ce changement de pensée est très intéressant, surtout étant donné le lien qu'il y a entre le fonctionnalisme computationnel et l'Intelligence Artificielle. Cela est d'autant plus important à étudier que, si le computationnalisme est faux, alors cela veut dire que le cerveau ne fonctionne pas de manière analogue à un ordinateur. Cela voudrait-il dire que l'objectif prométhéen de l'Intelligence Artificielle est voué à l'échec ?

Putnam a été nourri de diverses influences pour la mise en place de sa nouvelle conception du monde. L'une d'entre elles c'est le philosophe John McDowell, qui propose un modèle de l'esprit

[1] Hilary Putnam, « La Nature Des États Mentaux », *Les Études philosophiques*, n° 3 (1992): 323-35.

qui s'enracine dans une nouvelle forme de naturalisme et de réalisme. Le modèle de l'esprit de McDowell se démarque des autres qui existaient jusqu'alors et, comme Putnam malgré des désaccords sur certains points, nous allons le considérer comme étant la bonne description du fonctionnement de l'esprit.

Une question se pose alors. En effet, depuis le rejet du fonctionnalisme de Putnam, la recherche en IA a continué et a progressé pour en arriver aux résultats que nous voyons aujourd'hui. Or, le modèle de McDowell a permis le rejet du fonctionnalisme qui était pourtant conceptuellement lié au programme de recherche en IA, qui rendait possible la création d'un esprit humain artificiel. Ainsi, pour réaliser le projet prométhéen de reproduire un esprit humain il faut donc se demander : est-ce que l'IA peut reproduire un esprit tel qu'il est pensé par McDowell ?

Il faut que nous nous demandions si les différents types d'IA arrivent à reproduire les différentes capacités cognitives de l'esprit humain telles qu'elles sont théorisées par McDowell et, si c'est le cas, lesquelles. Les convictions de Putnam sur le fait que l'humain était analogue à un ordinateur n'ont pas résisté à la conception mcdowellienne de l'esprit, alors regardons ensemble si l'IA d'aujourd'hui, elle, résiste ou non.

Pour ce faire, nous allons d'abord étudier le modèle de l'esprit de McDowell et son cheminement de pensée. Cela nous amènera fatalement à faire le lien avec l'évolution de la pensée de Putnam qui finira dans par être un « réaliste naturel ». Il faudra également que nous comprenions en quoi ce réalisme est antagoniste au computationnalisme qu'il avait précédemment théorisé, donc antagoniste à la conception de l'esprit sous-entendu lors du lancement du programme de recherche en IA, ce qui nous permettra de mieux cerner la position du modèle de l'esprit de McDowell par rapport à ce programme de recherche. Ensuite nous pourrons étudier les deux principaux modèles d'intelligence artificielle présents aujourd'hui, qui sont l'IA symbolique et l'IA connexionniste, que ce soit leur histoire, leurs caractéristiques et leurs défauts. Nous pourrons alors argumenter sur les limites que rencontrent ces IA quant à la reproduction d'un esprit humain et conclure que l'IA n'est pas compatible avec l'esprit tel qu'il est conçu par McDowell.

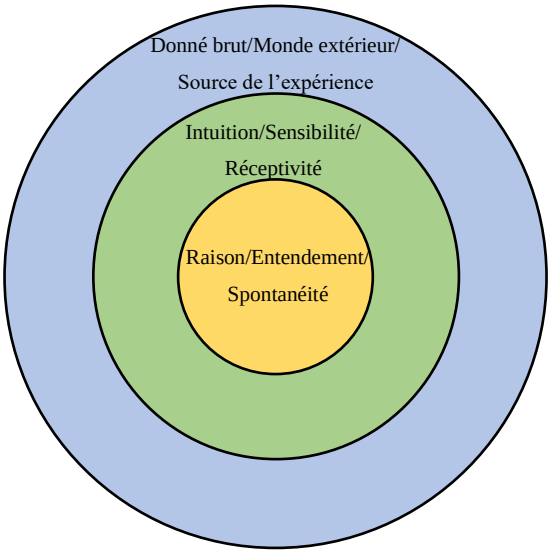
Chapitre 1 : L'esprit selon McDowell

Pour mieux comprendre ce que les IA parviennent à reproduire chez l'esprit humain, penchons-nous d'abord sur la théorie de l'esprit de McDowell ainsi que son cheminement de pensée.

Nous pouvons retrouver cette théorie dans son ouvrage *L'Esprit et le Monde*[1]. Dans celui-ci, qui compile un ensemble de conférences s'intitulant « Les conférences John Locke », prononcées à Oxford en 1991, McDowell s'inspire du système kantien tel qu'il a été développé par Kant dans la *Critique de la Raison Pure* (1781) afin de proposer sa propre conception de l'esprit. Sa position est une position naturaliste mais il s'agit d'un naturalisme qu'il caractérise lui-même de « modéré », se démarquant alors du naturalisme putnamien des années 1960. Sa principale interrogation porte sur le problème de la communication entre le contenu, ce que le monde extérieur présente à nos sens, notre réceptivité, prenant forme dans notre intuition, et nos concepts, qui sont le fruit de notre entendement et qui manipulent ce contenu de l'intuition. En somme, il se demande comment nous pouvons avoir des connaissances empiriques. Mais plutôt que de parler de contenu, qui fait écho à l'idée de contenu représentationnel, McDowell emprunte au penseur Davidson le terme de Donné. Le Donné, c'est le contenu brut, qui vient du monde extérieur via l'affection de notre sensibilité par ce dernier. C'est l'expérience nue. Pour McDowell, Kant ne s'est pas demandé comment le Donné est récupéré par l'entendement, formant ainsi un matériau précieux pour la raison qui manipule des concepts et qui rend des jugements. Nous ne faisons des jugements que via des concepts. Il faut donc penser ce qui relie le concept au Donné et qui nous permet de poser des jugements sur le monde, sur notre environnement. Qui plus est, McDowell considère que les animaux peuvent ressentir des choses comme la douleur ou la peur, à l'instar des humains. L'enjeu pour lui est donc également de concevoir un modèle de l'esprit qui est capable de rendre compte de la proximité qu'il y a entre les humains et les animaux tout en reconnaissant l'existence de capacités uniques présentes uniquement chez les premiers.

Pour mieux comprendre la pensée de McDowell nous pouvons représenter l'esprit et son interaction avec le monde via des cercles concentriques comme suit :

[1] John Henry McDowell, *L'esprit et le monde*, trad. par Christophe Al-Saleh, Analyse et philosophie (Paris: J. Vrin, 2007).



Comme nous pouvons le voir dans la figure ci-dessus, l'esprit peut être schématisé à l'aide de trois cercles concentriques. Il y a le cercle de la raison, là où se forment les différents jugements que l'on a et là où sont manipulés les concepts. McDowell reprend notamment l'idée de Wilfrid Sellars qui repose sur un « espace des raisons »[1] pour définir ce cercle. La raison est incluse dans un autre cercle qui est l'espace des intuitions. C'est ici que se forment les représentations sensibles qui prennent forme dans via les intuitions sensibles. Enfin, il y a le dernier cercle qui est le Donné brut. Nous pouvons parler également d'« extérieur ». Ce sont les informations brutes, telles quelles nous sont fournies à nos organes sensibles. Ce Donné brut devient un Donné une fois que les organes sensibles ont fourni les informations à l'esprit et que l'intuition sensible constitue un objet (qui est ce fameux Donné). Le Donné n'est pas le contenu représentationnel puisqu'il s'agit du contenu présent avant que l'esprit ne le « traite », avant que les concepts ne s'en emparent. Ce schéma va nous permettre d'illustrer et de faciliter la compréhension des différents modèles de l'esprit présentés par McDowell. Tout l'intérêt de l'ouvrage de McDowell est donc de comprendre comment les cercles communiquent ou ne communiquent pas entre eux.

Dans ce schéma nous avons repris beaucoup de termes associés à la pensée de Kant sur l'esprit humain. Et pour cause, comme nous l'avons dit plus tôt, Kant est le point de départ de la réflexion de McDowell.

[1] McDowell, page 37. Référence originale provenant de Wilfrid Sellars, *Empirisme et philosophie de l'esprit*, trad. par Fabien Cayla, Tiré à part (Paris: L'Éclat, 1992), page 80.

1. Le point de départ kantien et le Mythe du Donné

McDowell reprend une citation de Kant qui est la suivante : « Des pensées sans contenu sont vides, des intuitions sans concepts sont aveugles »[1]. Pour lui, cette phrase de Kant constitue le fondement pour un modèle de l'esprit et pour répondre à son questionnement. D'un côté, s'il n'y a pas de contenu, d'éléments que livre l'expérience formée par l'intuition, alors les pensées sont vides. On penserait mais on ne penserait rien puisqu'il n'y a rien à penser. Il n'y aurait pas de contenu représentationnel que les concepts puissent manipuler. Pour McDowell, c'est comme si nos pensées étaient « un jeu de concepts dépourvu de lien avec des intuitions, c'est-à-dire avec des éléments livrés par l'expérience »[2]. De l'autre côté, il y a l'idée que le contenu livré par l'expérience est impossible à connaître s'il ne prend pas forme via des concepts. Ainsi, l'intuition sans concepts est aveugle dans le sens où nous ne pouvons pas connaître (voir en quelque sorte) ce qui se présente à notre sensibilité.

Cependant, pour McDowell, il y a un autre élément dans la philosophie kantienne qui va à l'encontre de ces prescriptions et cela nous mènerait à une forme d'inquiétude. En effet, Kant décrit l'entendement comme l'exercice de la spontanéité. Il y a donc une liberté inhérente à l'espace des raisons. Or, si la liberté est totale, alors la pensée empirique l'est aussi et ne dépend pas de contraintes extérieures à la sphère conceptuelle. Ce faisant nous abandonnons alors l'idée qu'il y ait un fondement à la pensée empirique. La pensée serait certes libre mais « cela menace la possibilité même que les jugements d'expérience puissent être fondés de façon à être mis en relation avec une réalité extérieure à la pensée »[3]. Il faudrait alors séparer le contenu de la pensée et des concepts. Nous tomberions alors dans une forme de solipsisme puisqu'il n'y aurait plus aucun lien entre nous et l'extérieur. En soi, il n'y aurait plus que nous, raison de notre inquiétude.

Il apparaît donc comme nécessaire qu'il y ait une sorte de communication réciproque, une coopération entre l'intuition et les concepts. Cette nécessité est présente car ce que nous dit Kant nous permet d'instituer un dualisme entre les concepts et le contenu qui n'est pas représentationnel, contenu que Davidson nomme le Donné. Mais, s'il y a communication entre les concepts et l'intuition, alors nous pouvons supposer que la liberté absolue de l'entendement n'est peut-être pas

[1] McDowell, page 36. Référence originale provenant de Immanuel Kant, *Critique de la raison pure*, trad. par Alain Renaut, 2e éd. corr, GF 1142 (Paris: Flammarion, 2001), page 144.
[2] McDowell, page 36.
[3] McDowell, page 37.

si absolue que cela et qu'il y a des contraintes qui s'exercent sur les concepts. Or, il faut comprendre comment cette coopération est possible et pour se faire il faut notamment comprendre comment le Donné constitue un fondement dans la justification empirique. Ainsi, selon McDowell, le dualisme concept/intuition permet de

> […] reconnaître que notre liberté de développer nos concepts empiriques est bien contrainte de l'extérieur. Les justifications empiriques dépendent de relations rationnelles, qui sont dans l'espace des raisons. D'après une idée qui est censée nous rassurer, c'est dans les stimulations que le règne conceptuel reçoit de l'extérieur que les justifications empiriques trouvent une fondation ultime.[1]

C'est ce que McDowell nomme « l'idée de Donné »[2]. C'est l'idée selon laquelle le Donné constitue le fondement des jugements empiriques en contraignant les concepts. Il faut donc faire en sorte que cette idée du Donné soit possible et, pour McDowell, la solution la plus simple serait de penser que la sphère des raisons s'étend plus loin que la sphère des concepts de manière à pouvoir recevoir le Donné pur, le Donné indépendamment de toute appropriation conceptuelle. Il n'y aurait alors que des « présences nues »[3] qui seraient le fondement de la pensée empirique, ce sur quoi la raison et les concepts se rapporteraient. Ce serait la méthode la plus simple pour maintenir en place cette dualité concept/intuition. La séparation serait toujours marquée et il y aurait toujours une liberté dans l'entendement, une liberté qui se manifesterait par la spontanéité. Le Donné exercerait une influence sur nos pensées, nous permettant d'avoir un fondement sur lequel elles peuvent se rapporter tout en assurant l'exercice libre de la spontanéité.

Or, pour McDowell cela ne peut pas marcher. Il appelle cette conception le « Mythe du Donné » car, pour lui, cette conception n'apporte pas de justification mais des disculpations. Tout d'abord, il considère qu'il n'est pas clair de concevoir des relations extra-conceptuelles. Nous raisonnons via les concepts et supposer que la raison puisse recevoir du Donné pur serait donc parfaitement mystérieux. Cette relation serait impossible à conceptualiser et donc à connaître. Qui plus est, cela voudrait dire que la raison est elle-même soumise aux contraintes du monde extérieur. Elle y serait sensible du moins. La spontanéité n'aurait aucun contrôle dessus et ces contraintes extérieures ne seraient pas rationnelles. Cela ne serait qu'une pure relation causale, irrationnelle par nature et cela ne constitue en rien une manière convaincante pour que le contenu extra-

[1] McDowell, page 38.
[2] McDowell, page 38.
[3] McDowell, page 57.

conceptuel puisse s'intégrer de manière pertinente dans nos jugements empiriques. Ainsi, pour McDowell :

> […] une chose est d'être à l'abri du blâme, pour la raison que la situation dans laquelle nous nous trouvons découle en dernière instance de la force brute ; une autre est d'être en possession d'une justification. En définitive, l'idée de Donné fournit des disculpations là où nous attendions des justifications.[1]

Nous pourrions alors être tentés de rejeter le Mythe du Donné en considérant que le Donné n'affecte en rien les jugements empiriques et que la spontanéité agit librement sans aucun fondement de la sorte. Mais cela nous ramènerait à l'inquiétude qui est à l'origine du Mythe du Donné. En effet car une pensée sans contenu est vide ; il faut donc un fondement, un contenu. Pour échapper à cette oscillation entre Mythe du Donné et rejet du Donné (l'inquiétude), McDowell propose d'étudier d'autres modèles de l'esprit qui ont été théorisés.

2. Davidson et Evans, les deux choix possibles

Une des solutions pour sortir du Mythe du Donné est, selon McDowell, fournit par Davidson. Ce dernier défend une « théorie cohérentiste de la vérité et de la connaissance »[2]. Pour Davidson, le Donné qui s'exprime dans notre sensibilité n'est pas récupéré par les concepts. La sphère des concepts dans laquelle la spontanéité s'exerce est donc parfaitement délimitée, comme quand nous parlions de la dualité concept/intuition. Notre raison ne va pas au-delà de cette sphère non plus et c'est là la différence majeure avec le Mythe du Donné. Mais alors comment pouvons-nous avoir des jugements empiriques, faire en sorte que le contenu affecte notre pensée de manière à ce qu'elle puisse porter sur le monde ? Pour Davidson, c'est parce que la sensibilité affecte les concepts sur la pensée empirique sans qu'il n'y ait vraiment de contraintes rationnelles externes. Contrairement au Mythe du Donné, où l'espace des raisons dépassait celui des concepts et, ce faisant, était directement affecté par les contraintes causales venant de l'extérieur, la conception de Davidson suppose une affection des concepts sans que de contraintes s'appliquent à ces derniers. Par analogie, ce serait comme si l'on demandait à une personne d'avancer plutôt que de la pousser. La

[1] McDowell, page 40.
[2] McDowell, page 48.

personne reste entièrement libre de prendre en compte ce conseil ou non. Ici, pour les jugements et surtout les jugements empiriques, l'extérieur « conseille » plutôt que « contraint » les concepts.

Avoir une telle conception est lourde de conséquences car, si l'on suit la pensée de Davidson, nous ne pouvons pas sortir de nos croyances, nous ne pouvons donc pas avoir d'interactions avec le monde extérieur. Mais alors comment faire pour que les pensées portent sur le monde, pour que notre esprit puisse avoir une liaison rationnelle avec l'extérieur ? L'inquiétude est de retour et, pour y remédier, Davidson postule qu'il n'y a certes que des croyances mais que celles-ci sont vraies par essence. Pour cela, d'après McDowell, Davidson

> […] défend cette thèse en reliant la croyance à l'interprétation, et en soulignant qu'il est propre à l'interprétation qu'un interprète trouve que ses sujets ont pour l'essentiel des croyances correctes sur le monde avec lequel il les observe interagir causalement.[1]

La croyance serait donc en quelque sorte une interprétation des signaux causaux qui se présentent à la raison. Davidson semble donc apporter une solution pour que son système puisse expliquer l'existence de jugements empiriques qui ont une vraie emprise sur le monde, qui ne s'appuient pas sur rien (car, il ne faut pas l'oublier, « des pensées sans contenu sont vides »).

Néanmoins, il reste qu'on ne peut pas sortir de nos croyances, nous ne pouvons donc pas avoir d'interactions rationnelles avec le monde extérieur via notre sensibilité. Dire cela enlève le Donné du problème, effaçant ainsi l'inquiétude que nous avions au sujet de la communication entre les concepts et l'intuition. Or, pour McDowell, la question persiste. En effet, il demande « comment est-il possible que des exercices de la spontanéité portent sur une réalité qui est complètement extérieure à la sphère de la pensée ? »[2]. McDowell n'est pas d'accord avec le lien que Davidson fait entre la croyance et l'interprétation. Il s'agit selon lui d'une disculpation et non d'une justification. Davidson se coupe du monde, voulant ainsi éviter de tomber dans le Mythe du Donné mais cela pose un problème pour sa conception de nos croyances car, étant donné qu'il estime que l'expérience à un impact causal pour les jugements et croyances, l'expérience n'apporte pas de justifications car la causalité dont l'expérience dépend pour affecter la raison n'est pas rationnelle. Pour McDowell, défendre cette thèse n'est pas possible. Sans une contrainte rationnelle, il n'est pas possible que l'extérieur, l'environnement, puisse justifier nos pensées ou les adapter à ces justifications. Si l'on suit l'hypothèse de Davidson, nous pourrions à peu près croire n'importe quoi

[1] McDowell, page 49.
[2] McDowell, page 48.

et considérer cela pour vrai car aucune de nos croyances ne peut être confrontée à ce que Quine nomme le « tribunal de l'expérience ».

D'après McDowell, refuser la thèse de Davidson nous incite à retomber dans le Mythe du Donné, créant ainsi une oscillation dont il faudrait sortir. Il faut donc trouver une autre alternative à ces deux modèles de l'esprit et il propose alors de se pencher sur la pensée de Gareth Evans. Selon ce dernier :

> Les états informationnels acquis par un sujet dans la perception sont *non-conceptuels*, ou *non-conceptualisés*. Les jugements fondés sur ces états impliquent nécessairement une conceptualisation : en passant d'une expérience perceptive à un jugement sur le monde (typiquement exprimable dans une certaine forme verbale), on exerce des aptitudes conceptuelles fondamentales... Le processus de la conceptualisation ou du jugement prend le sujet dans un certain type d'état informationnel (avec un contenu d'un certain genre, c'est-à-dire un contenu non-conceptuel), pour l'amener dans un autre type | d'état cognitif (avec un contenu d'un autre genre, c'est-à-dire un contenu conceptuel).[1]

Donc, pour Evans, il y a un contenu qui se forme dans la sensibilité mais celui-ci est non conceptuel. Le contenu devient conceptuel que quand il y a un jugement. Le jugement a donc le statut de transition d'un contenu non-conceptuel à un contenu conceptuel. Evans se distingue des autres positions que l'on a vues juste avant car il sépare le conceptuel de l'intuition tout en considérant que le « système informationnel » permet de produire des états porteurs de contenu sans que la spontanéité intervienne dans l'opération. Ce sont des contenus non conceptuels. McDowell nous explique que ce qu'Evans nomme « système informationnel » est :

> […] le système des capacités exercées lors de la collection d'information sur le monde par le biais de nos sens (dans la perception), en recevant cette information des autres dans la communication (par le témoignage), et en conservant cette information dans le temps (avec la mémoire).[2]

Cette conception de l'esprit d'Evans a pour conséquence d'expliquer ce que nous avons en commun et ce qui nous différencie des animaux. Nous avons en commun des états informationnels mais les animaux, eux, n'ont pas de spontanéité. Ainsi, contrairement à nous, les animaux n'ont pas d'expériences perceptives car les expériences perceptives sont les états d'un sujet, d'un sujet exerçant sa spontanéité. Les animaux n'ont pas la capacité de saisir ce contenu non-conceptuel via des concepts et de former rationnellement un jugement d'expérience. Les éléments perceptifs d'une créature n'ayant pas de faculté de spontanéité ne constituent pas une expérience perceptive. Les

[1] McDowell, page 80-81. Référence originale provenant de Gareth Evans et al., *The Varieties of Reference* (Oxford, New York: Oxford University Press, 1982), page 227.
[2] McDowell, page 81.

états informationnels se forment dans l'intuition sans qu'un quelconque concept n'intervienne. Ils précèdent la spontanéité. Cependant, il faut noter que ces mêmes éléments perceptifs peuvent devenir une expérience perceptive chez les créatures qui, elles, ont une spontanéité.

Ainsi, la conception d'Evans se distingue du Mythe du Donné et de la conception de Davidson. Ces deux dernières conceptions donnent un rôle primordial à la spontanéité et à la raison. Cependant, si l'on s'accorde pour dire que les animaux n'ont pas de spontanéité, alors nous sommes obligés de dire qu'ils sont exclus de toute forme d'expériences, qu'elles soient « internes » comme le fait de ressentir de la douleur ou de voir une couleur, ou « externes » comme le fait de voir comment sont les choses, dans leurs formes ou leurs dispositions dans l'espace et le temps. Que ce soit pour le Mythe du Donné ou la conception de Davidson, la spontanéité est indispensable pour émettre un jugement donc pour percevoir le monde, ressentir de la douleur, etc. Or, force est de constater que les différentes créatures semblent être sensibles à leur environnement et Evans, lui, considère que les animaux ont quelque chose en commun avec nous, ce qui pourrait expliquer cette sensibilité à leur environnement.

La raison pour laquelle Evans situe les expériences en dehors de la sphère conceptuelle est qu'il y a énormément de détails dans le contenu de l'expérience. Et, selon lui, les concepts du sujet ne peuvent pas saisir l'intégralité de ces détails. Il y a donc des approximations qui sont faites comme par exemple lorsqu'il s'agit de discriminer les nuances de couleurs données dans l'expérience. De plus, il fait cela pour maintenir la liberté inhérente à la spontanéité. Si le contenu prend forme sans l'aide des concepts, alors la spontanéité ne subit pas de contraintes externes irrationnelles. Mais alors, comment les contenus non conceptuels peuvent fournir un fondement pour les jugements ? N'oublions pas qu'Evans isole la spontanéité comme la fait Davidson par exemple mais contrairement à ce dernier, où la spontanéité est affectée de manière non rationnelle, Evans postule qu'il y a un processus rationnel, transformant les expériences en raisons sur lesquelles les jugements peuvent s'appuyer. Plus précisément, il y a un effet de translation de l'expérience depuis l'espace de l'intuition à l'espace des concepts. Pour Evans, les expériences sont des dispositions à faire des jugements. Ces dispositions ne se réalisent dans les jugements que si la formule « toutes choses égales par ailleurs » est satisfaite, c'est-à-dire que si les croyances mobilisées dans le jugement s'associent à l'expérience, cette expérience permettant de rendre compte de la réalité du monde extérieure et incitant à se focaliser sur un élément de celle-ci. Ainsi, l'expérience se réalise dans mon jugement parce qu'elle m'expose à un objet ou un élément de

l'environnement et permet de focaliser mon jugement sur lui. L'expérience « incite » mon jugement. Ainsi, lorsqu'une disposition se réalise dans mon jugement, je peux dire « toutes choses égales par ailleurs » car l'expérience m'« incite » à focaliser mon jugement sur elle et par rapport à elle en ignorant tout le reste. Mon jugement est influencé par l'expérience, il se base sur celle-ci. Ma vision d'un arbre se réalise lorsque je dis « cet arbre est vert » car ce jugement était rendu disponible par cette expérience visuelle. L'expérience focalise l'esprit sur la vision de l'arbre, ce qui amène alors à dire que « toutes choses égales par ailleurs, cet arbre est vert ». Nous laissons de côté tous les autres paramètres pour ne juger que cet arbre. Cela permet à Evans de différencier l'expérience de la croyance. La croyance est mobilisée dans les jugements mais elle ne satisfait pas forcément la formule « toutes choses égales par ailleurs ». Elle ne rend pas compte d'un état du monde extérieur et n'incite donc pas à porter notre jugement sur un élément particulier de celui-ci. Par exemple, quand nous sommes victimes d'une illusion, notre croyance ne s'associe pas à une expérience, elle ne rend pas compte du monde extérieur. L'expression « toutes choses égales par ailleurs » n'est pas satisfaite car il n'y a pas de focus, d'incitation par une expérience à effectuer tel ou tel jugement. Nous ne pouvons pas étudier un seul élément du monde extérieur tout en ignorant les autres paramètres qui pourraient être pris en compte. Sans disposition qui se réalise, sans expérience, le jugement perd un rapport au monde que la croyance seule ne peut pas apporter. Il y a donc une indépendance entre l'expérience avec la croyance. Nous savons que l'expérience exerce une influence sur notre jugement lorsque l'expression « toutes choses égales par ailleurs » est satisfaite. Sinon, il n'y a que des croyances qui sont mobilisées. Nous retrouvons quelque chose de similaire avec Davidson pour qui les croyances interprètent les signaux non-rationnels provenant de l'extérieur. Ici, les contenus non-conceptuels issus de l'expérience « incitent » à faire tel ou tel jugement. Il y a une similarité dans ces deux approches.

Cependant la conception d'Evans pose un problème pour McDowell. Pour lui, elle n'est pas satisfaisante. D'abord parce que la phrase de Kant « des pensées sans contenu sont vides » réfute cette conception. La pensée, les jugements que l'on porte, n'ont pas de contenu. Il y a juste un contenu qui « incite » le jugement mais qui ne l'« exige » pas. Il n'y a rien sur quoi reposent fondamentalement les jugements empiriques. De plus, pour Evans, il y a bien du contenu issu des expériences formé par les intuitions. Mais comme l'a dit Kant : « des intuitions sans concepts sont aveugles ». Les intuitions sont donc aveugles selon Evans et dire cela, c'est dire qu'on n'a pas d'aperçu phénoménal du monde. Evans lui-même est d'accord avec cela, avec le fait que le sujet

doit avoir une conscience du lien entre sa perception et la réalité pour comprendre que le monde qu'il voit est le vrai monde. Or, pour McDowell :

> Un sujet ne dispose de cette compréhension d'arrière-plan que s'il est conscient pour lui-même de la manière dont son expérience se rapporte au monde, et nous ne pouvons pas nous représenter cette condition sans faire intervenir des capacités conceptuelles dans le sens le plus fort, sans faire intervenir, donc, la faculté de spontanéité.[1]

Pour McDowell, les concepts sont indispensables pour le contenu qui se forme dans la sensibilité. Il estime aussi que ce qu'il y a en commun entre les humains et les animaux n'est pas très clair. Il faudrait que nous soyons capables de trier et de mettre de côté toutes nos spécificités afin de découvrir un noyau commun avec les animaux. Mais cela semble impossible. La thèse selon laquelle nous tous, les êtres vivants, auraient un noyau non-conceptuel commun mais que nous, les humains, aurions une spécificité ne semble pas pertinente chez Evans aux yeux de McDowell. Pour ce qui est de l'argument selon lequel on ne peut pas saisir conceptuellement l'entièreté des détails que fournit l'expérience, McDowell estime que l'on peut former des concepts précis, comme les nuances de couleurs et que quand bien même nous n'avons pas autant de concepts de couleurs qu'il y a de nuances de couleurs, nous avons le concept de nuance lui-même qui nous permet de discriminer les couleurs entre elles et de les observer dans leurs détails. D'autant plus que, pour lui, la capacité de recognition est conceptuelle donc le contenu ne peut pas être non-conceptuel. Enfin, selon McDowell, Evans ne s'inquiète pas que le lien qu'il y a entre la pensée et l'expérience soit mystérieux. Il n'estime pas que c'est une contrainte causale et, ce faisant, il n'entre pas dans le Mythe du Donné. Cependant, il isole l'esprit du monde extérieur en isolant la spontanéité. La pensée ne peut alors pas se remettre en question, faire l'examen de soi, réviser ses jugements.

Tout comme la pensée de Davidson, l'esprit est isolé afin d'assurer la liberté de la spontanéité chez Evans. Ces positions ne sont pas tenables pour McDowell. Ainsi, s'il ne trouve pas de conception qui permette d'expliquer la manière dont les jugements empiriques s'effectuent tout en respectant les conditions kantiennes comme le besoin de contenu de la pensée, le besoin de concepts dans l'intuition et la liberté de la spontanéité, alors il faut qu'il formule la sienne. Et c'est ce qu'il va faire. Sa conception se forme par rapport aux autres que nous venons d'étudier et permet notamment de rendre compte de choses comme, par exemple, le rapport entre les humains et les animaux ainsi que le rapport qu'ont les humains avec le monde.

[1] McDowell, page 87.

3. La conception de l'esprit de McDowell

Contrairement au Mythe du Donné, à Davidson et à Evans, McDowell estime qu'il est indispensable qu'il y ait de la spontanéité dans l'intuition. Il reste fidèle à la citation de Kant « des pensées sans contenu sont vides, des intuitions sans concepts sont aveugles » et est même très intransigeant quant à son respect scrupuleux. Pour lui, les jugements que l'on porte sur le monde se font à travers la manipulation des concepts et des contenus. Or, ces contenus doivent être rationnels pour être manipulés, pour qu'ils affectent et exercent une contrainte sur les jugements. Cette contrainte ne serait pas le fruit d'une force inconnue, irrationnelle, comme la causalité ; elle est issue d'une relation rationnelle avec ce que fournit l'expérience. Par exemple, un contenu empirique (comme la représentation d'un arbre avec des feuilles vertes) fait l'objet d'une prise de conscience car il y a un lien rationnel qui est présent entre ce contenu et la raison et, ce faisant, il permet de contraindre les jugements de telle manière que la raison est amenée à considérer que la proposition « l'arbre a des feuilles bleues » est fausse, formulant alors une nouvelle proposition, une correction. L'expérience stimule rationnellement la spontanéité, elle crée des frottements pour que cette dernière ne soit pas hors sol, pour qu'elle se rapporte au monde. S'il y a contrainte et non incitation ou interprétation et si cette contrainte est rationnelle, alors il faut qu'il y ait des concepts qui interviennent lors de la formation du contenu afin que ce contenu soit un contenu conceptuel. Ainsi, « quand il m'apparaît que les choses sont d'une certaine façon, les capacités conceptuelles ont déjà été mises en œuvre »[1].

Un problème se pose cependant. L'exercice de la spontanéité est actif et libre alors que le sujet est passif quand il reçoit l'expérience. Pour prouver que les capacités mobilisées dans l'intuition sont conceptuelles il faut donc, selon McDowell, montrer que

> […] ces capacités servent également dans l'activité de juger. Et on peut valider l'identification entre les capacités à l'œuvre dans les apparences et les capacités à l'œuvre dans les jugements, en montrant un lien rationnel entre les apparences et la spontanéité, en montrant que les jugements sur la réalité objective peuvent bien trouver dans ces apparences des raisons, et que ces apparences, dans les circonstances appropriées (« toutes choses égales par ailleurs »), constituent des raisons à ces jugements.[2]

[1] McDowell, page 95.
[2] McDowell, page 95.

Et McDowell montre que c'est effectivement le cas. Comme nous l'avons vu avec l'exemple de l'arbre, le contenu formé dans la sensibilité fournit des raisons aux jugements. Ce sont des capacités qui sont à l'œuvre dans l'activité de juger.

Supposer que des capacités conceptuelles sont mobilisées dans l'expérience permet à McDowell de résoudre le problème de la liberté de la spontanéité. En effet, ce sont les concepts qui sont présents dans notre réceptivité qui subissent les actions de l'extérieur. Le contenu étant d'emblée conceptuel impose une contrainte à la spontanéité sans pour autant que celle-ci soit entravée. La spontanéité se fait librement, activement, en manipulant du contenu conceptuel dont une partie, au moins, tire son origine de l'expérience. On ne tombe ni dans une forme de solipsisme, ni dans une conception où la raison est directement soumise aux contraintes causales extérieures.

Le fait que la spontanéité se diffusant dans la sensibilité ne parait pas aller de soi vient pour McDowell du fait que nous nous distinguons des animaux via la possession de ladite spontanéité. Il y a, par exemple, chez Evans l'idée que nous avons un socle commun avec les animaux mais que nous avons, nous, quelque chose en plus, à savoir la spontanéité. Ainsi, nous partagerions seulement avec les animaux « une sensibilité perceptive aux traits de l'environnement »[1]. Or, pour McDowell, cette séparation n'est pas obligatoire. Au contraire, il faut même considérer que celle-ci n'existe pas vraiment pour éviter de tomber dans la situation d'Evans et Davidson.

Cette séparation découle de l'avènement d'une vision particulière du monde, apparue lors de l'essor de la science moderne. Cet essor a mené à la confrontation entre deux types d'intelligibilité :

> […] le premier type est cherché par ce que nous appelons la science naturelle, le second type est celui que nous trouvons quand nous situons quelque chose par rapport aux autres occupants de « l'espace logique des raisons », pour reprendre une expression suggestive de Wilfrid Sellars.[2]

Par le biais des sciences, nous pouvons étudier la nature, décrire les lois qui régissent le monde, et nous pouvons alors être tentés de décrire à travers ces mêmes sciences « l'espace logique des raisons » afin de trouver une forme d'intelligibilité chez les animaux et afin de réduire la nôtre en des termes naturels, causaux, non-rationnels. C'est en soi le but du projet naturaliste qui s'est développé à partir de la première moitié du XX^{ème} siècle. On retrouve notamment la démarche physicaliste qui est une théorie postulant que l'esprit n'a pas de substance propre. Tout ne serait que matière et états physiques. Il n'y aurait qu'une substance matérielle et donc les différents états

[1] McDowell, page 102.
[2] McDowell, page 104.

mentaux, c'est-à-dire les différents états dans lequel notre esprit peut se trouver, comme formuler un jugement, sont identiques à certaines combinaisons d'états physiques. Ainsi, formuler un jugement, c'est avoir une combinaison précise de neurones activée.

Or, pour McDowell, en faisant cela on vide la nature de signification car les lois naturelles ne sont que des lois irrationnelles. La nature se retrouve alors, selon lui, désenchantée dans le sens où celle-ci serait vidée de signification. Expliquer l'esprit avec les sciences aurait bien évidemment la même conséquence. À cela s'oppose ce que McDowell considère comme « l'espace logique des raisons ». Dans cette espace, nous créons le contenu rationnel, nous le comprenons, nous donnons des raisons, des justifications aux choses. Ce sont les jugements qui se basent sur notre spontanéité, la logique et les mathématiques. Pour McDowell, ces deux types d'intelligibilité n'ont pas, de prime abord, la même localisation dans le monde. Les sciences naturelles appartiennent à la nature, au monde extérieur et matériel. L'espace logique des raisons, lui, est présent dans notre esprit. Il s'agit de notre entendement, notre raison. Cependant, il s'avère être en réalité impossible de séparer les deux formes d'intelligibilité. Notre esprit se situe dans un corps qui est soumis aux lois de la nature. De plus, notre sensibilité et notre entendement sont censés être des capacités qui doivent être un minimum communes à nous et aux animaux. Il y a donc une incohérence qui apparaît car les opérations présentes dans l'entendement et, d'après McDowell, dans la sensibilité sont produites via des concepts mais cela devrait en réalité aussi être des opérations produites par les lois de la nature. Finalement, il faut parvenir à déterminer le statut de la spontanéité.

McDowell distingue trois statuts possibles. Le premier serait ce qu'il nomme le « naturalisme brut ». Selon cette conception, il faut que les capacités conceptuelles puissent être décrites en des termes naturels afin d'ancrer et de localiser les choses dans la nature. Le naturalisme brut ne peut pas expliquer des éléments comme la raison ou la justification mais il peut décrire et réduire l'espace des raisons, qui est ce qui permet l'existence de la raison et de la justification, en des termes scientifiques et naturels. Il ne s'agit pas de nier l'espace des raisons mais de dire que ce qu'il produit sert aussi à situer les choses dans la nature. Cela revient alors à considérer que la nature n'est pas totalement désenchantée car « selon cette conception, la naturalité n'exclut pas l'intelligibilité propre à la signification ».[1]

Une autre possibilité est celle que McDowell attribue à Davidson. Pour ce dernier, les concepts doivent se soumettre à l'intelligibilité propre à la signification et doivent donc rester dans l'espace

[1] McDowell, page 107.

des raisons. On ne peut pas réduire les concepts dans l'espace logique des lois et faire que leur but, leur fonction, est de situer les choses dans le règne de la loi. Mais, si l'on suit la pensée de Davidson, alors les choses-mêmes qui sont manipulées par la spontanéité via des concepts seraient, quant à elles, disponibles pour être décrites par les sciences afin d'être situées dans le règne de la loi. On pourrait donc décrire dans des termes scientifiques, naturels, les événements appartenant à l'espace des raisons. Les éléments de l'espace des raisons auraient donc des relations causales entre eux et avec d'autres éléments car les relations causales n'ont lieu qu'entre éléments qui sont dans le règne de la loi naturelle. Ainsi, selon Davidson (d'après McDowell), « une raison peut être une cause, même si ce n'est pas en vertu de son réseau de relations rationnelles qu'elle se tient dans des relations causales »[1].

Mais McDowell fait remarquer que cette possibilité implique l'idée que la sensibilité est naturelle et que la spontanéité, elle, est dans l'espace des raisons. Ce faisant, cela exclut la possibilité que la spontanéité puisse se diffuser dans les opérations de la sensibilité et, pour McDowell, on ne peut pas non plus dire que quelque chose instancie des concepts en vertu du fait que cette chose se trouve dans le règne de la loi. Ce n'est pas sa place dans le monde des lois naturelles qui fait qu'elle provoque l'apparition de tel ou tel concept. Donc pour McDowell, « il nous est interdit de soutenir que l'expérience tient précisément son contenu conceptuel du fait d'être un phénomène naturel »[2]. Or, les impressions sont des phénomènes naturels qui nous servent en tant que contenu conceptuel. Mais, « ce n'est pas parce qu'elles sont des phénomènes naturels qu'on peut caractériser les impressions en termes de spontanéité »[3]. Si on ne peut les caractériser en termes de spontanéité alors les impressions restent que des intuitions qui sont dans le règne de la loi et qui sont donc désenchantées. Ce faisant, les impressions sont forcément indépendantes de la spontanéité et cela est, pour McDowell, inacceptable comme nous l'avons vu auparavant.

Ainsi, plutôt que de parler de trois statuts possibles, il n'y en a que deux en réalité selon McDowell. Il y a donc soit le naturalisme brut, soit celui qu'il propose et qu'il nomme le « naturalisme modéré ». Pour lui, on ne peut pas réduire l'espace logique des raisons dans l'espace des lois naturelles mais nous retrouvons tout de même des capacités conceptuelles mobilisées dans la sensibilité. Cela va donc à l'encontre de l'idée qu'être naturel, c'est occuper une position

[1] McDowell, page 108.
[2] McDowell, page 109.
[3] McDowell, page 109.

uniquement dans le règne de la loi. On laisse ainsi une place à la spontanéité dans la nature car McDowell considère que ces événements sont des « actualisations de notre nature ».

Cependant, si on s'oppose au naturalisme brut, alors l'on risque de tomber dans ce que McDowell nomme « un platonisme rampant » au sein duquel nous nous considérerions comme faisant partie de la nature tout en ayant une part de nous, notre esprit, qui serait en dehors de celle-ci, dans une autre réalité et comme étant autonome du reste. McDowell, dans sa conception de l'esprit, évite de tomber dans cet écueil. Pour lui, rien de surnaturel car « les exercices de la spontanéité relèvent de notre manière de nous actualiser en tant qu'animaux »[1]. Même si la spontanéité et l'espace logique des raisons ne peuvent pas être réduits dans l'espace du règne des lois, les fondements du premier peuvent s'expliquer à partir du second. Pour ce qui est de l'espace logique des raisons en lui-même, McDowell considère qu'il

> […] n'est pas nécessaire d'intégrer les concepts relatifs à la spontanéité dans la structure du règne de la loi ; nous devons au contraire montrer qu'ils permettent d'intégrer des schémas dans un mode de vie.[2]

Pour mieux comprendre cet espace logique des raisons, il s'appuie alors sur la pensée d'Aristote, notamment l'éthique, et donne son interprétation de ce qu'est la « sagesse pratique » qu'il considère comme étant « une réactivité à certaines exigences de la raison »[3]. Ainsi, il y aurait des exigences que seraient présentes sans que nous en ayons conscience et c'est via la « sagesse pratique » que la prise de conscience s'opère. Celle-ci permet alors de « reconnaître et de créer le type d'intelligibilité permettant de placer quelque chose dans l'espace des raisons »[4]. Elle est le fruit d'une activité autoréflexive sur notre pensée afin que les règles la gouvernant soient constamment remises en question et affinées. McDowell, met en avant le rôle de l'éducation pour apprendre à penser et nous permettre de prendre conscience de l'espace logique des raisons. La « sagesse pratique » donne lieu à ce que McDowell appelle une seconde nature de ceux qui la possède. On peut donc expliquer l'espace des raisons via l'éducation qui apprend aux individus des règles de la pensée. Les habitudes qui en résultent sont une seconde nature et les moyens utilisés pour modeler les individus à ce mode de pensée sont ce que McDowell associe au terme allemand

[1] McDowell, page 112.
[2] McDowell, page 112.
[3] McDowell, page 112.
[4] McDowell, page 112.

« Bildung » que l'on peut traduire par « éducation » en français. Le Bildung accorde une autonomie de la signification.

A travers cette conception, McDowell réenchante la nature sans tomber dans un platonisme rampant ni dans une conception du monde préscientifique, où l'on considérait que la raison se trouvait partout dans la nature. Nous pouvons donc librement supposer que les opérations conceptuelles sont présentes dans la sensibilité sans que cela ne pose de problème. Ce naturalisme modéré peut aussi être considéré comme un « platonisme naturalisé »[1]. L'espace des raisons ne peut pas être décrit par le biais des lois naturelles mais il fait partie de la nature tout de même. Ce faisant, l'espace des raisons n'est pas isolé du reste du corps humain.

De plus, sa conception permet de corriger des manquements à la philosophie kantienne. En effet, Kant considère que le « je » consiste dans le fait que quelque chose persiste à travers le temps et qui relie les représentations. Mais cette persistance est seulement formelle et non substantielle, contrairement à l'ego cartésien. Or, pour McDowell, ce que dit Kant ne suffit pas. Cette persistance à travers les représentations n'est valable que si l'on suppose une chose vivante dans laquelle les événements ont lieu. Il faut donc une seconde nature car le concept de vie est le phénomène qui permet de créer une unité dans la spontanéité, un « je » qui est naturel. Ce concept de vie, présent dans l'espace des raisons, permet de trouver la signification de nous-mêmes comme étant une présence corporelle dans le monde. L'idée de seconde nature présuppose donc que l'on existe substantiellement dans le monde et qu'on agit dans celui-ci. Ainsi, « on peut concevoir les exercices des capacités relevant de la spontanéité comme des éléments dans le cours d'une vie »[2].

Forts de cette conception, nous pouvons maintenant nous demander ce qu'il en est concernant les animaux. Pour McDowell, c'est l'entendement et la spontanéité qui permet de prendre conscience du monde et de soi. Or, les animaux, eux, en semblent dépourvus. Pour mieux décrire comment pourrait être l'esprit des animaux, McDowell reprend les travaux de Gadamer, notamment son ouvrage *Vérité et méthode* (1960), défendant l'idée selon laquelle les animaux n'ont que des besoins immédiats et n'agissent que par rapport à leurs besoins biologiques. Pour Gadamer, ils ne vivent pas dans le monde, mais dans un environnement où seuls des problèmes et des besoins biologiques sont présents. Contrairement à eux, nous ne sommes pas seulement soumis à des impératifs biologiques et nous pouvons nous penser dans un monde. Cependant, il ne faut pas dire

[1] McDowell, page 126.
[2] McDowell, page 148.

que les animaux sont de simples automates, des créatures qui ne sont pas sensibles. En effet, pour McDowell, les animaux ont une sensibilité proto-subjective, c'est-à-dire une forme de sensibilité dépourvue de l'entendement humain. Dans un même temps, il reconnaît que les capacités conceptuelles intéressantes concernant les émotions, les états émotionnels et autres, « sont à l'œuvre ici que parce que leur mise en œuvre intègre le fait de comprendre qu'elles ne se réduisent pas à la première personne et au présent de l'indicatif »[1]. Cela laisse donc entendre que les animaux, créatures sensibles possédant une forme de proto-subjectivité (mais n'ayant pas de conscience à proprement parler) peuvent mobiliser des capacités conceptuelles relatives aux émotions et autres états émotionnels. Ils peuvent ainsi avoir peur ou avoir mal comme nous pouvons le constater au quotidien. C'est d'ailleurs ce que McDowell confirme en disant que :

> […] rien dans les concepts de douleur ou de peur n'implique qu'ils ne peuvent accrocher que là où il y a de l'entendement, et, donc, | une subjectivité pleine et entière. Il n'y a aucune raison de supposer qu'ils ne peuvent être appliqués selon un mode autre que la première personne qu'à quelque chose qui soit capable de se les appliquer à la première personne.[2]

Nous avons donc des caractéristiques de notre esprit en commun avec les animaux, comme la sensibilité ainsi que des états émotionnels, des émotions mobilisées par des concepts qui ne dépendent pas uniquement d'une subjectivité pleine, d'une spontanéité. De plus, pour McDowell, l'acquisition de notre seconde nature nous permettant d'avoir un entendement et une spontanéité peut être expliquée avec la théorie évolutionniste. Cela nous ancre alors d'autant plus dans le monde animal. Une piste pour expliquer cette acquisition est notamment l'apparition du langage qui permet de figer et transmettre les savoirs. Nous pouvons créer des traditions qui se transmettent et qui sont remises en question ou modifiées par les générations suivantes. En somme, c'est via le langage que le Bildung a pu apparaître et se développer au sein des communautés humaines.

Ainsi, l'être humain, comme les animaux, est présent dans le monde et vit dedans. Nous partageons beaucoup de choses avec les animaux (puisque nous en sommes nous-mêmes) mais, pour McDowell, ce qui nous différencie d'eux n'est pas le fruit de quelque chose d'extérieur à la réalité matérielle. Il n'y a rien d'extraordinaire en nous, juste des capacités acquises à travers l'évolution, une seconde nature qui nous permet de saisir les données de la sensibilité via des concepts afin de se les approprier et de former des jugements empiriques qui se rapportent au monde. L'extérieur a donc un impact sur nous, un impact rationnel. Il provoque des frottements

[1] McDowell, page 157.
[2] McDowell, page 158.

contre l'activité libre de notre spontanéité, l'exhortant ainsi de réviser ses jugements afin qui ceux-ci s'accordent à la réalité des choses extérieures.

4. L'Intelligence Artificielle et la conception de l'esprit de McDowell

Comme nous l'avons vu plus tôt, le programme de recherche sur l'Intelligence Artificielle s'est développé dans un contexte bien particulier. Il est apparu à une époque où le naturalisme philosophique est revenu sur le devant de la scène, conjointement à une avancée spectaculaire des sciences naturelles. Ces dernières étant très prolifiques, il apparut alors que cette pratique et méthode scientifique furent les plus pertinentes pour comprendre le monde et nous-mêmes. Avec de nombreux penseurs comme Carnap, Quine ou encore Russel, ainsi que le développement de la philosophie analytique qui prône une philosophie plus proche des sciences, la volonté (car l'on pensait que c'était possible) de naturaliser l'esprit humain, c'est-à-dire de pouvoir décrire son fonctionnement et ce dont il s'agit en des termes scientifiques et fonctionnels est apparue. L'idée que le cerveau humain puisse être lui aussi une machine de Turing, provoqua l'émergence du fonctionnalisme computationnel avec Hilary Putnam. Le fonctionnalisme computationnel est plus subtil que le physicalisme que nous avons abordé plus tôt. Il prétend décrire les états mentaux, les différents états dans lesquels notre esprit peut se trouver, via les fonctions que ces états jouent au sein du système pris dans son ensemble, c'est-à-dire notre corps. Le fonctionnalisme cherche donc à décrire l'esprit humain en des termes scientifiques et fonctionnels, donc toujours de manière causale et naturelle. L'Intelligence Artificielle est apparue dans ce cadre philosophique là et les deux principaux types d'IA, c'est-à-dire l'IA symbolique et l'IA connexionniste, qui se sont développés au fil des décennies ont pour objectif (ou du moins ils l'avaient au départ) de reproduire l'esprit et l'intelligence humaine.

Or, ici, nous avons un modèle de l'esprit différent du naturalisme brut d'alors. Ici, il s'agit d'un naturalisme modéré qui affirme l'indépendance et la naturalité des capacités rationnelles. Comme nous l'avons dit au début de ce chapitre, ce modèle de l'esprit de McDowell est une des raisons du rejet du fonctionnalisme computationnel par son propre créateur, Hilary Putnam, et même de la conception réaliste sous-entendue. Si Putnam a remis en question sa vision de l'esprit humain ainsi que sa manière de voir la réalité, il convient alors de se demander si l'IA est toujours cohérente dans le modèle mcdowellien, dans cette vision particulière de l'esprit et du monde.

Pour que l'IA soit cohérente dans ce modèle, il faudrait qu'elle soit capable de reproduire la sensibilité ainsi que l'intuition humaine et puisse rendre compte des différentes capacités conceptuelles mobilisées aussi bien dans l'intuition sensible que dans la spontanéité. En somme, elle doit être capable de former du contenu par le biais des informations reçues et faire en sorte que ce contenu soit déjà conceptuel. Mais elle doit aussi être capable de reproduire les relations rationnelles en jeu dans l'esprit, afin de pouvoir rendre compte des raisonnements humains, tout en partant d'une base naturelle, c'est-à-dire irrationnelle car le cerveau se trouve dans l'espace des lois. L'IA doit donc être en mesure de développer des capacités d'apprentissage lui permettant d'acquérir des modes de raisonnement non-naturels.

Pour savoir si l'IA est capable de reproduire ces éléments de l'esprit selon McDowell, il va donc falloir que nous étudiions les deux principaux types d'IA ainsi que leur fonctionnement. Mais, avant cela, il semble essentiel que nous nous demandions quelle influence McDowell a eu sur Putnam et quelle pensée nouvelle a émergé chez ce dernier. Cela nous permettra d'encore mieux saisir les subtilités du modèle de l'esprit de McDowell et tous les présupposées philosophiques sous-jacents. Mais, surtout, nous pourrons comprendre en quoi ce modèle rejoint la pensée de Putnam et s'oppose au fonctionnalisme.

Chapitre 2 : Putnam et le réalisme naturel

Hilary Putnam a théorisé le fonctionnalisme computationnel pour ensuite le rejeter. Nous avons dit que McDowell a influencé ce changement de pensée mais regardons cela de plus près afin de mieux saisir en quoi la pensée de McDowell s'inscrit dans une rupture avec les positions de la philosophie analytique majoritaires du XX^{ème} siècle, notamment avec l'idée de décrire et d'expliquer l'esprit humain au travers des sciences.

1. Le réalisme scientifique et son enfant le fonctionnalisme computationnel

Putnam fut au début de sa carrière ce qu'il a nommé un « réaliste scientifique ». Les partisans du réalisme scientifique « soutiennent qu'il est justifié de croire à la vérité approximative de nos théories scientifiques »[1]. Ce réalisme se rapproche (ou plutôt est le dérivé) de ce que Putnam appelle le réalisme métaphysique. Il le nomme également « réalisme moderne » dans son ouvrage *La Triple Corde*[2] car il s'agit selon lui de la forme de réalisme apparut au XVII^{ème} siècle en Europe en pleine période de la philosophie moderne avec des penseurs comme Descartes. Ainsi, selon Putnam, le réalisme métaphysique repose sur l'idée suivante :

> […] le monde est constitué d'un ensemble fixe d'objets indépendants de l'esprit. Il n'existe qu'une seule description vraie de « comment est fait le monde ». La vérité est une sorte de relation de correspondance entre des mots ou des symboles de pensée et des choses ou des ensembles de choses extérieures. J'appellerai ce point de vue externalisme, parce qu'il adopte de préférence une perspective qui est celle du point de vue de Dieu.[3]

Le réalisme métaphysique incarne l'idée selon laquelle il n'y a qu'une seule vérité possible, qu'une manière pour nos connaissances de correspondre à la réalité. C'est comme si nous pouvions adopter un point de vue de dieu, un point de vue objectif permettant de rendre compte de la réalité du monde et de la correspondance effective entre nos pensées et cette réalité extérieure. Notons qu'il y a une notion d'interface entre l'esprit et ce monde. Cette interface, ce sont nos

[1] Pierre-Yves Rochefort, « Putnam (A) », in *L'encyclopédie philosophique*, 2017, https://encyclo-philo.fr/putnam-a.
[2] Hilary Putnam, *La triple corde*, éd. par Pierre Fasula, trad. par Pierre Fasula et al., Analyse et philosophie (Paris: Librairie philosophique J. Vrin, 2017).
[3] Hilary Putnam, *Raison, vérité et histoire*, trad. par Abel Gerschenfeld, Propositions (Paris: Édition de Minuit, 1984), page 61.

représentations qui proviennent du monde dit réel. Le monde crée en nous des représentations et notre pensée a un contact direct avec celles-ci. Notre contact avec le monde est donc, lui, indirect. Pour Putnam, ce réalisme vient en partie de Descartes, lequel s'est longuement posé la question de l'adéquation entre ce qu'il voit et la réalité effective du monde.

C'est donc dans le cadre du réalisme scientifique (durant ses débuts en tant que chercheur académique en philosophie) que Putnam mit en place le fonctionnalisme computationnel en 1967, via son article « La nature des états mentaux »[1]. Selon cette théorie, notre esprit fonctionnerait de manière analogue à une machine de Turing c'est-à-dire à un automate qui exécute des programmes. Une machine de Turing est une machine abstraite qui exécute des algorithmes. Elle contient généralement quatre paramètres : il y a le paramètre d'entrée (ou *input*) qui correspond aux données que reçoit le système, il y a le paramètre de l'état actuel correspondant à la valeur ou au nom de l'état dans laquelle le système se trouve à l'instant T, il y a le paramètre de sortie (ou d'*output*) qui correspond aux comportements provoqués par la combinaison des *inputs* et de l'état actuel et enfin il y a le paramètre de l'état suivant (lié à l'*output*) qui correspond au nouvel état dans lequel le système va basculer à la suite de la résolution comportementale.

Ainsi, selon la théorie de Putnam, l'esprit est une machine de Turing. Les états mentaux sont des états fonctionnels c'est-à-dire qu'un état est un état mental si et seulement s'il a un rôle fonctionnel au sein du système qu'est notre corps. Des états comme la douleur par exemple sont des états mentaux car ils remplissent la fonction d'informer l'organisme que celui-ci a subi un dommage, ce qui a pour effet de modifier notre comportement en nous faisant dire « ouille » par exemple. Comme pour une machine de Turing, chaque organisme fonctionne alors selon une table qui comprend les quatre paramètres décrits plus hauts dans des configurations uniques définissant par avance que telle combinaison d'états et d'*inputs* provoquera tels *outputs*. Cela implique que la table est écrite à l'avance. Il paraît alors possible en principe de la trouver, de la modéliser. Le fonctionnalisme s'inscrit donc dans un projet de naturaliser l'esprit humain car il stipule notamment que les relations entre les différents paramètres du système sont des relations causales, des relations soumises aux lois naturelles, donc explicables via les sciences physiques.

[1] Putnam, « La Nature Des États Mentaux ».

2. Le réalisme interne et la révolution de l'externalisme sémantique

En 1975, Putnam rédige l'article « La signification de "signification" »[1] dans lequel il théorise ce qu'il nomme « l'externalisme sémantique ». Cette théorie stipule que la référence de nos termes, leurs extensions, c'est-à-dire ce qu'ils désignent, détermine au moins en partie leur signification. C'est donc l'environnement du locuteur, naturel et social, qui détermine la signification des termes qu'il utilise. Pour Putnam, il y a une « division du travail linguistique » c'est-à-dire que, pour les cas d'espèces naturelles par exemple, ce sont les experts ceux qui sont capables, avec des méthodes et des instruments, de confirmer que l'objet désigné par le terme que l'on emploie correspond bien à ce à quoi le terme fait normalement référence. En effet, car la plupart des autres individus disposent d'une signification déformée ou incomplète en raison d'un manque de connaissances provenant des méthodes de vérification. Ces significations incomplètes, Putnam les appelle « stéréotypes ». Par exemple, nous pouvons nous demander si tel métal que l'on a trouvé est de l'or ou non car nous ne sommes pas en mesure de le déterminer par nous-mêmes. La population n'est généralement pas capable de distinguer l'or du vermeil (ou de distinguer l'orme et le hêtre pour ce qui est des arbres). Elle se remet donc à l'avis des experts. Ceux-ci peuvent alors étudier la composition atomique de ce métal afin de dire au reste de la population si cet objet est bien de l'or ou non. Notre utilisation du mot « or » dépend donc de l'avis de ces experts qui pourrons nous corriger et nous amener à affiner notre stéréotype. Mais ce ne sont pas ces scientifiques qui déterminent le sens du mot « or », c'est l'environnement qui a fixé cette signification. Leurs méthodes ne sont finalement que des méthodes de vérification et de confirmation. C'est l'environnement qui fixe causalement la signification de nos termes. Certes, nous créons socialement les mots et décidons de le rapporter à un type d'objet en particulier présent dans notre environnement. Nous déterminons donc la référence « normale » de nos termes, mais c'est l'environnement qui va agir *in fine* sur la signification. Celle-ci n'est pas une simple construction mentale, l'environnement exerce des contraintes causales sur les experts et les personnes afin de rectifier et d'affiner nos significations dans le but d'exclure ou non des objets en tant que références valables aux termes employés. C'est l'environnement qui dicte la bonne signification des termes.

[1] Hilary Putnam, « La signification de « signification » », in *Textes Clés de philosophie de l'esprit Vol. II : Problèmes et perspectives*, trad. par Dominique Boucher, Textes clés (Vrin, 2003).

L'externalisme sémantique représente une découverte majeure dans la pensée de Putnam car cela va l'amener à rejeter progressivement le réalisme scientifique pour défendre ce qu'il nomme le « réalisme interne ». Ce réalisme consiste au fait de :

> […] soutenir que la question « De quels objets le monde est-il fait » n'a de sens que dans une théorie ou une description. Beaucoup de philosophes « internalistes », mais pas tous, soutiennent aussi qu'il y a plus d'une théorie ou description « vraie » du monde. La « vérité » est pour l'internalisme une sorte d'acceptabilité rationnelle (idéalisée) - une sorte de cohérence idéale de nos croyances entre elles et avec nos expériences telles qu'elles sont représentées dans notre système de croyances - et non une correspondance avec des « états de choses » indépendants de l'esprit ou du discours. Il n'y a pas de point de vue de Dieu qui soit connaissable ou utilement imaginable ; il n'y a que différents points de vue de différentes personnes, qui reflètent les intérêts et les objectifs de leurs descriptions et leurs théories. (« Théorie de la vérité-cohérence », « Non-réalisme », « Pragmatisme », sont quelques-uns des noms que l'on donne au point de vue internaliste. Mais ils ont tous des connotations inacceptables du fait de leurs autres applications historiques).[1]

Selon cette conception, il n'y a pas une seule description vraie du monde, une seule vérité. La vérité n'est plus une correspondance entre la pensée et le monde mais une cohérence entre nos différentes croyances. C'est donc une conception internaliste puisque l'on ne sort pas de notre corps. Il n'y a pas de point de vue objectif, de point de vue de dieu. Le réaliste interne considère son corps et son mental comme un système de croyances internes qui n'a de contact avec l'extérieur que via des représentations qui se forment en lui. Nous ne voyons le monde extérieur que par le biais d'une interface, d'un intermédiaire. Cette notion d'interface est donc la seule caractéristique commune entre le réalisme métaphysique et le réalisme interne. Ce réalisme n'abandonne pas la notion d'interface entre nous et le monde, au fait que nous n'avons que des représentations. Au contraire, si on est un réaliste interne, alors on considère que chacun est cantonné à ses propres représentations lui offrant une vision unique du monde.

Ainsi, l'externalisme sémantique rend parfaitement compte de ce réalisme interne puisque la détermination de la référence de nos termes dépend de l'environnement au sein duquel chacun évolue. Ainsi, chaque personne, de par ses expériences vécues, finit par avoir un réseau de croyances qui lui est unique. Il est d'autant plus vrai que ce réseau est unique puisque chacun va former en lui un stéréotype, une signification imparfaite et subjective, liée à chaque signification. Deux personnes n'ont donc peut-être pas exactement la même signification au terme « chat » car cette signification est en partie déterminée causalement par l'environnement de chacun. Ma

[1] Putnam, *Raison, vérité et histoire*, page 61-62.

signification du mot « chat » ne sera pas la même que celle d'un individu vivant au Japon puisque nous n'aurons pas vu les mêmes chats. Mais, qui plus est, je dispose d'un stéréotype associé au mot « chat » qui m'est unique. Cela signifie donc que ma signification du mot « chat » diffère même de celles des autres français avec lesquels je partage le même environnement. Ainsi, personne n'a le même réseau de croyances, chacun a sa propre vision du monde.

Le réalisme interne va amener Putnam à remettre en question sa propre théorie du fonctionnalisme computationnel. La critique la plus importante de Putnam à l'encontre du fonctionnalisme computationnel se retrouve dans son ouvrage *Représentation et Réalité*[1]. Il y expose les différentes formes de fonctionnalisme qui sont apparues depuis son article « La nature des états mentaux »[2] et démontre pour chacune d'entre elles l'impossibilité de rendre compte de la réalité de l'esprit humain. Nous aborderons certains de ces arguments lorsque nous nous demanderons en quoi le réalisme interne (et par extension le réalisme naturel) est antagoniste au fonctionnalisme initialement théorisé par Putnam.

Cette critique de sa propre théorie computationnelle de l'esprit forme également un argument à l'encontre d'une forme de réalisme métaphysique utilisant le fonctionnalisme computationnel afin de tenter d'expliquer, réduire de manière objective tous les esprits en des termes scientifiques. Or, les arguments de Putnam montrent que la possibilité de réduire toutes les capacités de notre esprit en des termes physiques et computationnels est illusoire. Il montre également que des choses comme la signification de nos termes ou plus globalement notre manière de voir le monde dépend de notre histoire personnelle, de nos expériences vécues et de notre environnement. Il ne semble donc pas qu'il soit possible de soutenir une conception du monde selon laquelle il n'y a qu'une seule vérité, une seule manière de faire correspondre notre pensée à la réalité. Il semble que défendre un point de vue divin, un point de vue objectif soit fallacieux. *Représentation et réalité* symbolise donc parfaitement le deuxième Putnam, un Putnam du réalisme interne.

[1] Hilary Putnam, *Représentation et réalité*, trad. par Claudine Tiercelin, NRF essais (Paris: Gallimard, 1990).
[2] Putnam, « La Nature Des États Mentaux ».

3. Le troisième Putnam ou le réalisme naturel

Durant les années 1970, Putnam commença à étudier les œuvres de William James. Cela l'amena également à se pencher sur les œuvres de John Langshaw Austin ainsi que sur celles de McDowell. Nous avions vu que le réalisme métaphysique issu de Descartes présenté par Putnam et le réalisme interne sont des positions qui considèrent qu'il y a une interface entre nous et le monde, que tout ce dont nous avons accès sont des représentations formées via les sens. Or, Putnam va découvrir les problèmes mis en avant par McDowell comme celui de la communication, du rapport qu'il y a entre les pensées et le monde débouchant sur une oscillation entre idéalisme voire solipsisme et réductionnisme radical. Progressivement, Putnam va alors commencer à se convaincre « du fait que ces problèmes traditionnels reposaient tous sur une conception fautive de la perception »[1]. Austin, James et McDowell vont lui montrer qu'il y a une troisième voie dans le réalisme qui est possible. Et, cette troisième voie, Putnam en avait besoin car, que ce soit via l'externalisme sémantique ou via son rejet des différentes théories de l'esprit physicalistes, réductionnistes ou idéalistes, il voyait, comme McDowell, un mouvement inlassable de va-et-vient. En effet, il dit :

> […] je ne voyais notamment ni comment défendre le réalisme ni comment il pouvait y avoir une quelconque autre manière de comprendre la relation du langage à la réalité. Il me semblait que je me m'étais empêtré dans des antinomies sans issue ![2]

Cette troisième voie, Putnam va l'appeler « réalisme naturel ». Il reprend cette étiquette à « William James, lorsqu'il exprime le désir d'obtenir une théorie de la perception qui rende justice au "réalisme naturel de l'homme commun" »[3]. Son nouveau réalisme s'inspire donc fortement de James mais aussi d'Austin et McDowell. Il reconnaît d'ailleurs l'influence de McDowell dans la formation de cette nouvelle pensée :

> Bien que je ne souhaite pas rendre McDowell responsable de mes formulations dans ces conférences, je tiens à reconnaître l'étendue de l'influence de son œuvre, qui a renforcé mon propre intérêt pour le réalisme naturel dans la théorie de la perception - un intérêt qui fut d'abord réveillé par une réflexion sur les thèses de William James.[4]

[1] Putnam, *La triple corde*, page 147.
[2] Putnam, page 35.
[3] Putnam, page 29.
[4] Putnam, page 20.

La pensée de McDowell a donc exercé une influence décisive dans ce troisième revirement de la pensée de Putnam. En effet, il affirme désormais la volonté de sortir de cette oscillation entre réalisme et idéalisme, volonté également partagée par McDowell notamment au travers de son ouvrage *L'esprit et le monde*. Il faut que cette nouvelle voie dont Putnam a besoin soit radicale. Il faut que celle-ci ne s'inscrive ni dans la conception réaliste moderne, ni dans l'idéalisme. Il faut rejeter toute conception qui s'en approche, il faut faire table rase :

> J'éprouve aujourd'hui plus fortement que jamais le besoin de trouver une « troisième voie » entre le réalisme moderne et l'idéalisme dummettien ; mais il faut qu'une telle troisième voie coupe court, comme l'a instamment répété McDowell, à l'idée qu'il existe une antinomie, et non qu'elle colle simplement ensemble des éléments du réalisme moderne et des éléments de l'image idéaliste. Toute conception qui conserve une notion semblable à la notion traditionnelle des *sense-data* ne saurait nous offrir une voie de sortie ; toute conception de ce genre nous met, *in fine*, face à un problème qui paraît insoluble.[1]

Nous pouvons alors nous demander pourquoi les positions idéalistes et matérialistes sont dans l'impasse. Pourquoi n'arrive-t-on pas à sortir de cette inlassable oscillation ? Pour Putnam, c'est la pensée de McDowell, dans la lignée de James et Austin, qui expose la source du problème :

> Selon McDowell, le présupposé central responsable de ce désastre est l'idée selon laquelle il doit exister une interface entre nos capacités cognitives et le monde extérieur - ou, autrement dit, l'idée selon laquelle nos capacités cognitives ne peuvent atteindre les objets par elles-mêmes.[2]

La connaissance du responsable de ce « désastre » est ce qui est constitutif du nouveau réalisme naturel putnamien. La caractéristique principale de ce réalisme est le rejet de l'idée d'interface, de représentations, de *sense-data,* c'est-à-dire des données issues des sens que notre cerveau va utiliser pour former des qualia, des images. Il faut abandonner l'idée d'interface et accepter l'idée que l'esprit ait un rapport direct avec le monde, c'est-à-dire que qu'il y ait du concept dans la sensibilité.

Le réalisme moderne (ou métaphysique) avait déjà reçu des critiques de la part de Putnam mais, ici, il va plus loin. Tout d'abord, Putnam reconnaît que le réalisme interne ne tient pas la route. À la base, le réalisme interne s'oppose au réalisme métaphysique comme nous l'avons vu, notamment via l'externalisme sémantique mais aussi parce que Putnam considère le réalisme métaphysique comme inintelligible :

[1] Putnam, page 40.
[2] Putnam, page 29.

> […] je l'ai identifié [le réalisme interne] avec le rejet de la présupposition réaliste traditionnelle (1) d'une totalité fixe de tous les objets ; (2) d'une totalité fixe de toutes les propriétés ; (3) d'une ligne de démarcation entre les propriétés que nous « découvrons » dans le monde et les propriétés que nous « projetons » sur lui ; (4) d'une relation fixe de « correspondance » supposée définir la vérité. J'ai rejeté ces quatre présupposés non parce qu'ils sont simplement faux, mais parce qu'ils sont en fin de compte inintelligibles.[1]

Cependant, le réalisme interne tombe, selon Putnam, dans les mêmes travers, le même formalisme inintelligible du réalisme métaphysique. Le formalisme épistémique du réalisme interne est le même que l'épistémologie naturelle car il y a toujours sous-entendu l'idée d'une interface entre l'homme et l'extérieur. Cette notion d'interface, l'idée qu'il y a d'une part la réalité et de l'autre nos représentations séparées par la barrière des sens, cela est pour Putnam inintelligible. Ainsi, le réalisme interne est également inintelligible et Putnam avoue lui-même ne pas vraiment savoir ce qu'est ce réalisme *in fine* :

> Aussi, que je sois ou non encore un réaliste interne, cela est, j'imagine, aussi peu clair que de savoir ce que je comprenais sous cette étiquette malheureuse.[2]

Mais alors nous pouvons nous demander en quoi la notion d'interface, de *sense-data*, est inintelligible. Pour Putnam, cette notion, introduite notamment par Russell, n'est qu'une autre manière de parler de représentations et remonte alors à Descartes. La théorie du *sense-datum* postule que des évènements dans le cerveau produisent des *sense-data*, des données issues des sensations, mais Putnam nous fait remarquer que personne ne sait vraiment comment cela fonctionne. En effet, les partisans immatérialistes comme Descartes

> […] ne s'accordaient même pas sur la question de savoir si les « sense-data » étaient des parties des esprits individuels (comme dans la conception de Hume), s'ils appartenaient d'une manière ou d'une autre à ces esprits sans en être des parties (remarquez le possessif dans l'expression berkeleyenne: « les esprits et leurs idées »), s'ils étaient des particuliers ou des qualités (Nelson Goodman pense même que les qualités sont des particuliers)' ou, au vingtième siècle, si le même sense-datum pouvait raisonnablement être «perçu immédiatement» par plus d'un esprit, ou même s'il était concevable que certains sense-data puissent exister sans être perçus.[3]

Scientifiquement parlant, une telle conception de l'esprit n'est pas intelligible. Il faudrait être en mesure d'expliquer comment et pourquoi un tel processus a lieu dans notre esprit. Sans cela, la théorie des *sense-data* possède un aspect mystérieux.

[1] Putnam, page 40.
[2] Putnam, page 40.
[3] Putnam, page 55-56.

Pour apporter une explication, Putnam explique que nous pourrions défendre une conception matérialiste de la théorie des *sense-data*. En effet, il y a de nombreuses positions philosophiques qui sous-tendent les *sense-data* comme la théorie de l'identité. Or, cette théorie et toutes celles qui en dérivent posent problème également. D'après Putnam, la théorie de l'identité est « la théorie selon laquelle les sensations et les pensées ne sont que des processus cérébraux »[1]. Toujours selon Putnam, l'une des propositions les plus populaires donnant un sens à la théorie de l'identité est que « l'"identité" est une "token-identité anomale" »[2]. Cette proposition a été formulé par Davidson qui pense que :

> […] l'on ne peut pas trouver d'identités « de type à type » entre des événements tombant sous une description psychologique et des événements tombant sous une description physique, mais il affirme néanmoins que chaque événement individuel où quelqu'un pense telle et telle chose ou fait telle et telle expérience est « identique » à un événement physique ! Dans le jargon de Davidson, chaque « token » d'événement mental est identique à un « token » d'événement physique.[3]

Ainsi, un certain état mental apparaît quand un certain état physique apparaît. Nous pouvons alors dire que l'état mental est identique à l'état physique. Nos processus mentaux et leurs relations sont donc identiques à des processus physiques et leurs relations physico-chimiques. Cela permet alors de rendre les *sense-datum* intelligibles selon Putnam car :

> Si les sense-data (les « qualia ») sont eux-mêmes des événements cérébraux plutôt que des effets immatériels d'événements cérébraux, alors on évite certainement le problème qui porte sur la manière dont un événement matériel est censé causer un événement immatériel dans l'esprit.[4]

Les *sense-data* émergent donc dans l'esprit lorsque certains états physiques apparaissent dans le cerveau, notamment à la suite de stimulation sensorielles. Les *sense-data* sont là car certains états physiques sont présents. Il y a donc une explication de pourquoi et comment ce processus a lieu, une explication matérialiste. Sans cerveau, pas de *sense-data*. Finalement, cela revient à considérer que les phénomènes mentaux seraient localisés uniquement dans le cerveau et l'esprit et n'auraient qu'un rapport indirect avec l'extérieur. Ces phénomènes auraient des relations cognitives spéciales entre eux et il y aurait des chaines de relations causales qui les relieraient aux objets extérieurs.

[1] Putnam, page 57.
[2] Putnam, page 59.
[3] Putnam, page 64-65.
[4] Putnam, page 57.

Mais si les états mentaux sont identiques à des états physiques, où se trouvent les états physiques de la conscience ? Comme le constate Putnam : « les cerveaux ont des centres du langage, des aires pour différentes sortes de mémoire, etc., mais aucun centre de la conscience »[1]. Ce que montre Putnam, c'est que nous ne savons pas réellement de quelle identité entre les états mentaux et les états physiques il s'agit. De même, si identité il y a, alors elle doit être possible en vertu d'une loi naturelle. Mais qu'elle est cette loi ? Nous ne le savons pas. Ici aussi, il y a un aspect mystérieux concernant la théorie de l'identité. C'est une théorie qui n'est pas pleinement intelligible.

Putnam ne se concentre pas que sur la théorie de l'identité. Il cherche en effet à montrer que toutes les théories ayant pour sous-entendu l'idée de représentation mentale, l'idée d'interface entre notre conscience et le monde, ne sont pas pleinement intelligibles. Or, si ces conceptions ne sont pas intelligibles, ce n'est pas vraiment sérieux de les utiliser. Au contraire, comme le pense Putnam, il vaut mieux abandonner cette idée d'interface pour revenir à une notion plus simple, c'est-à-dire supposer à nouveau une interaction directe entre l'esprit et le monde.

Ainsi, comme toute théorie stipulant une notion d'interface, le réalisme interne ne fonctionne pas. Il ne permet pas de proposer une alternative crédible au réalisme métaphysique car, comme ce dernier, il n'est pas totalement intelligible. C'est pour cela que l'intérêt de Putnam pour cette nouvelle forme de réalisme qu'est le réalisme naturel est important. Selon lui, il vaut mieux abandonner l'idée d'interface pour revenir à une notion plus simple c'est-à-dire supposer à nouveau une interaction directe entre l'esprit et le monde. C'est ce que propose le réalisme naturel qui est un réalisme direct, une manière de voir notre rapport au monde sans interface, une relation directe avec ce qui est extérieur à nous. Le réalisme naturel peut alors être considéré comme la forme intelligible, la forme évoluée, du réalisme interne. Finalement, il s'agit de revenir à une forme de philosophie du sens commun, une philosophie naïve mais qui n'est plus si naïve que ça.

C'est ce que Putnam nomme « une seconde naïveté »[2]. Il s'agit d'un retour à une pensée prémoderne, pré-cartésienne, une pensée d'apparence plus simple mais bien plus consistante et sensée. Pour qu'il y ait un tel retour de la naïveté dans la perception par exemple, il faut qu'il y ait le réalisme naturel. Un tel retour au simple ne permet pas de postuler une notion d'interface entre le nous et le monde. En fait, la seconde naïveté est le fruit d'une réflexion qui a duré quatre cents

[1] Putnam, page 58.
[2] Putnam, page 34.

ans, c'est l'accomplissement du réalisme naturel, d'un nouveau pas franchi dans le monde philosophique. Comme Putnam le dit :

> Parvenir au réalisme naturel, c'est voir l'inanité et l'inintelligibilité d'une image qui impose une interface entre nous et le monde. C'est une manière d'accomplir la tâche de la philosophie, la tâche que John Wisdom nomma autrefois un « voyage du familier au familier ».[1]

Le réalisme naturel peut également être nommé « réalisme pragmatique » même si Putnam n'emploie jamais vraiment cette appellation. Ce réalisme a un aspect pragmatique car il prend en compte les choses qui affectent nos vies, nos manières de vivre et de voir le monde quotidiennement. C'est une philosophie qui se rapproche de la philosophie populaire et la prend en compte. Ainsi, pour Putnam :

> […] s'il y avait une grande idée du pragmatisme, ce serait bien l'insistance sur le fait que ce qui a du poids dans nos vies devrait également en avoir en philosophie.[2]

Par ce retour à une forme de naïveté, Putnam rejoint la pensée de McDowell selon laquelle les capacités de l'esprit dérivent du cerveau et du long processus de l'évolution tout en n'étant pas réductible à l'espace du règne des lois. Pour Putnam, le discours sur l'esprit est un discours sur certaines de nos capacités, ce qui s'accorde avec l'espace logique des raisons que McDowell présente comme étant une capacité que nous avons et que nous pouvons développer via l'éducation. Putnam parle d'autres capacités comme la capacité à se souvenir, imaginer ou s'attendre à. Et, comme McDowell, Putnam défend la place centrale qu'a le langage pour notre développement cognitif. Pour lui :

> Le langage étend les capacités mentales que nous partageons avec les animaux de façon presque infinie ; notre capacité de construire des théories scientifiques sophistiquées n'en est qu'un exemple parmi d'autres. Qu'on pense au rôle des constantes logiques telles que les mots tous et aucun. Un animal ou un enfant qui n'a pas encore appris l'usage de ces mots peuvent avoir des attentes que nous, qui avons acquis ces mots, pouvons décrire et décrivons grâce à eux.[3]

Le langage étend nos capacités. Le langage altère notre pensée, notre manière de voir le monde. Il permet à notre pensée de mieux discriminer et de faire de nouvelles expériences. Le langage permet aussi, comme le montre l'externalisme sémantique, à l'environnement d'agir sur notre manière de penser notamment de penser le monde. Les différents langages n'ont pas tous

[1] Putnam, page 71.
[2] Putnam, page 106.
[3] Putnam, page 90.

exactement le même rapport au monde car ils se sont développés dans des environnements différents. Le langage est donc, pour Putnam comme pour McDowell, un élément central dans le développement de l'esprit humain.

Le réalisme naturel s'accorde avec le naturalisme modéré de McDowell. Il ne doute pas que les capacités de l'esprit se forment dans le cerveau qui, lui, est soumis aux lois physiques. Il présuppose aussi l'externalisme sémantique et le holisme dans la détermination des croyances. C'est un réalisme qui place les humains au sein du monde, il les place comme des animaux qui ont certaines capacités que l'on regroupe sous l'appellation d'esprit. L'environnement agit sur nous jusque dans notre manière de penser et nous agissons sur lui en retour. La relation entre nous et l'extérieur est directe, il n'y a pas d'intermédiaires, de *sense-data* qui relèvent de la science-fiction. Ce qui agit sur nous, et ce sur quoi nous agissons en retour, ne sont pas des illusions, des mirages, mais bien une réalité tangible au sein de laquelle nous vivons.

Ce réalisme est supposé aller à l'encontre du projet de réduire l'esprit en des termes computationnels. Pour comprendre cela, nous devons d'abord nous demander pourquoi le réalisme interne est antagoniste au computationnalisme. En effet, comme nous l'avons vu, le réalisme naturel peut être considéré comme une évolution du réalisme interne. Or, il s'avère que Putnam était partisan du réalisme interne quand il a présenté des arguments à l'encontre du fonctionnalisme computationnel. Donc, si nous comprenons les arguments dans un cadre réaliste interne, alors nous les comprendrons aussi dans un cadre réaliste naturel.

Chapitre 3 : Le réalisme interne et naturel contre le fonctionnalisme

Pour mieux comprendre pourquoi le réalisme interne est antagoniste au fonctionnalisme computationnel, nous devons examiner ce premier de plus près. En effet, nous avons les caractéristiques principales de ce réalisme mais, les arguments que mobilise Putnam à l'encontre du fonctionnalisme requiert des précisions pour que nous puissions pleinement les comprendre. C'est précisions, ces concepts et ces thèses sont d'autant plus importantes qu'elles conservent leur pertinence dans le réalisme naturel.

1. Le réalisme interne implique l'externalisme sémantique et le holisme sémantique

Dans le cadre du réalisme interne, Putnam tient pour vrai deux théories majeures : l'externalisme sémantique et le holisme sémantique. Nous avons ce qu'était l'externalisme sémantique mais, maintenant, tâchons de comprendre le holisme sémantique.

Tout d'abord, il y eut le holisme épistémologique. Le holisme épistémologique fut initialement présenté par Duhem dans son ouvrage *La Théorie physique. Son objet, sa structure*[1], notamment le chapitre 6 dans lequel il explique que lorsque l'on cherche à démontrer qu'une proposition est inexacte à l'aide d'une expérience, ce n'est pas uniquement la proposition en litige qui est concernée mais tout un ensemble de théories qui sont acceptées et sous-entendues par le scientifique. En effet, d'après Duhem, « le physicien qui exécute une expérience ou en rend compte reconnaît implicitement l'exactitude de tout un ensemble de théories »[2]. Cela a alors pour conséquence que

> […] si le phénomène prévu ne se produit pas, ce n'est pas la proposition litigieuse seule qui est mise en défaut, c'est tout l'échafaudage théorique dont le physicien a fait usage ; la seule chose que nous apprenne l'expérience, c'est que, parmi toutes les propositions qui ont servi à prévoir ce phénomène et à constater qu'il ne se produisait pas, il y a au moins une erreur ; mais où gît cette erreur, c'est ce qu'elle ne nous dit pas.[3]

[1] Pierre Duhem, *La théorie physique. Son objet, sa structure*, Bibliothèque idéale des sciences sociales (Lyon: ENS Éditions, 2016).
[2] Pierre Duhem, « La théorie physique et l'expérience », in *La théorie physique. Son objet, sa structure*, Bibliothèque idéale des sciences sociales (Lyon: ENS Éditions, 2016).
[3] Duhem, « La théorie physique et l'expérience ».

Ainsi, la localisation de l'erreur dans un ensemble de théories est sous-déterminée par l'expérience. Lorsqu'une expérience réfute un ensemble de théories, nous savons qu'une des hypothèses faisant partie de l'ensemble doit être modifiée mais nous ne pouvons pas savoir qu'elle est hypothèse en question.

Inversement, Duhem explique qu'une expérience ne peut confirmer une théorie en particulier mais toujours un « ensemble théorique » pris comme un tout. Il n'est donc jamais possible de confirmer une théorie en particulier mais il est possible de confirmer l'ensemble théorique. Cependant, les hypothèses des sciences physiques ne sont pas forcément contradictoires entre elles et peuvent toutes être confirmées par les mêmes expériences. Il apparaît alors impossible selon Duhem, qu'une expérience décisive puisse isoler et confirmer de manière indubitable un ensemble théorique particulier. En effet, si jamais nous voulions y parvenir, Duhem nous explique la méthode ainsi :

> Énumérez toutes les hypothèses qu'on peut faire pour rendre compte de ce groupe de phénomènes ; puis, par la contradiction expérimentale, éliminez-les toutes, sauf une ; cette dernière cessera d'être une hypothèse pour devenir une certitude.[1]

Or, cela n'est pas possible car il faudrait être capable d'énumérer toutes les hypothèses et il s'avère qu'il n'y a aucun moyen d'être sûr que nous avons formulé toutes les hypothèses possibles. Peut-être qu'il y en a une qui échappe à notre imagination, nous ne pouvons pas le savoir. Ainsi, si nous ne pouvons jamais être sûr que nous avons épuisé toutes les possibilités d'hypothèses possibles, nous ne pouvons jamais confirmer de manière indubitable un ensemble théorique particulier. Peut-être qu'il existe une autre hypothèse résistant aux mêmes expériences, mais qui, elle, résisterait, contrairement à l'ensemble théorique que nous tenions pour vrai à la base, à une expérience supplémentaire. Ainsi, pour Duhem :

> La contradiction expérimentale n'a pas, comme la réduction à l'absurde employée par les géomètres, le pouvoir de transformer une hypothèse physique en une vérité incontestable ; pour le lui conférer, il faudrait énumérer complètement les diverses hypothèses auxquelles un groupe déterminé de phénomènes peut donner lieu ; or, le physicien n'est jamais sûr d'avoir épuisé toutes les suppositions imaginables.[2]

Il y a donc un holisme concernant les théories scientifiques présenté ici par Duhem. Les théories ne peuvent être réfutées ou confirmées isolément, elles dépendent d'un tout, d'un ensemble théorique. Il apparaît également (et c'est une conséquence de cet holisme) qu'il est impossible de

[1] Duhem.
[2] Duhem.

réfuter une théorie en particulier ni de confirmer un ensemble théorique en particulier. L'expérience sous-détermine les théories.

Vient alors ensuite Quine qui, dans son article « Les deux dogmes de l'empirisme »[1], développe une théorie similaire d'holisme, si bien qu'on appelle holisme épistémologique la théorie Duhem-Quine. Cependant la version de Quine se distingue de celle de Duhem par sa radicalité.

Dans l'article, Quine propose une critique de deux dogmes constitutifs du positivisme logique, dont l'un des principaux représentants est Carnap, qui sont :

- L'idée qu'il y a une opposition entre les énoncés analytiques et les énoncés synthétiques,
- L'idée que la signification d'un énoncé est liée à son mode de vérification sensoriel.

Ainsi, il estime d'abord qu'il n'y a pas de démarcation stricte entre les énoncés analytiques et les énoncés synthétiques. Un énoncé analytique est un énoncé dont le prédicat est contenu dans le sujet de manière explicite ou implicite. Un énoncé analytique est vrai en vertu de la signification. Et, parmi les énoncés analytiques, Quine distingue les énoncés logiques et les énoncés analytiques par synonymie. C'est ce dernier type d'énoncé qu'il va étudier afin de comprendre et expliquer l'analyticité car il estime que la synonymie peut donner du sens à la signification. Ainsi, pour Quine « on peut dire qu'un énoncé est analytique simplement lorsqu'il est synonyme avec un énoncé logiquement vrai »[2]. Pour lui, si on arrivait à montrer que deux énoncés ont la même signification, donc mettre en lumière la synonymie, et dont un de ces énoncés est logiquement vrai, alors on arriverait à dire qu'un énoncé analytique est vrai en vertu de la signification. La synonymie entre deux expressions a quelque chose de plus que la simple coextensivité. Nous avons tendance à supposer que la synonymie permet aux expressions d'être interchangeables dans des propositions sans que cela ne change la valeur de vérité. Or, pour Quine, c'est le cas aussi pour les expressions coextensives. Les expressions coextensives sont interchangeables et conservent la valeur de vérité des propositions. Donc, pour Quine :

[1] Willard Van Orman Quine, « Les deux dogmes de l'empirisme », in *De Vienne à Cambridge : L'héritage du positivisme logique de 1950 à nos jours*, éd. par Pierre Jacob, Tel (Gallimard, 1980), 93-121.
[2] Quine, page 112.

> Il nous faut donc reconnaître que la substituabilité mutuelle *salva veritate*, conçue dans le cadre d'une langue extensionnelle, n'est pas une condition suffisante de la synonymie cognitive dans le sens requis pour dériver l'analyticité.[1]

Il faut quelque chose en plus et, pour Quine, c'est la modalité de l'expression qui fait la différence. En effet, selon lui :

> Si une langue contient l'adverbe intensionnel « nécessairement » (au sens mentionné) ou d'autres particules de ce genre, alors la substituabilité mutuelle *salva veritate* dans une telle langue constitue effectivement une condition suffisante de la synonymie cognitive.[2]

Nous pouvons distinguer deux expressions synonymes de deux expressions coextensives si nous pouvons substituer les expressions au sein de propositions sans que cela ne change rien, que ce soit au niveau de la valeur de vérité et au niveau de la modalité. Deux expressions qui sont interchangeables en utilisant des modalités sont synonymes. La modalité nous permet de mieux concevoir le lien entre synonymie et analycité car il existe des modalités comme la nécessité et il s'avère qu'une vérité analytique est une vérité nécessaire. Si, en substituant des expressions, la nécessité est conservée et la proposition garde sa valeur de vérité, alors cela veut dire que les deux expressions sont synonymes. Elles partagent la même nécessité analytique.

Cependant, cela pose problème pour Quine. En effet, pour lui, « une telle langue n'est intelligible que dans la mesure où la notion d'analyticité est déjà préalablement comprise »[3]. Il développe son argument :

> On peut dire des termes singuliers qu'ils sont synonymes d'un point de vue cognitif, lorsque le jugement d'identité qu'on forme en insérant « = » entre eux est analytique. On peut simplement dire que certains énoncés sont synonymes du point de vue cognitif, lorsque leur biconditionnel (obtenu en les reliant par « si et seulement si ») est analytique.[4]

Ainsi, pour Quine, utiliser des modalités implique l'idée d'analycité. Par exemple, le mot « nécessairement » implique l'idée de nécessité et donc d'analycité. Il y a donc un problème car on voulait prouver l'analycité via la synonymie mais il s'avère que c'est l'inverse que nous sommes en train de faire. Il semble qu'il soit impossible de donner un sens à la notion d'analycité. C'est ce que Quine conclut :

> Il semblait d'abord qu'en ayant recours au royaume des significations, on définirait tout naturellement l'analyticité. Puis le recours aux significations s'est transmué en recours à la

[1] Quine, page 105.
[2] Quine, page 105.
[3] Quine, page 105.
[4] Quine, page 105.

> synonymie ou à la définition. Mais la définition s'est avérée n'être qu'un vœu pieux, et la synonymie n'être intelligible qu'en recourant à l'analyticité elle-même. Donc, nous voilà revenus au problème de l'analyticité.[1]

Cette opposition entre analytique et synthétique n'est donc peut-être qu'une chimère. En tout cas, elle n'est pas justifiée puisque nous n'arrivons pas à savoir ce qu'est l'analycité.

En ce qui concerne le deuxième dogme, Quine aborde l'idée que nos énoncés sont traduisibles dans la langue des données sensibles. Il se base notamment sur le projet réductionniste radical mené par Carnap qui « se donne pour tâche de spécifier une langue des données sensibles et de montrer comment on peut y traduire le reste du discours ayant un sens, énoncé par énoncé »[2]. D'après Quine, Carnap a fait de nombreuses avancées mais son programme était voué à l'échec. Il remarque d'ailleurs que Carnap a lui-même finit par abandonner un réductionnisme aussi radical. Pour Quine le dogme du réductionnisme est lié au clivage entre analytique et synthétique, ce qui a pour conséquence que ces deux dogmes sont identiques. Dans le réductionnisme radical, il y a l'idée qu'en général

> […] la vérité des énoncés dépend à la fois du langage et des faits extra-linguistiques ; nous avons vu que cette observation évidente peut conduire, sinon logiquement en tout cas (malheureusement) naturellement, au sentiment qu'on peut analyser la vérité d'un jugement en deux composantes, l'une linguistique, l'autre factuelle. Si l'on est empiriste, la composante factuelle se réduit à un ensemble de confirmations sensorielles. Dans le cas extrême, où la composante linguistique est la seule qui compte, un énoncé vrai est analytique. Mais j'espère qu'on arrive maintenant à apprécier combien la distinction entre l'analytique et le synthétique a obstinément survécu en dépit de l'absence de son tracé.[3]

Comme Carnap, Quine rejette l'idée de chercher la signification de termes. Ils privilégient donc de travailler sur les énoncés. Cependant, Quine rejette l'idée que l'on puisse confirmer ou infirmer les énoncés pris isolément. Lui, propose « l'idée que nos énoncés sur le monde extérieur sont jugés par le tribunal de l'expérience sensible, non pas individuellement, mais seulement collectivement »[4]. Pour lui :

> La totalité de ce qu'il est convenu d'appeler notre savoir ou nos croyances, des faits les plus anecdotiques de l'histoire et de la géographie aux lois les plus profondes de la physique atomique ou même des mathématiques pures ou de la logique, est une étoffe tissée par l'homme, et dont le contact avec l'expérience ne se fait qu'aux contours. Ou encore, pour

[1] Quine, page 106.
[2] Quine, page 113.
[3] Quine, page 116.
[4] Quine, page 115.

> changer d'image, l'ensemble de la science est comparable à un champ de forces, dont les frontières seraient l'expérience.[1]

Ainsi, selon cette conception holistique, on peut réévaluer des énoncés proches des bords via les données sensorielles. Les énoncés au centre du champ, eux, peuvent être révisés par d'autres énoncés mais pas vraiment par les données sensorielles directement. Il y a une gradation entre les vérités les plus proches des bords et celles les plus éloignés. Plus on s'éloigne, moins nous avons de lien avec le bord et moins nous aurons tendance à être remis en question. Il reste néanmoins que, selon Quine,

> […] le champ total est tellement sous-déterminé par ses frontières, c'est-à-dire par l'expérience, qu'on a toute liberté pour choisir les énoncés qu'on veut réévaluer, au cas où intervient une seule expérience contraire. Aucune expérience particulière n'est, en tant que telle, liée à un énoncé particulier situé à l'intérieur du champ, si ce n'est à travers des considérations d'équilibre concernant la totalité du champ.[2]

Nous retrouvons ainsi la même sous-détermination mise en avant par Duhem. Le holisme de Quine et celui de Duhem sont très similaires à ceci près que Quine est plus radical dans sa démarche. Pour lui, ce ne sont pas juste les théories scientifiques qui sont mobilisées par le holisme et affectées par la sous-détermination, mais ce sont la totalité de nos connaissances et croyances. Nos connaissances sont donc liées à toutes les autres et l'expérience sous-détermine ce tout de connaissances. Une expérience récalcitrante engage l'entièreté de notre réseau de connaissances et de croyances.

Mais il n'y a pas que le holisme épistémologique qui est présenté par Quine dans son article « Les deux dogmes de l'empirisme ». En effet, il est communément admis que Quine a introduit une autre forme d'holisme dans le texte, le holisme sémantique. Le fait de retrouver le holisme sémantique dans « Les deux dogmes de l'empirisme » est débattu et contesté, du moins le fait de retrouver pleinement le holisme sémantique tel qui est défini aujourd'hui comme nous pouvons le voir dans l'ouvrage *Holism: A Shopper's Guide*[3] de Fodor et Lepore. Ces derniers contestent le fait qu'il y ait l'apparition du holisme sémantique tel qu'il est défini aujourd'hui dans le texte de Quine, mais ils reconnaissent tout de même qu'il y a du holisme sémantique dans le texte, du moins des

[1] Quine, page 117.
[2] Quine, page 117.
[3] Jerry A. Fodor et Ernest Lepore, *Holism: A Shopper's Guide*, éd. par Ernest LePore (Cambridge, Mass., USA: Blackwell, 1992).

prémisses. Pour Fodor et Lepore, s'il y a une thèse du holisme sémantique dans le texte de Quine, on la trouverait dans le passage suivant :

> L'idée de définir un symbole en contexte par l'usage représenta, comme on l'a vu, un progrès, par rapport à l'empirisme terme à terme de Locke et Hume. Avec Bentham, on en vint à reconnaître l'énoncé, plutôt que le terme, comme l'unité requise par une critique empiriste. Ce que je prétends maintenant, c'est que même, en prenant pour unité l'énoncé, nous employons un tamis trop fin. L'unité de signification empirique est la totalité de la science.[1]

Notons que la référence à Bentham était une référence à Frege dans la version originale. En réalité, l'article a pris plusieurs formes. Il y eut d'abord une référence à Russel, puis à Frege et, enfin, à Bentham. Mais, pour Fodor et Lepore, cela ne change rien à leurs propos[2]. Ainsi, selon eux, il y a trois considérations qui feraient penser qu'il y a une formulation du holisme sémantique dans ce passage, au lieu de juste être une réitération du holisme épistémologique. Ces considérations sont :

> […] premièrement, la référence à Frege (qui, vraisemblablement, parlait d'unités de signification plutôt que d'unités de confirmation) ; deuxièmement, la critique traditionnelle selon laquelle « Deux dogmes » est un *locus classicus* du holisme sémantique ; et troisièmement, vu que le réductionnisme est explicitement considéré comme une doctrine à la fois sémantique et épistémologique dans « Deux dogmes », il est naturel d'interpréter son refus également à la fois comme une doctrine épistémologique et sémantique.[3]

Ainsi, l'article de Quine introduit non seulement une nouvelle forme de holisme épistémologique mais également le holisme sémantique.

Qu'est-ce donc alors que le holisme sémantique ? Pour Michaël Esfeld, « le holisme sémantique applique ce qui vaut pour des concepts théoriques à tous les concepts »[4]. En ce sens, le holisme sémantique a une portée plus large que le holisme épistémologique. Ainsi, Esfeld reprend les deux thèses constitutives du holisme sémantique présentées par Wilfrid Sellars que l'on retrouve dans l'ouvrage *Empirisme et philosophie de l'esprit*[5] :

> a. *Une thèse sur les conditions de maîtrise des concepts* : on ne peut pas acquérir de concepts pris isolément. Maîtriser un concept implique de maîtriser un certain

[1] Quine, « Les deux dogmes de l'empirisme », page 116.
[2] Fodor et Lepore, *Holism,* page 216.
[3] Fodor et Lepore, page 40-41 : « […] first, the reference to Frege (who, presumably, really was talking about the units of meaning rather than the units of confirmation); second, the critical tradition according to which "Two dogmas" is a locus classicus for semantic holism; and third, since reductionism is explicitly viewed as both a semantic and an epistemological doctrine in "Two dogmas," it's natural to construe its denial there as both an epistemological and a semantic doctrine too ».
[4] Michaël Esfeld, *La philosophie de l'esprit : Une introduction aux débats contemporains*, Cursus (Armand Colin, 2020), page 181.
[5] Sellars, *Empirisme et philosophie de l'esprit*.

> nombre d'autres concepts, y compris ceux qui fixent les conditions standard d'application du concept en question. Il n'est donc pas possible d'apprendre un langage mot par mot. D'après Sellars, l'enfant apprend en même temps une série de concepts qui constituent un langage rudimentaire. Puis il élargit ses capacités linguistiques en apprenant de nouveaux concepts et de nouvelles règles d'inférence (§ 19, p. 47 dans l'édition française).
>
> b. *Une thèse sur le contenu des concepts* : le contenu d'un concept se définit par des relations inférentielles à d'autres concepts, y compris aux concepts qui déterminent les conditions standard d'application du concept en question.[1]

Il semble que ce soit le holisme sémantique hérité de Quine que mobilise Putnam dans son argumentation à l'encontre du fonctionnalisme. C'est d'ailleurs ce que Paul Bernier remarque dans son compte rendu de l'ouvrage *Représentation et réalité* puisqu'il dit : « Pour ce qui concerne la thèse du holisme sémantique, Putnam s'appuie sur la thèse quinienne du holisme épistémologique. »[2]. Cependant Bernier ajoute que l'on doit noter

> […] qu'il n'est pas évident que cette thèse épistémologique puisse être directement transposée au niveau de l'analyse sémantique, surtout lorsqu'il s'agit d'analyser les notions de croyance et de désir qui sont le fait de la psychologie du sens commun.[3]

Nous avons vu cependant qu'il est possible de trouver au moins l'ébauche de la thèse du holisme sémantique dans les écrits de Quine. Il ne s'agit donc pas vraiment d'une transposition mais de l'utilisation d'une autre thèse à part entière.

Pour Putnam, le holisme sémantique s'accorde avec l'externalisme sémantique. Certes, l'environnement fixe causalement la référence des termes et donc la signification de ceux-ci. Cependant, la signification qu'a un individu, signification imparfaite que Putnam appelle le stéréotype comme nous l'avions indiqué précédemment, peut changer en fonction de l'environnement, de comment l'individu fait l'expérience du monde qui l'entoure et de l'état de son réseau conceptuel. En effet, la signification du terme présente dans l'esprit de l'individu est ancrée dans un réseau conceptuel, un réseau de croyances et de connaissances cohérentes propre à cet individu. Personne n'a jamais totalement la signification dans son entièreté dans la tête, c'est d'ailleurs pour cela que Putnam disait : « vous aurez beau retourner le problème dans tous les sens, rien à faire, les « significations » ne sont pas dans la tête ! »[4]. La signification que l'on a dans notre

[1] Esfeld, page 182-183.
[2] Paul Bernier, « Compte rendu de [Hilary Putnam, Représentation et réalité, traduction française par Claudine Engel-Tiercelin, Paris, Éditions Gallimard, 1990, 226 pages.] », *Philosophiques* 18, nº 2 (1991): 191-95, page 192.
[3] Bernier, page 192.
[4] Hilary Putnam, « La signification de « signification » », in *Textes Clés de philosophie de l'esprit Vol. II : Problèmes et perspectives*, trad. par Dominique Boucher, Textes clés (Vrin, 2003), page 57.

tête n'est pas complète et n'est peut-être pas adéquate. Elle est corrigée par nos pairs, comme les experts, qui seront en mesure de dire si l'objet que l'on désigne avec tel terme est vraiment la référence à ce terme. Nous nous trompons régulièrement et c'est l'environnement, via les experts, qui va nous permettre de corriger notre stéréotype du terme. Il n'empêche que chacun peut avoir son propre stéréotype, sa propre vision d'un phénomène en accord avec son réseau conceptuel.

Maintenant que nous voyons ce qu'est la théorie du holisme sémantique, son origine et le lien que nous pouvons faire avec l'externalisme sémantique, nous pouvons nous demander pourquoi le Putnam du réalisme interne tient pour vrai et utilise ces deux théories. La réponse est que le holisme sémantique et l'externalisme sémantique permettent de rendre le réalisme interne concret. Tâchons d'expliquer pourquoi.

Le réalisme interne est apparu comme une rupture par rapport au réalisme métaphysique que Putnam défendait jusqu'alors. Historiquement, cette évolution s'est notamment faite dans le cadre d'une prise de position de Putnam à l'encontre du projet de naturaliser la notion de vérité soutenue par Hartry Field. Selon Field, ce projet serait possible grâce à l'externalisme sémantique, qui reconnaît un lien causal entre le terme et l'objet qu'il désigne, car nous pouvons alors stipuler qu'il est possible de naturaliser la notion de référence des termes en vertu de ce lien causal. De plus, en reprenant la théorie de la vérité de Tarski qui conçoit la notion de vérité comme étant réductible à la simple de référence des termes, Field déduit qu'il est donc possible de naturaliser la notion de vérité. Réduire la notion de vérité en des termes causaux est d'autant plus important que cela permettrait de faire une grande avancée pour un projet plus important encore, celui de naturaliser l'esprit via des termes physiques et computationnels. Putnam, lui, n'ayant jamais envisagé que l'externalisme sémantique puisse être utilisé pour naturaliser l'esprit humain, remit en question ses positions philosophiques pour adopter le réalisme interne.

Le réalisme interne de Putnam n'est pas une conséquence de l'externalisme sémantique. En réalité, il tire son origine d'un argument bien particulier qui n'a pas à priori pas de rapport avec l'externalisme sémantique. Cet argument s'appelle l'argument « modèle-théorique ». Il défend l'idée qu'il est impossible d'avoir un point de vue objectif, un point de vue de dieu. Selon cet argument, il n'y a pas de moyen de trouver la théorie idéale parmi des théories scientifiques concurrentes :

> Si on admet, à la suite de Quine, que notre système de croyances est sous-déterminé par l'expérience, nous devons habituellement avoir recours à des critères relativement

> indépendants de l'expérience pour sélectionner une interprétation parmi une diversité d'interprétations éligibles pour un même ensemble de données expérimentales.[1]

Le choix de la théorie se fait alors selon des arguments normatifs, comme la simplicité ou la beauté esthétique de celle-ci, ce que Putnam nomme les « critères d'acceptabilité rationnels ». Et, pour Putnam, nous n'avons pas de moyens de savoir si nous avons raison de choisir cette théorie plutôt qu'une autre, hormis en invoquant d'autres critères d'acceptabilité rationnels. En somme, nous ne pouvons pas savoir si une théorie est vraiment la bonne.

Mais l'argument va plus loin. En effet, Putnam s'inspire également du théorème de Löwenheim-Skolem afin d'établir que

> […] les conditions de vérité des phrases de notre langage (lesquels nous sont donnés par les contraintes opérationnelles et théoriques) sous-déterminent la référence des termes qui les constituent.[2]

Ainsi, pour Putnam, les conditions de vérité des termes ou expressions, c'est-à-dire le fait que l'usage du terme ou de l'expression soit correct dans une situation donnée, sous-déterminent les références à ces termes ou expressions. C'est en lien direct avec la sous-détermination de la référence mise en avant par Quine dans le chapitre 2 de son ouvrage *Le mot et la chose*[3]. En effet, dans un exercice de traduction radicale comme le fait de traduire le mot « gavagai », Quine a montré que l'observation ne nous permet pas de dire si le peuple indigène parle d'un « lapin » ou d'une « tranche de lapin » ou du concept de la « lapinité ». Toutes ces traductions sont possibles et toutes sont correctes. Nous pouvons utiliser chacune d'entre elles sans que cela n'affecte notre compréhension ou la vérité de la phrase ainsi traduite. Mais, pour Putnam, cette sous-détermination est présente pour toutes les expressions utilisées par les individus même s'ils font partie de la même communauté linguistique. Il entreprend de montrer cela dans son ouvrage *Raison, vérité et histoire*[4] avec l'exemple de l'expression « un chat est sur le paillasson ». Il montre, en effet, qu'il pourrait y avoir un dialogue entre des personnes pour qui l'expression veut effectivement dire qu'il y a un chat sur le paillasson et des personnes pour qui l'expression signifie qu'une cerise est dans un arbre sans que ces personnes ne se rendent compte qu'elles ne parlent pas de la même chose. Pour

[1] Rochefort, « Putnam (A) ».
[2] Rochefort.
[3] Willard Van Orman Quine, *Le mot et la chose*, trad. par Joseph Dopp et Paul Gochet, Champs. Essais (Paris: Flammarion, 2010).
[4] Putnam, *Raison, vérité et histoire*.

Putnam, il est toujours possible de modifier la structure du langage de manière à ce que personne ne puisse discerner l'interprétation que l'individu a d'une expression ou d'un terme.

Dès lors, forts de ces arguments modèle-théorique, comment savoir si nous parlons des mêmes choses ? Comment savoir si une théorie est objectivement celle qui correspond à la réalité lorsque l'on ne sait pas la départager d'autres théories concurrentes hormis en invoquant des arguments normatifs ? Comment parler d'un point de vue objectif lorsque nous sommes résolument isolés au sein même de notre propre langage, de notre réseau conceptuel ? Cet argument est fondamental pour Putnam dans son passage à un réalisme interne. Accepter l'argument amène inéluctablement à considérer qu'il ne peut y avoir de point de vue objectif surplombant et qu'il ne peut pas y avoir une seule et unique description vraie du monde. Chacun est enfermé dans ses stéréotypes, avec son propre réseau conceptuel et, quand bien nous parvenons à communiquer avec les autres et à dire des choses vraies, nous ne savons pas si nous faisons référence exactement aux mêmes objets du monde. Il y a une vérité mais celle-ci relève d'une cohérence interne au sein de notre réseau conceptuel.

Cependant, pour que cet argument devienne crédible, pour que celui-ci sorte de son aspect purement théorique, il faut accepter une conception holistique de nos croyances et connaissances ainsi que l'externalisme sémantique. Cela signifie donc que le réalisme interne, s'il est bien compris, implique le holisme sémantique et l'externalisme sémantique. En effet, le holisme sémantique fournit une conception de l'esprit selon laquelle tous nos concepts sont interconnectés et dépendent les uns des autres. De plus, nos théories sont sous-déterminées par les expériences empiriques. Une telle conception de l'esprit permet donc de rendre compte de cette notion de vérité en tant que cohérence interne au sein d'un réseau de croyances car les individus n'acquièrent pas tous les mêmes concepts ni dans les mêmes conditions. Le fait que nos réseaux conceptuels sont sous-déterminés par l'expérience permet, quant à lui, de rendre compte de cette impossibilité de trouver la théorie idéale. Les liens conceptuels ne sont pas les mêmes d'un individu à l'autre ce qui a pour conséquence que chaque individu dispose d'un réseau conceptuel, d'un réseau de croyances et de connaissances qui lui est propre. Cela est d'autant plus pertinent si, comme Putnam, l'on considère pour vrai l'externalisme sémantique qui, rappelons-le, nous dit que l'environnement fixe au moins en partie la signification de nos termes. Notre réseau conceptuel varie donc en fonction de notre environnement et de nos expériences vécues. Qui plus est, bien que l'externalisme sémantique nous dise que l'environnement a un impact causal sur les significations, il dit aussi que

les individus ont des stéréotypes qui vont varier d'un individu à l'autre en fonction de leur réseau conceptuel propre. Il y a donc une forme d'indépendance du stéréotype à l'égard du monde et c'est cela qui peut nous amener à considérer que les valeurs de vérité des expressions sous-déterminent les références. Si le lien causal qu'entretient l'environnement avec les significations nous empêchait d'avoir des stéréotypes différents, en somme s'il permettait à nous tous d'avoir les mêmes significations dans nos têtes, alors les conditions de vérité des expressions détermineraient les références. Mais ce n'est pas le cas. Nous avons des réseaux conceptuels différents, nous ne voyons ni ne comprenons le monde exactement de la même manière et, pourtant, nous parvenons à communiquer entre nous. Cela laisse donc supposer qu'il est possible d'entretenir un discours vrai et cohérent sans pour autant avoir les mêmes références que les autres. Enfin, l'externalisme sémantique introduit par Putnam nous permet de ne pas tomber dans une forme de solipsisme, car il affirme notre appartenance à un monde puisqu'il y a un lien causal entre la signification de nos termes et ce dernier. Cela empêche donc toute conclusion radicale selon laquelle il n'y aurait que notre esprit et rien d'autre.

Si nous considérons le réalisme interne pour vrai, alors nous devons accepter l'externalisme sémantique et le holisme sémantique. Le réalisme interne implique de concevoir l'esprit comme un réseau de concepts qui dépend de l'environnement naturel et social de chaque individu. Les individus ont chacun leurs propres manières de voir le monde et il n'est pas possible de savoir si ces visions concordent entre elles. Il faut donc admettre que la vérité n'est qu'une forme de cohérence rationnelle au sein du réseau de croyances et de connaissances de chacun. En somme, il n'y a pas de point de vue de dieu. Plusieurs descriptions vraies du monde sont possibles et peuvent cohabiter sans que cela ne pose de problème dans les rapports et la communication entre les humains. Un anglais n'a pas la même description du monde qu'un brésilien étant donné qu'ils se trouvent dans deux environnements différents, même si cette différence est infime. Cependant, leurs descriptions respectives du monde sont vraies car il y a une cohérence interne au sein de leur réseau conceptuel respectif. La notion de vérité n'est plus une correspondance avec une réalité objective extérieure à l'individu mais est le fruit d'une cohérence entre nos concepts et avec les expériences que nous faisons. Chacun a une vision du monde qui lui est propre et personne ne peut nous dire qu'elle est fausse, si ce n'est Dieu.

Ainsi, le réalisme interne, pour ne pas simplement être une conception purement abstraite, implique le holisme sémantique et l'externalisme sémantique. Et, étant donné que ces deux thèses

sont utilisées par Putnam pour rejeter le computationnalisme, nous pouvons également conclure que le réalisme interne est antagoniste au computationnalisme.

2. Pourquoi le fonctionnalisme computationnel est-il antagoniste au réalisme interne ?

Le fonctionnalisme postule la thèse de la réalisation multiple selon laquelle un même état computationnel peut se retrouver dans des organismes, des systèmes différents. Cet argument donne du poids au fonctionnalisme car il permet notamment de rendre compte du fait que des organismes, aussi différents soient-ils, soient capables d'avoir des états mentaux similaires comme la douleur par exemple. Or, avec le réalisme interne, cette thèse ne peut plus être soutenue. En effet, si l'on combine l'externalisme sémantique et le holisme sémantique, il paraît alors difficile d'affirmer que des espèces différentes puissent avoir un même état mental en raison de leurs différences d'environnement, de relations sociales et de constitutions anatomiques. Pour Putnam, cela va même plus loin puisque, selon lui, il ne pas y avoir un même état mental chez plusieurs individus de la même espèce et cela pour les mêmes raisons :

> Le nombre de neurones que compte votre cerveau n'est plus exactement le même que le nombre de neurones que compte celui de n'importe qui d'autre, et les neurologues nous disent qu'il n'y a pas deux cerveaux qui soient « câblés » de la même manière. Le « câblage » dépend de l'histoire de la maturation et de la stimulation provoquées par l'environnement sur le cerveau de l'individu.[1]

Cela va donc à l'encontre d'un fonctionnalisme qui, selon Putnam, est antérieur à toutes les autres formes de fonctionnalisme computationnel. Ce fonctionnalisme se base de manière stricte sur le formalisme des machines de Turing. Il postule ainsi qu'il ne peut y avoir qu'un état computationnel par attitude propositionnelle. Ce fonctionnalisme ne fut jamais pris au sérieux, y compris par Putnam lui-même. Cependant, toutes les autres formes de fonctionnalisme computationnel se sont développées en partie par rapport à ce fonctionnalisme-là. Donc, pour Putnam, réfuter ce fonctionnalisme constitue un argument pour rejeter tous les autres. Ainsi, d'après ce fonctionnalisme hypothétiquement plus originaire que le fonctionnalisme originaire putnamien, chaque espèce à des tables de Turing certes différentes mais pouvant avoir des états fonctionnels et donc des états mentaux identiques. Nous devons donc conclure que ce

[1] Putnam, *Représentation et réalité*, page 141-142.

fonctionnalisme computationnel est trop restreint puisqu'il paraît impliquer qu'il ne peut y avoir qu'un unique état fonctionnel et mental associé à la douleur. Or, selon le réalisme interne, chaque individu a des états mentaux uniques, ce qui complexifie toute explication de la présence de phénomènes similaires comme la douleur chez de nombreux individus et espèces différentes. C'est ce que Ned Block appelle le « chauvinisme » dans son article « Troubles with functionalism »[1]. Dans cet article, Block présente la théorie physicaliste de l'esprit comme étant une théorie critiquée par le fonctionnalisme. En effet, d'après lui, le fonctionnalisme reproche au physicalisme d'être

> […] une théorie chauviniste : elle refuse les propriétés mentales de systèmes qui les possèdent de fait. En disant par exemple que les états mentaux sont des états du cerveau, les physicalistes excluent injustement ces pauvres créatures sans cervelle qui ont néanmoins un esprit.[2]

Cependant, Block démontre que :

> L'argument même que le fonctionnalisme utilise pour condamner le physicalisme peut être tout aussi bien appliqué contre le fonctionnalisme : en effet, toute version du fonctionnalisme qui évite le libéralisme tombe, comme le physicalisme, dans le chauvinisme.[3]

Pour Block, le fonctionnalisme, peu importe la forme qu'il prend, finit soit par être libéral, c'est-à-dire qu'il ne discrimine pas assez les entités qui peuvent avoir des états fonctionnels et des états mentaux, ou par être chauvin, c'est-à-dire qu'elle discrimine trop les entités qui peuvent avoir des états fonctionnels et des états mentaux. Pour lui, aucune forme de fonctionnalisme ne peut esquiver à la fois le libéralisme et le chauvinisme.

Dans ce cas, ce fonctionnalisme est trop chauviniste car si chaque individu a des états mentaux différents, il devient difficile de postuler que, nous, ainsi que de nombreux êtres vivants, puissent tous ressentir de la douleur par exemple. Cette théorie fonctionnaliste postule qu'il ne peut y avoir qu'un seul état mental associé à la douleur et le réalisme interne, lui, postule que cet état ne peut se trouver théoriquement que chez un unique individu. Et, quand bien même tous les humains, malgré leur différence de vécu et d'environnement, auraient le même état mental associé à la

[1] Ned Block, « Troubles with Functionalism », *Minnesota Studies in the Philosophy of Science* 9 (1978): 261-325.
[2] Block, page 265 : « […] a chauvinist theory: it withholds mental properties from systems that in fact have them. In saying mental states are brain states, for example, physicialists unfairly exclude those poor brainless creatures who nonetheless have minds ».
[3] Block, page 265 : « the very argument which functionalism uses to condemn physicalism can be applied equally well against functionalism; indeed, any version of functionalism that avoids liberalism falls, like physicalism into chauvinism ».

douleur, il est impossible de pouvoir affirmer que d'autres animaux puissent l'avoir aussi. Les différences de vécu et d'environnement sont trop importantes.

Une solution serait alors de postuler que des états mentaux différents puissent être du même type. Ainsi, d'après Putnam, si l'on veut décrire un état mental en des termes fonctionnalistes, il faudrait dresser une liste disjonctive de tous les états computationnels qui seraient coextensifs présents chez chaque individu sur Terre. Mais cela pose deux problèmes. Premièrement, dresser une telle liste ne ferait que faire une description d'états qui sont coextensifs. Pour que cela soit une réduction il faudrait, selon Putnam, montrer que l'état fonctionnel obéit aux mêmes lois que l'état mental et que les effets expliqués par l'état mental soient également expliqués par l'état fonctionnel. Or, la liste disjonctive ne permet pas cela. Elle nous fournit une description mais pas une explication. Deuxièmement, une telle liste s'avèrerait être théoriquement infinie et donc hors de notre portée. Ainsi, pour Putnam :

> Les êtres sentants physiquement possibles se présentent tout simplement sous des « configurations » trop nombreuses, physiquement et computationnellement parlant, pour que quelque chose comme un fonctionnalisme de la forme « un état computationnel par attitude propositionnelle » soit vrai.[1]

Ce recours à une liste disjonctive fut aussi une solution apportée par les physicalistes pour défendre leur théorie. Cela amène alors Block à notifier que :

> […] ce ne sont là que les motifs pour lesquels les fonctionnalistes rejettent généralement acerbement les théories disjonctives parfois avancées par des physicalistes désespérés. Si les fonctionnalistes sourient soudainement sur des états radicalement disjonctifs pour se sauver du chauvinisme, ils n'auront plus aucun moyen de se défendre du physicalisme.[2]

Ainsi, comme nous pouvons le voir, non seulement cette solution ne fonctionne pas selon Putnam dès lors que l'on accepte le cadre réaliste interne mais, quand bien même elle fonctionnerait, Block estime que ce n'est, de toute manière, pas une bonne manière de sauver le fonctionnalisme.

Un autre argument de taille contre le fonctionnalisme computationnel provient du problème de l'équivalence entre les termes et donc entre les états mentaux. Avec l'externalisme sémantique et le holisme sémantique, nous avons vu qu'il est contestable de stipuler que des individus puissent

[1] Putnam, *Représentation et réalité*, page 145.
[2] Block, « Troubles with Functionalism », page 316 : « […] these are just the grounds on which functionalists typically acerbically reject the disjunctive theories sometimes advanced by desperate physicalists. If functionalists suddenly smile on wildly disjunctive states to save themselves from chauvinism, they will have no way of defending themselves from physicalism ».

avoir le même état mental. Maintenant, il s'agit de se pencher sur la possibilité de savoir si des termes sont synonymes dans l'exercice d'interprétation d'un langage ou si des théories scientifiques sont coréférentielles. S'il était possible de définir la relation de synonymie entre des termes de langage différents, il serait alors possible de stipuler qu'il y a des équivalences entre des états mentaux, équivalence qui serait alors elle aussi réduite en des termes computationnels. Si des termes sont équivalents (ou synonymes), cela veut dire qu'ils ont la même signification et la même référence malgré leurs différents environnements respectifs. Pour les théories scientifiques, dire qu'elles sont coréférentielles implique qu'elles prétendent expliquer le même phénomène et qu'elles sont respectivement vraies dans leurs environnements respectifs. Ainsi, selon Putnam, si jamais nous voulions interpréter un langage en étant computationnaliste, il faudrait être en mesure de posséder des théories ou des schémas d'inférences que les personnes de la communauté linguistique (celle dont on cherche à interpréter des énoncés) ont en commun. Nous pourrions alors réduire le principe de synonymie en une relation calculable ou computationnellement définissable. Cela revient à dire que pour savoir si deux termes sont synonymes ou si deux théories sont coréférentielles afin de pouvoir réduire cette relation à quelque chose de définissable computationnellement, il faut être capable de décrire toutes les croyances et toutes les spécificités de l'environnement dans lequel est apparu chaque terme ou chaque théorie. Mais, étant donné que chaque croyance est liée à d'autres croyances au sein d'un réseau, il s'avère que, pour décrire une croyance (comme une définition de terme ou une information sur notre environnement), nous avons besoin de décrire les croyances qui justifient la première et les croyances qui justifient les secondes, etc. Nous tombons alors dans une immense boucle de justification. Il y a trop de choses à définir et à réduire car la vision du monde propre à chacun est constituée d'une infinité de relations entre des éléments, ce qui rend la tâche de description impossible. Ainsi, pour Putnam :

> […] il est faux de supposer qu'en principe on peut dire ce qu'un terme désigne dans un environnement à partir d'une description suffisamment complète de cet environnement, faite en fonction d'un ensemble standardisé de paramètres physiques et computationnels, à moins d'élargir la notion d'environnement du locuteur de manière à inclure tout l'univers physique.[1]

Pour ce qui est des théories physiques, il faudrait être en mesure de connaître toutes les théories possibles, ce qui semble impossible également. En réalité, il faudrait être omniscient comme un dieu, ce qui est impossible. L'argument de Putnam consiste à dire qu'aucun humain ne peut

[1] Putnam, *Représentation et réalité*, page 149.

connaître toutes les théories possibles et encore moins être capable de décrire l'univers entier. Or, si visiblement l'être humain n'a pas recours à une telle méthode dans la pratique, alors pourquoi, d'après le fonctionnalisme computationnel, l'esprit le ferait ? Et pourquoi une IA devrait le faire si elle veut se rapprocher d'une intelligence humaine ? Il semble que ce qu'implique le fonctionnalisme soit disproportionné avec les capacités réelles du cerveau humain. Ce fonctionnalisme considère que chaque individu possède en lui, l'entièreté des significations, l'entièreté des états mentaux, ce qui amène au cas de figure ridicule ou la connaissance de l'univers entier devrait être présente en chacun. Grâce à l'externalisme sémantique, le réalisme interne montre que nous ne sommes pas complètement maîtres de nos significations et de nos états mentaux. Il y a un rapport et une présence au monde qui était oblitérée par le computationnalisme originel.

Enfin, rappelons-nous, le réalisme naturel de Putnam défend l'idée qu'il n'y a pas d'interface entre notre esprit et le monde. Les notions de représentation et de *sense-data* ne sont pas totalement intelligibles selon lui. Un des arguments principaux qu'avance Putnam contre le fait d'utiliser une interface entre nous et le monde est qu'il n'y a pas de formalisme, d'explication claire de ce qu'est cette interface, de ce qu'elle apporte et de comment ça marche. Or, ce genre d'argument, il le fait aussi à l'encontre du computationnalisme dans *Représentation et Réalité* puisqu'il affirme qu'en principe, le fonctionnalisme est censé se baser sur le formalisme des machines de Turing tout en constatant que, factuellement, aucune forme de fonctionnalisme ne s'est réellement basée sur ce formalisme. Comme nous l'avons vu, un fonctionnalisme qui se base sur le formalisme des machines de Turing doit postuler qu'il ne peut y avoir qu'un état computationnel par attitude propositionnelle. Or, un tel fonctionnalisme ne peut pas marcher et il a donc été rapidement abandonné par les théoriciens de l'esprit. Ce qu'a théorisé Putnam dans « La nature des états mentaux » est déjà une forme de fonctionnalisme qui ne respecte pas vraiment ce formalisme puisqu'il ne stipule pas qu'il n'y a qu'un état computationnel par attitude propositionnelle. Au contraire, il stipule qu'il peut y avoir plusieurs états computationnels par attitude propositionnelle. On a donc un formalisme computationnel spécifique qui n'est pas vraiment défini.

Toutes les théories fonctionnalistes se basent alors soit sur un formalisme computationnel qui n'a jamais vraiment été défini, soit sur des théories psychologiques. Et se baser sur des théories psychologiques pose également problème pour Putnam car, pour savoir si une théorie psychologique est « réalisée », il suffit d'observer si le comportement de l'individu est conforme

aux prédictions de la théorie. Nous tombons alors dans une forme de béhaviorisme, ce qui cause un problème de libéralisme, à ne pas comprendre dans le sens du libéralisme dont parle Block[1]. Il ne s'agit pas ici de considérer qu'il y a trop de choses qui auraient des états mentaux. Il s'agit plutôt de dire que toute théorie psychologique qui a le type de structure des automates probabilistes et qui prédira le comportement d'un individu sera réalisée. Il y aura donc trop de réalisations possibles, trop de théories qui pourraient expliquer le fonctionnement de l'esprit sans que nous puissions choisir laquelle est la bonne. En somme, il y a trop de théories psychologiques possibles. Finalement, que ce soit via le formalisme computationnel ou via une théorie psychologique, nous ne savons pas vraiment de quoi nous parlons. Pour Putnam, « jusqu'à présent, nous n'avons que les descriptions les plus vagues de ce qu'est supposé être le formalisme computationnel »[2]. Or, cela l'amène à conclure que « sans un formalisme computationnel, la notion d'"état computationnel" n'a aucun sens »[3]. C'est une des critiques principales du Putnam du réalisme interne à l'encontre du computationnalisme. Pour lui, nous ne savons pas vraiment de quoi nous parlons quand nous disons que les individus fonctionnent selon le modèle computationnaliste. Cette théorie est dépourvue de sens.

Ainsi, les théories fonctionnalistes computationnelles peuvent être considérées comme soit trop chauvines, soit trop libérales. De plus, pour le Putnam du réalisme interne, la réduction en des termes computationnels de l'équivalence entre les états mentaux demanderait d'être en mesure de décrire l'univers visible. Il a donc trop de choses à prendre en compte, à réduire en des termes computationnels et l'on peut légitimement douter que le cerveau soit capable d'accomplir un tel exploit. Mais, surtout, il s'avère que toutes les théories fonctionnalistes computationnelles ne se basent sur aucun formalisme clairement défini. Pour Putnam, nous ne savons pas vraiment de quoi nous parlons lorsque nous discutons à propos du computationnalisme.

[1] Block, « Troubles with Functionalism ».
[2] Putnam, *Représentation et réalité*, page 144.
[3] Putnam, page 144.

3. Transposition des arguments au cadre du réalisme naturel

Le réalisme naturel, en tant qu'évolution du réalisme interne, ne rompt pas les liens avec les thèses de l'externalisme sémantique et du holisme sémantique. Il les implique lui aussi. Or, les arguments développés par le Putnam du réalisme interne à l'encontre du fonctionnalisme se basent notamment sur l'externalisme sémantique et le holisme sémantique. Il s'avère que donc que ces arguments restent toujours cohérents dans le cadre du réalisme naturel. Ainsi, pour un réaliste naturel aussi, le fonctionnalisme computationnel ne peut pas être une théorie de l'esprit valable. L'argument du manque de sens des théories fonctionnalistes est lui aussi toujours valable dans ce réalisme. Nous pouvons donc maintenir la conclusion que les cerveaux et les corps des humains ainsi que des animaux ne fonctionnent pas de la sorte.

De par cette continuité entre réalisme interne et réalisme naturel, il y a des phénomènes mis en avant par le réalisme interne qui peuvent être expliqués par le modèle de l'esprit de McDowell et Putnam, contrairement au fonctionnalisme. Il y a par exemple un aspect normatif dans la signification des termes notamment dans le cas de la synonymie. Il y a l'estimation de la simplicité de la théorie ou d'autres outils comme le « principe de charité » qu'on appelle aussi le « bénéfice du doute », etc. Ainsi, nous connaissons bien plus de choses sur les plantes que les individus du XVII^{ème} siècle et pourtant nous considérons que leur mot « plante » signifie la même chose que notre mot « plante » aujourd'hui. On utilise la charité dans les interprétations, on laisse de côté, quand on interprète, certaines différences de croyances. L'interprétation se plie donc à des procédures normatives. Pour Putnam, les concepts, comme le concept de plante, ne conservent pas une essence à travers le temps, ils conservent une identité. En effet, selon lui :

> […] (à moins d'être philosophe ou historien des sciences doté d'un tour d'esprit philosophique), nous ne disons pas que les gens, il y a deux cents ans, « vivaient dans un monde différent » ou que leurs notions sont « incommensurables » avec les notions que nous avons aujourd'hui : littéralement (ce qui bien sûr n'est jamais le cas !) cela voudrait dire en effet que nous ne pourrions pas interpréter une simple lettre écrite par quelqu'un il y a deux cents ans. Bref, pour nous, le concept de plante a une identité à travers le temps mais pas d'essence.[1]

Or, le fonctionnalisme ne peut pas vraiment rendre compte de ces aspects normatifs. Au contraire, selon le fonctionnalisme, le terme plante devrait mobiliser un même état mental peu

[1] Putnam, page 40-41.

importe l'individu et l'époque. Il semble difficile de pouvoir expliquer de manière computationnelle le fait que l'on ait regroupé sous le même terme des définitions et des concepts totalement différents. Ces aspects normatifs sont le fruit de l'environnement social de l'individu. C'est au sein d'un groupe, d'une société, que l'on décide de faire preuve de charité envers les concepts du passé ou que l'on choisit une théorie scientifique parmi plusieurs choix car elle est plus simple à manipuler ou plus élégante esthétiquement. Ce sont des normes et des conventions sociales qui ne sont pas déterminées fonctionnellement au sein de chaque système individuel. Dans le modèle de McDowell, l'aspect social dans le développement de nos capacités mentales formant l'esprit est primordial. Vivre dans le monde, dans un groupe et se développer au sein de celui-ci permet de comprendre que l'on puisse acquérir ces outils normatifs.

Un autre élément important selon Putnam est l'imagination. Pour Putnam, c'est l'imagination qui nous permet d'excéder notre réseau de croyances, de comprendre le monde. Par exemple, comprendre la phrase « le Soleil est à 150 millions de kilomètres de la Terre » revient à faire des « sauts conceptuels ». Faire des sauts conceptuels c'est « se projeter en imagination dans de nouvelles manières de penser »[1]. C'est le seul moyen possible pour saisir de tels énoncés sans avoir à modéliser l'univers entier, tel un dieu. Pour Putnam, ce sont avant tout des capacités propres à l'être humain qui le rendent capable d'apprendre de nouvelles choses et d'interpréter des croyances. L'imagination est la capacité à dépasser notre réseau de croyances interne pour en former de nouvelles ou pour apporter des significations qui n'auraient pas pu être trouvées via des expériences sensibles ou des raisonnements logiques. L'imagination est surement elle aussi éduquée socialement. Elle est une conséquence de notre présence au sein d'une société dans le monde et de notre ouverture à celui-ci. Or, pour le fonctionnalisme, si l'on n'accepte pas la tâche prométhéenne de décrire l'univers, alors l'imagination semble être une capacité irréductible. Il semble impossible de rendre compte du fait de pouvoir excéder notre réseau de croyances internes pour en créer de nouvelles.

Enfin, il y a le problème des qualia. C'est un problème que Putnam aborde dans son ouvrage *Raison, vérité et histoire*. Par exemple, comment pouvons-nous savoir si Lucas a le même quale que Théo quand ils regardent tous les deux une tomate rouge ? Peut-être que Lucas a le quale du rouge quand il voit la tomate alors que Théo, lui, a le quale du vert. Ce que Théo désigne comme étant rouge, il le voit en réalité comme étant vert. Et inversement, quand les deux regardent de

[1] Putnam, page 171.

l'herbe, Lucas a le quale du vert alors que Théo a le quale du rouge. En réalité, nous ne pouvons pas le savoir. De plus, le fonctionnalisme ne peut pas décrire les qualia en des termes computationnels car les qualia ne peuvent pas être des états fonctionnels étant donné qu'ils n'ont pas de rôles fonctionnels. Si les qualia avaient des rôles fonctionnels, Lucas et Théo seraient dans des états mentaux différents quand ils regardent une tomate puisque que cela voudra dire qu'il y a un *input* qui diffère chez chacun d'eux. Or, ce n'est pas le cas. Les états mentaux de chaque individu se trouvant devant une tomate ont les mêmes causes (*inputs*) et ont les mêmes effets (*outputs*), comme, par exemple, le fait de dire « cette tomate est rouge ». Étant donné ces similarités, il paraît impossible de distinguer les états mentaux des individus. Il faut donc conclure que les qualia n'ont pas de rôles fonctionnels étant donné que ces derniers, les ressentis qualitatifs, peuvent différer sans que cela ne change quoi que ce soit dans la communication, dans les réactions des individus. Pour Putnam, « si le rôle fonctionnel était identique au caractère qualitatif, alors on ne pourrait pas dire que la qualité de la sensation a changé »[1]. Si les qualia jouaient un rôle fonctionnel alors nous serions obligés de dire soit que Lucas et Théo n'ont pas la même sensation quand ils voient une tomate car le ressenti qualitatif serait alors déterminant pour causer l'état mental approprié, ce qui les empêcherait de dire tous les deux « cette tomate est rouge », soit que la sensation, le quale, qu'ont Lucas et Théo est identique quand ils voient une tomate rouge. Mais, dans la réalité, nous ne pouvons affirmer ni l'une ni l'autre des possibilités. Le fait de ne pas parvenir à rendre compte des qualia constitue donc un problème majeur pour le computationnalisme.

Les qualia ont un statut particulier. Ce sont des ressentis qualitatifs privés et ineffables. Cela pose un problème de taille pour toute théorie cherchant à réduire en des termes scientifiques, dans des termes du monde des lois naturelles, ces ressentis. Dire que voir du rouge active une certaine zone du cerveau ne nous indique pas pourquoi le rouge est comme il est et pourquoi c'est ce ressenti qualitatif et pas un autre. Il y a donc un manque, un gap, entre l'explication physique, scientifique, et le ressenti subjectif. Il y a un gouffre explicatif.

Le naturalisme modéré de McDowell ne semble pas non plus pouvoir donner l'explication de pourquoi le rouge est tel qu'il est. Cependant, tout le propos de ce naturalisme est de dire qu'il n'est pas possible de réduire en des termes de lois naturelles de telles capacités de l'esprit. Le modèle de l'esprit de McDowell peut nous amener à dire que les qualia sont le fruit de capacités extra conceptuelles issues de l'évolution mais, hormis cette origine supposée, il n'y a pas d'explications

[1] Putnam, *Raison, vérité et histoire*, page 95.

scientifiques de ce que sont les qualia. Qui plus est, l'affirmation de l'absence d'interface, de *sense -data*, c'est-à-dire l'affirmation d'un rapport direct au monde nous permet d'affirmer que des expériences de pensées comme celles du spectre de vision inversée ne peuvent pas être réelles. Ce rapport direct au monde, le fait que nous soyons des animaux dans un environnement, faisant partie d'une espèce commune qui s'est développé au fil des millénaires, nous permet de faire la supposition pragmatique que Lucas et Théo ont tous les deux le quale du rouge quand ils voient une tomate.

4. Pour conclure

Le réalisme naturel reprend les mêmes arguments qui vont à l'encontre du computationnalisme originaire développé par Putnam alors qu'il était un réaliste interne. C'est donc un réalisme qui est lui aussi antagoniste à cette théorie fonctionnaliste de l'esprit. Finalement, le computationnalisme initial de Putnam décrit les systèmes, les individus, comme étant isolés du monde. Ils ne sont que des machines, fonctionnant avec leurs tables de Turing qui leur sont propres et n'interagissant avec le monde que par input et output. Notre contact avec l'extérieur se résume à des *sense-data*, à une relation indirecte et cela invisibilise totalement le fait que nous faisons partie du monde et que nous vivons en son sein. Nous ne sommes pas isolés, ce genre de conception solipsiste n'est pas tenable. Savoir que nous faisons partie du monde avec d'autres individus nous permet de rendre compte de la nature de notre esprit. Par le développement de ces capacités, nous pouvons sortir de notre environnement proche, nous projeter dans le monde afin de comprendre et de connaître les choses. C'est ce qui nous différencie des animaux et qui rend le computationnalisme insuffisant pour fournir une description correcte de l'esprit. L'esprit est un regroupement de capacités qui proviennent du cerveau et tirent, certes, leur origine de la nature, du monde des lois naturelles, mais elles ne sont pas réductibles à ce dernier. Putnam a montré que des capacités comme l'imagination ou l'intention ne sont pas réductibles en des termes computationnels. Les capacités de l'esprit humain font partie d'un autre espace, l'espace logique des raisons, qui est hors d'atteinte pour les sciences physiques quand bien même ces capacités se forment dans le cerveau qui, lui, fait partie de l'espace des lois de la nature. Nous comprenons mieux en quoi McDowell a exercé une influence qui n'est pas des moindres sur le dernier Putnam. Il fait partie de la mouvance qui apporte une réponse convaincante aux yeux de Putnam au problème de la communication entre l'esprit et le

monde. Il apporte un naturalisme modéré qui s'accorde à un réalisme naturel, réalisme qui, lui, implique l'externalisme sémantique et le holisme sémantique. Le modèle de l'esprit de McDowell s'inscrit donc dans un rejet du fonctionnalisme computationnel initié par Putnam et ouvre une troisième voie pour sortir du débat réaliste/idéaliste qui dure depuis des siècles.

Si le modèle de l'esprit de McDowell est antagoniste au fonctionnalisme computationnel tel qu'il fut présenté initialement par Putnam, nous pouvons alors nous demander légitimement si l'Intelligence Artificielle est compatible avec un tel modèle. Est-ce que rejeter l'idée selon laquelle l'esprit fonctionne de manière analogue à un ordinateur amène à rejeter l'idée qu'un ordinateur puisse « reproduire » un esprit humain ? Pour savoir cela, tâchons maintenant d'étudier de plus près les deux courants principaux de la recherche en Intelligence Artificielle afin de mieux comprendre ce que l'on entend par IA.

Chapitre 4 : Les deux principaux genres de l'Intelligence Artificielle

Tout d'abord, qu'est-ce que l'Intelligence Artificielle ? Aurélie Jean considère que l'Intelligence Artificielle « regroupe les méthodes de calculs numériques, sur ordinateur, qui reproduisent un certain type d'intelligence : l'intelligence analytique »[1]. L'intelligence analytique c'est le type d'intelligence qui permet la résolution de problèmes ou l'analyse de données. C'est la forme d'intelligence à laquelle nous avons tendance à penser lorsque l'on emploie le mot « intelligent ». L'Intelligence Artificielle cible donc une forme précise d'intelligence, du moins pour l'instant.

De plus, avant d'être des programmes, des logiciels, des robots, des produits, l'Intelligence Artificielle est avant tout un programme de recherche. Bien que l'appellation d'Intelligence Artificielle n'y soit pas présente, il semble que ce soit Alan Turing qui ait lancé le programme de recherche en IA via son article « Computing machinery and Intelligence »[2]. Dans cet article, il pose la question « est-ce que les machines peuvent penser ? ». Cette question étant complexe ne serait-ce que par le fait qu'il n'est pas clairement défini ce que c'est que penser, Turing imagine alors un jeu qu'il appelle le jeu de l'imitation. Il décrit ce jeu comme ceci :

> Il se joue à trois : un homme (A), une femme (B) et un interrogateur (C) qui peut être de l'un ou l'autre sexe. L'interrogateur se trouve dans une pièce à part, séparé des deux autres. L'objet du jeu, pour l'interrogateur, est de déterminer lequel des deux est l'homme et lequel est la femme. Il les connaît sous les appellations X et Y et, à la fin du jeu, il doit déduire soit que « X est A et Y est B », soit que « X est B et Y est A ». L'interrogateur peut poser des questions à A et B de la manière suivante :
>
> C : X peut-il ou peut-elle me dire, s'il vous plaît, quelle est la longueur de ses cheveux ?
>
> A supposer à présent que X soit vraiment A, alors A doit répondre. La finalité du jeu pour A est d'essayer d'induire C en erreur. Sa réponse pourrait donc être :
>
> A : « Mes cheveux sont coupés à la garçonne et les mèches les plus longues ont à peu près vingt centimètres de long. »
>
> Pour que le ton de la voix ne puisse pas aider l'interrogateur, les réponses devraient être écrites ou, mieux, dactylographiées. L'installation idéale serait un téléimprimeur communiquant entre les deux pièces. A défaut, les questions et réponses peuvent être répétées

[1] Aurélie Jean, « Une brève introduction à l'intelligence artificielle », *médecine/sciences* 36, nº 11 (1 novembre 2020): 1059-67.
[2] Alan Mathison Turing, « Computing machinery and Intelligence », *Mind* LIX, nº 236 (1 octobre 1950): 433-60.

> par un intermédiaire. L'objet du jeu pour la joueuse (B) est d'aider l'interrogateur. La meilleure stratégie pour elle est probablement de donner des réponses vraies. Elle peut ajouter à ses réponses des choses telles que :
>
> « Je suis la femme, ne l'écoutez pas ! », mais cela ne servira à rien, car l'homme peut faire des remarques similaires.
>
> Nous posons maintenant la question : « Qu'arrive-t-il si une machine prend la place de A dans le jeu ? L'interrogateur se trompera-t-il aussi souvent que lorsque le jeu se déroule entre un homme et une femme ? » Ces questions remplacent la question originale : « Les machines peuvent-elles penser ? ».[1]

À l'époque, l'idée qu'un ordinateur réussisse le jeu, c'est-à-dire qu'il ne se fasse pas trouver par l'interrogateur, ne pouvait être que spéculatif étant donné la puissance de calcul des machines qui n'étaient alors qu'à leur balbutiement. Cependant, cela n'empêchait pas Turing d'être optimiste à l'idée que les ordinateurs finiront par gagner au jeu. En effet il dit :

> Examinons, en premier lieu, la question sous sa forme la plus précise. Je crois que dans une cinquantaine d'années il sera possible de programmer des ordinateurs, avec une capacité de mémoire d'à peu près 10^9, pour les faire si bien jouer au jeu de l'imitation qu'un interrogateur moyen n'aura pas plus de 70% de chances de procéder à l'identification exacte après cinq minutes d'interrogation.
>
> Je crois que la question originale « Les machines peuvent-elles penser ? » a trop peu de sens pour mériter une discussion. Néanmoins, je crois qu'à la fin du siècle l'usage, les mots et l'éducation de l'opinion générale auront tant changé que l'on pourra parler de machines pensantes sans s'attendre à être contredit. Je crois de plus qu'il ne sert à rien de dissimuler ces croyances. L'idée populaire selon laquelle les savants avancent inexorablement d'un fait bien établi à un autre, sans être influencés par des hypothèses non vérifiées, est absolument fausse. Pourvu que nous sachions clairement quels sont les faits prouvés et quelles sont les hypothèses, aucun mal ne peut en résulter. Les hypothèses sont de grande importance puisqu'elles suggèrent d'utiles voies de recherches.[2]

Affirmer cela ne se fait pas sans sous-entendu concernant le fonctionnement de l'esprit humain. Effectivement, Turing conçoit le cerveau humain comme un calculateur, comme quelque chose d'analogue à un ordinateur. Pour lui :

> Le lecteur doit accepter comme un fait établi que les ordinateurs peuvent être, et ont été, construits suivant les principes que nous avons décrits, et qu'ils peuvent en fait imiter de très près les actions d'un calculateur humain.[3]

[1] Alan Turing et Jean-Yves Girard, *La machine de Turing*, trad. par Julien Basch et Patrice Blanchard, Points Sciences (Points, 1999), page 135-136.
[2] Turing et Girard, page 148-149.
[3] Turing et Girard, page 142.

L’appellation de « calculateur humain » et l’idée qu’un ordinateur puisse l’imiter supposent donc que l’esprit fonctionne de manière computationnelle, avec une table, un algorithme qui régit son comportement. En tout cas, il est indéniable que Turing pensait que les machines seraient un jour capables d’égaler les hommes au moins dans les domaines purement intellectuels.

Comme nous l’avons dit précédemment, l’article de Turing a lancé un vaste programme de recherche sur l’Intelligence Artificielle. Cependant il faut tout de même préciser qu’officiellement,

> L'acte de naissance de l'Intelligence Artificielle (IA) correspond à un programme de rencontres organisées à Dartmouth College (Hanover, New Hampshire, USA) ayant réuni une dizaine de personnes pendant l'été 1956 […]. C'est à cette occasion que l'expression "Artificial Intelligence" (choisie par McCarthy) fut utilisée pour la première fois de manière systématique pour désigner le nouveau champ de recherche ; elle était cependant loin de faire l'unanimité parmi les chercheurs présents, certains ne voyant là que du traitement complexe d'informations.[1]

L’appellation « Intelligence Artificielle » n’allait pas de soi, même parmi les créateurs du programme de recherche. Tout était alors à prouver pour les défenseurs de la vision de Turing. Sur quoi se concentre ce programme de recherche alors ? Pour Daniel Andler, dans son ouvrage *Intelligence artificielle, intelligence humaine : la double énigme*[2], il travaille sur trois thèmes fondamentaux :

> Le premier concerne le quoi : en quoi consiste l'intelligence ? Quelle est au juste la capacité possédée par l'homme que l'on voudrait conférer aux ordinateurs ? Le deuxième et le troisième le comment : comment l'intelligence est-elle produite chez l'homme ? comment peut-elle être produite dans des systèmes artificiels tels que l'ordinateur ?[3]

Les réponses données par les pionniers du programme de recherche à ces questions étaient très proches de la pensée de Turing. En effet, ces réponses sont :

> (1) L'intelligence consiste en la capacité de résoudre des problèmes, d'accomplir des tâches.
>
> (2) L'intelligence est produite, chez l'humain, par son appareil cognitif, c'est-à-dire par son cerveau en tant que système de traitement de l'information.
>
> (3) Tout système de traitement de l'information, naturel ou artificiel, structurellement équivalent au système humain, possède de ce fait l'intelligence.

[1] GdRIA du CNRS, ouvrage coordonné par Sébastien Konieczny et Henri Prade, *L’intelligence artificielle. De quoi s’agit-il vraiment ?* (Cepadues, 2020), page 7.
[2] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, NRF Essais (Gallimard. Paris, 2023).
[3] Andler, page 24.

(4) En particulier, un ordinateur convenablement configuré pourrait donc produire de l'intelligence.[1]

Cela correspond encore à la définition donnée de l'IA aujourd'hui. Par exemple, comme nous l'avons vu dans l'article d'Aurélie Jean, l'IA se focalise sur l'intelligence analytique c'est-à-dire la capacité à résoudre des problèmes, à agir, à traiter des informations. Mais aussi et surtout, nous voyons en quoi le fonctionnalisme computationnel de Putnam théorisé en 1967 a pu fournir une conception de l'esprit en accord avec ces réponses et les objectifs voulus par Turing notamment parce que cette théorie s'inspire des travaux de ce dernier. Rappelons-le, le computationnalisme présente l'esprit comme analogue à un ordinateur. Chaque individu serait alors une machine programmée selon une table précise. Comprendre le fonctionnement de l'esprit reviendrait alors à trouver cette table, trouver cet algorithme qui nous anime.

Il est donc compréhensible que le computationnalisme se soit développé conjointement au programme de recherche en Intelligence Artificielle et il n'est alors pas si surprenant que, parallèlement aux difficultés qu'a traversé l'IA dans les années 1980 et 1990, le fonctionnalisme computationnel soit remis en question. En effet, le programme de recherche en IA a dû faire face à ce qu'Andler nomme des « renoncements ». Ceux-ci ont été progressifs mais on retrouve par exemple le renoncement à l'idée que l'IA comprenne les symboles qu'elle manipule, le renoncement à l'idée qu'une seule IA puisse reproduire toute l'intelligence en général, le renoncement à l'idée que l'IA soit à l'image de l'intelligence humaine ou encore le renoncement à l'idée que l'IA soit indépendante à l'égard de son créateur. La fin du XX^{ème} siècle a donc été un siècle de renoncement et de déceptions concernant la recherche en Intelligence Artificielle. Finalement, cette période de difficultés a provoqué un renoncement à une intelligence artificielle « forte » pour se focaliser sur une intelligence artificielle « faible ». Une IA forte c'est le but rêvé de Turing et des pionniers du programme de recherche. C'est une IA qui serait en mesure de comprendre ce qu'elle fait, ce qu'elle manipule, ce qu'elle communique. C'est une IA qui serait en mesure d'innover, de surprendre son créateur en apprenant et en réagissant différemment de ce qui était prévu à la base. L'IA faible quant à elle, c'est une IA qui manipule les symboles mais qui ne comprend pas ce qu'elle fait. Elle ne peut pas surpasser les connaissances que son créateur lui a insufflées, elle ne peut pas excéder sa condition initiale. Il y a un lien apparent entre IA faible et IA forte. L'IA forte est ce vers quoi l'IA faible cherche à tendre. Cela laisse donc supposer qu'il

[1] Andler, page 24-25.

est possible de passer d'une IA faible à une IA forte. Or, cela ne semble pas si simple. Comme le dit Aurélie Jean :

> Ce passage de l'IA faible à l'IA forte est régi par la théorie du point de singularité technologique. Celui-ci suppose l'existence, dans un futur même lointain, d'un point de basculement technologique, où les machines auront une intelligence générale équivalente ou supérieure à celle de l'humain… avec la conscience d'exister.[1]

Depuis le début du programme de recherche en Intelligence Artificielle un tel point de bascule ne semble jamais avoir été franchi. Pour l'instant nous n'avons donc pas encore eu d'IA forte. Mais diverses tentatives, diverses voies sont étudiées pour parvenir à ce but. Il y en a deux principales qui se démarquent : ce sont l'IA symbolique et l'IA connexionniste. Ces deux types d'IA sont présents depuis le tout début du programme de recherche. En effet, le programme de rencontres organisées à Dartmouth College en 1956 fut organisé

> […] à l'initiative de deux jeunes chercheurs qui, dans des registres différents, allaient fortement marquer le développement de la discipline : John McCarthy et Marvin Minsky, le premier défendant une vision purement logique de la représentation des connaissances, le second travaillant alors sur les neurones formels et les perceptrons, et qui privilégierait plus tard l'usage de représentations structurées (appelées en anglais "frames") de stéréotypes de situations pouvant inclure différents types d'information.[2]

Les deux voies sont là. Avec McCarthy, nous aurons l'IA symbolique et, avec Minsky, il s'agira de l'IA connexionniste.

1. L'IA symbolique ou la voie de la raison

Pourquoi IA « symbolique » ? Selon Andler, c'est parce qu'elle

> […] repose tout entière sur l'idée que l'intelligence consiste en la capacité de manipuler des symboles. Cela vaut pour l'ordinateur, mais aussi pour la pensée humaine : car la même question se pose à son propos.[3]

Cette idée ne vient pas de nulle part. Elle est issue d'une hypothèse fondatrice pour le programme de recherche en Intelligence Artificielle et en science cognitive. Cette hypothèse est la suivante :

---

[1] Jean, « Une brève introduction à l'intelligence artificielle ».
[2] GdRIA du CNRS, ouvrage coordonné par Sébastien Konieczny et Henri Prade, *L'intelligence artificielle. De quoi s'agit-il vraiment ?*, page 7.
[3] Andler, *Intelligence artificielle, intelligence humaine*, page 83.

> […] tout système cognitif, naturel ou artificiel, procède en effectuant des calculs sur des représentations. Dans le vocabulaire de la philosophie de l'esprit, c'est ce qu'on appelle la conception « computo-représentationnelle » de la cognition.[1]

Cette hypothèse que l'on nomme l'« hypothèse des systèmes symboliques physiques » ne vient pas non plus de nulle part. Elle a notamment été formulée par deux des principaux fondateurs du programme de recherche en IA, Herbert Simon et Allen Newell. Selon eux :

> Un système symbolique physique a les moyens nécessaires et suffisant pour l'action intelligente générale.
>
> Par « nécessaire », nous entendons que tout système faisant preuve d'intelligence générale s'avérera, après analyse, être un système de symboles physiques. Par « suffisant », nous entendons que tout système de symboles physiques d'une taille suffisante peut être organisé davantage pour faire preuve d'intelligence générale. Par « action intelligente générale », nous souhaitons indiquer la même portée de l'intelligence que celle que nous observons dans l'action humaine : dans toute situation réelle, un comportement approprié aux objectifs du système et adaptable aux exigences de l'environnement peut se produire, dans certaines limites de vitesse et de complexité.[2]

Pour eux, l'intelligence réside dans les systèmes symboliques physiques. Cela veut donc dire que notre cerveau est lui aussi un système symbolique physique. Or, si cela est vrai, si l'on admet que notre système cognitif effectue des opérations et des calculs sur des symboles, alors « on peut donc charger l'ordinateur de reproduire tout processus intellectuel, si complexe qu'il puisse être »[3]. Ainsi, tout ce qui est intelligent manipule des symboles. L'intelligence en elle-même consiste en la manipulation de symboles. Le respect de cette hypothèse n'est pas propre à l'IA symbolique. Cependant, c'est surement la voie qui tient le plus strictement cette hypothèse pour vraie.

En effet, c'est la voie consistant à développer des systèmes capables de telles manipulations symboliques, notamment via des algorithmes. L'IA symbolique est l'intelligence artificielle qui se « programme » qui requiert la simple utilisation de langage de programmation, un langage permettant d'écrire une suite d'instructions à base de « si … alors … », de « tant que … faire … », etc. Cette suite d'instructions, c'est un algorithme computationnel, une forme de table d'instruction

---

[1] Andler, page 69.

[2] Allen Newell et Herbert A. Simon, « Computer science as empirical inquiry: symbols and search », *Communications of the ACM* 19, nº 3 (1 mars 1976): 113-26, page 116 : « A physical symbol system has the necessary and suficient means for general intelligent action.
By "necessary" we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence. By "general intelligent action" we wish to indicate the same scope of intelligence as we see in humian action : that in any real situation behavior approprate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity ».

[3] Andler, *Intelligence artificielle, intelligence humaine*, page 85.

que l'ordinateur va alors suivre. Un élément important est que l'ordinateur manipule des symboles, certes, mais il ne comprend pas les chiffres ou les lettres qu'il manipule. Nous pouvons comprendre cela notamment à travers l'expérience de pensée de la chambre chinoise de John Searle, apparut dans son article « Minds, Brains and programs »[1]. Dans cette expérience, il imagine une personne enfermée dans une chambre et qui ne connaît pas le chinois. Il a cependant un manuel, lui permettant de répondre à des phrases en chinois par d'autres phrases en chinois. Le processus de réponse ne se base que sur la syntaxe des phrases et non leurs sens. Pour telle combinaison de symbole, la personne va répondre une autre combinaison précise. Il s'avère justement que cette personne reçoit des phrases en chinois et répond alors avec les instructions de son manuel. En réalité, ce sont des questions que sont envoyées à la personne enfermée dans la chambre et, du point de vue du locuteur qui envoie ces questions, la personne enfermée dans la chambre se comporte comme un individu qui parlerait vraiment chinois. Or, il ne comprend pas le chinois, il ne fait qu'appliquer des règles préétablies. Cette expérience de pensée tend donc à démontrer que l'IA symbolique n'est pas consciente du sens des symboles et ne comprend pas ce qu'elle manipule. En effet, comme la personne dans la chambre, elle ne fait que suivre des règles syntaxiques préalablement établies. L'ordinateur comprend seulement des marques auxquelles il réagit en fonction de ce que son programme lui dicte. Le symbole renvoie à quelque chose mais seul nous, les humains, pouvons comprendre ce que ces symboles désignent ou signifient. C'est l'un des renoncements qui ont été faits à l'égard de l'IA durant la seconde moitié du XX^{ème} siècle. Nous avons renoncé à l'idée que les IA comprennent ce qu'elles manipulent.

Aussi, pour Andler, il y a différents types de processus que l'on retrouve dans les cerveaux ou dans les ordinateurs. On distingue un premier niveau qui est inférentiel, conscient, logique, qui est celui du traitement de l'information et un second niveau qui est, lui, associatif, automatique et inconscient. L'IA se place donc sur le niveau fondamental. L'IA symbolique s'applique au premier type de processus. Ainsi :

> De manière générale, les règles doivent se conformer à la rationalité.
>
> La rationalité s'entend en deux sens principaux. Le premier est la conformité aux normes de la raison - celles de la logique comprise au sens large ; le second la capacité d'agir en vue d'atteindre ses objectifs. Le rapport entre les deux acceptions est celui de l'adéquation des moyens aux fins : pour se donner les meilleures chances d'atteindre ses objectifs, il est recommandé, toutes choses égales par ailleurs, de se conformer aux normes de la raison. Les processus rationnels sont aussi ceux qui optimisent les chances de l'agent de parvenir à ses

[1] John R. Searle, « Minds, brains, and programs », *Behavioral and Brain Sciences* 3, n° 3 (septembre 1980): 417-24.

> fins. L'IA vise l'utilité : c'est dans la mesure où elle sert à quelque chose que l'intelligence l'intéresse. Et être utile, c'est permettre au système de trouver des solutions aux problèmes qui se présentent ; ou encore de prendre, dans toute situation, la meilleure décision possible.
>
> Les algorithmes qui font passer d'un état au suivant — c'est-à-dire qui réalisent une inférence — doivent donc respecter les normes logiques : ils ne doivent réaliser que de « bonnes » inférences, à savoir qui préservent la vérité.[1]

L'IA symbolique, pour permettre de résoudre des problèmes, respecte strictement les règles rationnelles. Elle intervient dans le domaine du raisonnement.

Pour ce faire, et c'est ce qui distingue l'IA symbolique des autres types d'intelligences artificielles, elle respecte deux hypothèses fondamentales. La première hypothèse est que

> Les états sont typiquement ce qu'on a appelé au chapitre précédent des propositions, c'est-à-dire des expressions décrivant un certain fait, un certain état de choses ou de situation. Ces expressions sont l'homologue, dans notre système cognitif, de phrases du langage naturel ; pour le dire d'une manière imagée, mais que certains philosophes et psychologues comprennent littéralement, les propositions de notre pensée sont écrites dans un « langage de la pensée », ou « mentalais ». Il nous est naturel, aujourd'hui, de considérer qu'elles sont « codées » dans un langage informatique quelconque. Ces propositions sont, selon le cas, considérées par l'agent comme vraies, fausses, douteuses, ou encore comme des buts à atteindre ou à éviter, à espérer ou à redouter, ou des situations à approuver ou désapprouver.[2]

L'IA symbolique reprend donc une thèse notamment formulée par Fodor qui est qu'il y a un langage de la pensée constituant les ressources manipulables pour les opérations de notre esprit ou d'un ordinateur. Se trouver dans un état, c'est donc formuler une proposition en mentalais. Cette proposition peut être vraie ou fausse, elle exprime une croyance ou un désir. L'IA symbolique est donc une IA qui se positionne au niveau des croyances et des désirs. Ce langage de la pensée amène à réaliser des actions comme formuler des phrases dans le langage naturel par exemple. Nous, les humains, comprenons le sens de ces actions ou de ces mots formulés. Il en est tout autre pour les intelligences artificielles symboliques.

L'idée qu'il y a un langage de la pensée qui est manipulé pour nos capacités cognitives laisse penser qu'il y a un algorithme, une suite d'instructions pour passer d'un état propositionnel à un autre. Cela amène donc à la deuxième hypothèse. Celle-ci

> […] concerne la question de savoir s'il est en notre pouvoir de découvrir ces algorithmes. Une chose est de postuler qu'ils existent, une autre de penser que nous saurons les identifier. Longtemps nous avons su que la Lune avait une face cachée sans que nous puissions l'observer ; de nombre de phénomènes du passé nous savons qu'ils ont eu lieu mais que nous

[1] Andler, page 79-80.
[2] Andler, page 78.

> n'en aurons jamais connaissance. L'IA fait l'hypothèse qu'il en va autrement des processus cognitifs qu'elle cherche à reproduire.[1]

Cette seconde hypothèse stipule donc que nous finirons par identifier les algorithmes responsables des différents états propositionnels. Nous finirons par connaître tous les processus cognitifs.

Pour trouver ces algorithmes, il y a deux voies possibles. La première est la voie anthropique, la première parue historiquement car elle cherche à découvrir les processus de la cognition humaine pour arriver à un certain résultat. C'est la première approche car l'un des buts initiaux de l'IA est de comprendre la cognition humaine. C'était l'objectif des pionniers qui voulaient créer une intelligence de synthèse, créée par l'homme et « identique » à l'intelligence naturelle. Pour trouver des algorithmes, on se base alors sur l'introspection. La seconde voie est l'approche ananthropique. Ici, on ne s'intéresse pas à l'humain. On recherche des algorithmes qui marchent, qui donnent le bon résultat, peu importe le processus pour y parvenir. Cette approche permet plus de latitude dans la recherche d'algorithmes car le répertoire de solutions est plus large. Aussi, elle rend mieux compte des processus secondaires comme la perception sous toutes ses modalités ou la motricité. Ces deux voies sont souvent combinées par les chercheurs pour parvenir à leurs fins.

2. Les problèmes de l'IA symbolique

Maintenant que nous avons les principales caractéristiques de cette IA, penchons-nous de plus près sur ses problèmes et défauts. Ce type d'IA fut celui qui reçut quasiment toute l'attention et toutes les ressources durant le XX^{e} siècle. C'est donc ce type d'IA qui procéda aux différents renoncements que nous avions spécifiés précédemment. Or, cela est tout sauf anecdotique quand nous visons le projet de reproduire une intelligence humaine. Ne serait-ce que le cas de la compréhension. La compréhension notamment des symboles que nous manipulons est une partie fondamentale de l'intelligence humaine. L'IA, n'en étant pas capable, se retrouve séparé de l'intelligence humaine par un fossé.

Aussi, le fait de respecter strictement les règles de rationalité fait que si un algorithme ne donne pas le résultat escompté, alors on change cet algorithme. L'IA symbolique dépend des

[1] Andler, page 86.

connaissances et des règles que nous, humains, lui insufflons. Elle tombe donc dans le problème du contrôle. L'informaticien choisit les règles à appliquer et, si l'algorithme ne va pas, l'informaticien change d'algorithme. L'IA symbolique doit importer la solution depuis l'intelligence humaine.

L'IA symbolique n'a pas su apporter des résultats satisfaisants ce qui a causé une perte de confiance de la part des organismes qui finançaient la recherche. Pour développer les IA symboliques, les chercheurs ont d'abord appliqué la stratégie des petits pas. Les algorithmes étaient développés pour résoudre des problèmes ou passer des tests tout d'abord dans des conditions et des environnements simplifiés. On faisait travailler les ordinateurs dans des « micro-mondes » dans l'objectif de le déployer dans le vrai monde après. Cependant, selon Andler, cette stratégie a rencontré une barrière infranchissable. En effet, il y a deux raisons à cela. La première est :

> […] ce qu'on appelle l'explosion combinatoire : le nombre d'étapes à accomplir croît exponentiellement avec le nombre de cas à considérer — plus le monde est vaste, plus nombreuses sont les trajectoires possibles. Si rapides que soient les ordinateurs, ceux d'alors comme ceux d'aujourd'hui, si grandes leurs mémoires, ils finiraient toujours par être débordés.[1]

Une solution à ce problème serait d'inculquer à l'IA une sorte de « sens commun » en lui fournissant des informations suffisamment riches afin qu'il puisse trouver le bon chemin sans avoir besoin d'explorer toutes les possibilités. Cependant, cela amène à un autre problème, qui concerne tout type d'IA, que les chercheurs ont nommé le problème de la « représentation de la connaissance ». En effet :

> […] comment identifier, formaliser puis organiser ces informations une bonne fois en sorte que, pour chaque tâche dont il doit s'acquitter, le système trouve celles dont il a besoin et celles-là seulement, ce qui est indispensable pour qu'il ne se perde pas, ici encore, dans une combinatoire explosive ?[2]

Or, il semble que, pour le moment, aucune solution à ce problème ne soit trouvée.

La deuxième raison est qu'il n'est pas possible de déterminer toutes les conséquences d'une action. L'action provoque une infinité d'effets sur le monde, importants ou non. Mais il y a également une infinité d'éléments du monde qui ne change pas. La solution à cela serait de faire en sorte que le système néglige une partie des effets, qu'on le contraigne à ne pas prendre en compte les facteurs non pertinents. Cependant :

[1] Andler, page 96.
[2] Andler, page 96.

> Il y aurait […] autant de contraintes qu'il y a de tâches, ce qui imposerait de programmer ces contraintes à partir de la donnée de la tâche, c'est-à-dire à procéder d'avance et pour chaque cas à l'examen qu'il s'agissait d'éviter.[1]

Les algorithmes d'IA symboliques ne parviennent donc pas à bien à être mobilisés dans le monde réel.

Peut-être que le problème vient des fondements mêmes de l'Intelligence Artificielle symbolique. En effet, toujours d'après Andler, beaucoup de penseurs ont conclu que l'IA symbolique était dans une impasse. En effet :

> […] c'est bien ce qu'ont pensé des auteurs aussi différents que Hubert Dreyfus, philosophe d'inspiration phénoménologique, critique de la première heure, et Marvin Minsky, co-fondateur de la discipline. Opposés en tout, ils s'accordaient sur une chose : l'IA des commencements était condamnée.[2]

Quels sont les arguments en faveur de cette conclusion dramatique ? Tout d'abord, et nous l'avons déjà abordé, il y a le problème du sens. Les systèmes d'IA symboliques ne comprennent pas, ne savent pas, ne connaissant pas ce sur quoi portent les opérations qu'ils effectuent. Ce sont des machines dépourvues d'un « intérieur » qui pourrait permettre cela. Il y a une cécité sémantique qui est un obstacle majeur pour la recherche en Intelligence Artificielle. Les chercheurs ont d'ailleurs fini par renoncer à l'idée que l'IA puisse comprendre.

Le second argument a notamment été abordé par Hubert Dreyfus dans son ouvrage *Intelligence artificielle : mythes et limites*[3]. Il défend l'idée que l'IA symbolique « serait peut-être radicalement erronée, il faudrait alors le modifier en profondeur, voire l'abandonner complètement »[4]. L'IA symbolique ne pourra surement jamais parvenir à reproduire une intelligence humaine car « ce n'est pas ainsi que fonctionne l'esprit humain »[5]. Pour affirmer cela, Dreyfus s'appuie sur des « descriptions phénoménologiques ou des observations empiriques apparemment incompatibles avec la conception générale de l'IA classique ou avec les modèles particuliers qu'elle avait construits »[6].

Ces arguments ne furent pas décisifs aux yeux des chercheurs en IA car :

---

[1] Andler, page 98.
[2] Andler, page 99.
[3] Hubert L. Dreyfus, *Intelligence artificielle : Mythes et limites*, trad. par Rose-Marie Vassallo-Villaneau (Flammarion, 1984).
[4] Andler, *Intelligence artificielle, intelligence humaine*, page 101.
[5] Andler, page 101.
[6] Andler, page 101.

> D'une part, ils n'établissaient pas de manière irréfutable que même compris comme description de leur phénoménologie, les processus mentaux étaient conformes à la description qu'avançaient les critiques. D'autre part, cette description n'était pas incompatible avec les modèles de l'IA symbolique, dont on pouvait imaginer qu'ils se situaient à un autre niveau, de la même manière qu'un gaz peut être compris comme un milieu homogène remplissant tout le contenant et comme un ensemble peu dense de molécules s'entrechoquant au hasard. Par ailleurs, pour l'approche ananthropique, la manière dont s'y prend l'esprit humain était sans pertinence.[1]

Cependant, ces arguments permirent de mettre en évidence le fait que l'IA symbolique n'était pas l'unique hypothèse valable en matière d'intelligence artificielle. Il n'était dès lors plus obligatoire de créer des modèles compatibles avec l'IA symbolique. De nouveaux types d'IA purent alors se développer, notamment l'IA connexionniste.

3. L'IA connexionniste ou l'ère du neurocalcul

L'IA connexionniste est aussi vieille que l'IA symbolique. En effet, dans l'article « A logical calculus of the ideas immanent in nervous activity »[2] de Warren S. McCulloch et Walter Pitts paru en 1943, l'on retrouve le point de départ du neurocalcul. Dans l'article, McCulloch et Pitts font différents postulats sur ce qu'est censé être l'activité d'un neurone. Grâce à cela, ils énoncent et démontrent différents théorèmes concernant les réseaux de neurones ce qu'ils leur permettent de venir à la conclusion que « les événements neuronaux et les relations entre eux peuvent être traités au moyen de la logique propositionnelle »[3]. En effet, pour eux :

> Pour la psychologie, quelle que soit sa définition, la spécification du réseau apporterait tout ce qui peut être réalisé dans ce domaine, même si l'analyse était poussée jusqu'aux unités psychiques ultimes ou "psychons", car un psychon ne peut être rien de moins que l'activité d'un seul neurone. Comme cette activité est intrinsèquement propositionnelle, tous les événements psychiques ont un caractère intentionnel ou "sémiotique". La loi du "tout ou rien" de ces activités et la conformité de leurs relations à celles de la logique des propositions assurent que les relations des psychons sont celles de la logique à deux valeurs des propositions. Ainsi en psychologie, qu'elle soit introspective, comportementale ou physiologique, les relations fondamentales sont celles de la logique à deux valeurs.[4]

---

[1] Andler, page 101.
[2] Warren S. McCulloch et Walter Pitts, « A Logical Calculus of the Ideas Immanent in Nervous Activity », *The Bulletin of Mathematical Biophysics* 5, nº 4 (1 décembre 1943): 115-33.
[3] McCulloch et Pitts, page 115 : « […] neural events and the relations among them can be treated by means of propositional logic ».
[4] McCulloch et Pitts, page 131 : « To psychology, however defined, specification of the net would contribute all that could be achieved in that field even if the analysis were pushed to ultimate psychic units or "psychons," for a psychon

Cet article présente un neurone formel permettant *in fine* de reproduire la logique classique. Les mathématiques et le raisonnement étant intrinsèquement liés à la logique, il est alors possible de faire des calculs avec ce type d'objet. Le neurocalcul est né. Il faut cependant noter que cet article est aujourd'hui daté et ne correspond plus à la compréhension actuelle des neurones dans la recherche en psychologie.

Cette idée de neurone formel continuera son chemin jusqu'en 1957 où Frank Rosenblatt inventa le perceptron, un algorithme mobilisant des fonctions mathématiques et faisant office de neurone formel ou plutôt de réseau de neurones le plus simple. D'après Andler :

> [...] le perceptron inventé par Rosenblatt en 1957 ne différait de l'ordinateur apparu dix ans plus tôt que par sa grande simplicité - c'était une machine du même genre, et la question était de savoir laquelle était la mieux à même, au prix de perfectionnements, de reproduire un calcul des idées complet, c'est-à-dire une intelligence.[1]

Cependant, bien que Rosenblatt était convaincu du potentiel de sa création, il était seul face aux autres fondateurs de l'IA qui, eux, privilégiaient l'approche symbolique. Le neuroalcul fut donc relégué au second plan mais continua tout de même à se développer. Des chercheurs comme Minsky entreprirent de surmonter la grande simplicité du perceptron en complexifiant sa structure et en ajoutant des couches de neurones supplémentaires. Cela a notamment consisté en l'ajout de couches qui ne sont pas en contact avec l'environnement. Ces couches constitueraient le support de la partie cogitative et délibérative de la pensée.

Cependant, il fallait une méthode pour « éduquer » les réseaux multicouches car le processus d'éducation semblait impossible à réaliser étant donné le temps de calcul que cela nécessitait. La solution a été la redécouverte d'une méthode appelée « rétropropagation », qui existait depuis les années 1960 et que l'on consolida à la première moitié des années 1980.

À force de perfectionnement, il fut établi dans les années 1980

> […] un résultat mathématique montrant que les réseaux multicouches possèdent une forme d'universalité, un résultat qui équivaut, dans le contexte du neurocalcul, à l'existence d'une machine de Turing universelle dans le contexte de l'IA symbolique. En un mot, le neurocalcul semblait tenir enfin l'outil passe-partout qu'il espérait depuis Rosenblatt.[2]

---

can be no less than the activity of a single neuron. Since that activity is inherently propositional, all psychic events have an intentional, or "semiotic," character. The "all-or-none" law of these activities, and the conformity of their relations to those of the logic of propositions, insure that the relations of psychons are those of the two-valued logic of propositions. Thus in psychology, introspective, behavioristic or physiological, the fundamental relations are those of two-valued logic ».

[1] Andler, *Intelligence artificielle, intelligence humaine*, page 124.

[2] Andler, page 125-126.

Une machine de Turing universelle étant une machine pouvant théoriquement exécuter tout type de programme, les réseaux de neurones devinrent alors eux aussi pertinents pour réaliser des tâches complexes, variées et, pourquoi pas, pour reproduire un esprit humain.

À partir des années 1990 et surtout 2000, l'IA connexionniste s'est popularisée notamment grâce à l'amélioration de la puissance de calcul des machines et de la collecte massive de données :

> […] la résurgence actuelle des réseaux de neurones a été surtout permise par l'augmentation des capacités de calcul des ordinateurs (avec à présent des processeurs dédiés à ces calculs) et la disponibilité massive des données nécessaires à ces algorithmes d'apprentissage profond.[1]

Les données, nécessaires à ce type d'IA, ont été rendues disponibles par les grands opérateurs internet américains et chinois. Pour ce qui est de la puissance de calcul et des progrès technologiques, il faut souligner l'importance qu'a eu l'industrie du jeu vidéo qui a inventé la carte graphique, supposant de faire fonctionner des réseaux multicouches en un temps réduit.

À partir des années 2010, l'appellation « réseau de neurones » devint « *deep learning* » et ce type d'IA devint le plus populaire avec des résultats remarquables comme, par exemple, l'algorithme AlphaGo qui gagna une partie de go contre l'un des meilleurs joueurs mondiaux. Aujourd'hui, l'IA semble se démocratiser dans tous les domaines de la vie quotidienne. Que ce soit la reconnaissance et génération d'images, la génération ou correction de texte, l'animation de personnages non joueurs dans les jeux vidéo, le *deep learning* est présent partout.

Essayons alors de comprendre, succinctement et schématiquement, le fonctionnement d'un réseau de neurones. Comme nous l'avons vu, les réseaux sont constitués d'automates qui sont des neurones formels, des perceptrons. D'après Andler :

> Chacun de ces automates (les « unités » ou « nœuds » du réseau) peut être dans l'un de deux états, inactif (codé par 0) ou « actif » (codé par 1) ; il calcule son état à chaque instant à partir des stimulations qu'il reçoit des unités auxquelles il est connecté ; et il transmet une stimulation, fonction de son état, à certaines unités. Les interactions entre les nœuds du réseau sont véhiculées par des connexions physiques et modulées par des coefficients appelés « poids », analogues des poids synaptiques modulant l'influence qu'un neurone exerce sur un autre via son axone.
>
> Les modèles de base du deep learning sont organisés en couches, chaque couche exerçant une influence sur la suivante.
>
> Ils sont le plus souvent utilisés comme des systèmes « feed-forward » dans lesquels l'information circule dans la même direction, depuis l'entrée (l'input) jusqu'à la sortie

[1] GdRIA du CNRS, ouvrage coordonné par Sébastien Konieczny et Henri Prade, *L'intelligence artificielle. De quoi s'agit-il vraiment ?*, page 11.

> (l'output), en traversant successivement les couches internes. Certains réseaux, dits récurrents, peuvent aussi comprendre des boucles qui « recyclent » une sortie en nouvelle entrée.[1]

Nous avons donc un fonctionnement très technique et difficilement prédictible par les non spécialistes à cause du nombre élevé de calculs, là où il était plus facile de concevoir l'IA symbolique comme consistant en des algorithmes. Un réseau est d'abord créé et préparé, puis entrainé et enfin déployé. Lors de la création, nous déterminons le nombre de couches, le nombre d'unités, c'est-à-dire le nombre d'automates constitués de neurones formels et qui correspondent à des nœuds du réseau, et les connexions entre les unités. Pour préparer le réseau, on le confronte à des données qui sont aménagées de manière qu'elles soient acceptables par le réseau c'est-à-dire sous forme de vecteurs de certaines dimensions. Il y a donc une différence fondamentale avec l'IA symbolique puisque ce dernier type d'IA travaille sur des symboles alors qu'ici, nous travaillons sur des vecteurs, un ensemble de valeurs mathématiques. Ensuite, le réseau est entraîné. En effet, comme nous avons pu le suggérer précédemment, le *deep learning* a cette particularité de pouvoir apprendre. Il faut lui apprendre à faire ce que l'on veut qu'il fasse. Le réseau doit donc passer par une phase d'apprentissage qui est l'étape la plus complexe du processus. Il faut noter que cet apprentissage est supervisé. L'IA connexionniste n'apprend pas toute seule. Il faut une présence humaine. Pour faire l'apprentissage il faut constituer une base de données. Cette base « comprend un nombre considérable, voire astronomique, d'items du genre de ceux que l'on veut rendre le réseau capable de traiter »[2]. Ainsi :

> L'entraînement consiste alors à faire défiler les exemples autant de fois que nécessaire, en modifiant progressivement les poids des connexions, à la lumière des erreurs commises par le réseau, jusqu'au moment où toutes ses réponses sont correctes.
>
> De manière abstraite et générale, ce qu'il s'agit d'apprendre, c'est à détecter, au sein d'une classe C d'entités (objets, individus, événements, situations...), celles qui ont l'une ou l'autre d'une famille de propriétés choisies en raison de leur intérêt théorique ou pratique. On veut être capable de répondre à toute question de la forme suivante : X a-t-il P ? où X est un membre de la classe C en question, et P l'une des propriétés considérées.[3]

Enfin, le réseau est déployé, c'est-à-dire qu'il est utilisé comme un algorithme dans lequel nous entrons un ensemble de valeurs booléennes et qui donnera comme résultat un vecteur comprenant les états des unités de la dernière couche.

---

[1] Andler, *Intelligence artificielle, intelligence humaine*, page 105-106.
[2] Andler, page 107.
[3] Andler, page 108.

Comme nous pouvons le constater, le *deep learning* a pour fonction première de classer des objets dans diverses catégories. Il range les objets qu'on lui fournit dans des boîtes, des catégories qu'on a préalablement déterminées. Cependant, il ne fait pas que cela car, s'il est bien entrainé pour classer correctement certains types d'objets, alors, quand nous lui présenterons un nouvel objet qui ne fait partie de la base d'apprentissage, il sera en mesure de le classer et donc de nous dire dans quelle catégorie il appartient. Ce type d'IA, pour classer, doit prédire. Andler prend pour exemple une compagnie d'assurance. Il dit :

> Prenons le cas d'une déclaration de sinistre. L'assureur cherche à savoir s'il est ou non frauduleux. Un réseau peut être entrainé sur un grand nombre de dossiers précédents, sur le caractère frauduleux ou non desquels des experts humains, suite à enquête, ont pu se prononcer. S'il fonctionne correctement, le réseau auquel on soumet une nouvelle déclaration pourra la classer, avec une probabilité de succès élevée, comme frauduleuse ou légitime.
>
> En ce sens étendu de la vision, les tâches de catégorisation peuvent être comprises comme des tâches de complétion, et réciproquement.[1]

Or, la complétion peut être comprise comme une forme de prédiction. En effet, « ranger un objet particulier perçu (par exemple visuellement) dans une catégorie, c'est "prédire" sa catégorie à partir de son apparence visuelle »[2].

Nous voyons donc que l'IA connexionniste se distingue nettement de l'IA symbolique de par sa structure même. Nous avions vu qu'il y avait deux niveaux dans les processus cognitifs. Il y a un niveau inférentiel, conscient, logique, et un second niveau associatif, automatique et inconscient. Nous avions également vu que l'IA symbolique se situe dans le premier niveau. L'IA connexionniste, elle, se situe dans le second niveau. En effet, d'après Andler :

> Une différence importante entre approche symbolique et approche connexionniste, on l'a vu, porte sur les fonctions cognitives que chacune place au centre de sa visée. Pour l'IA symbolique, ce sont les fonctions supérieures (raisonnement, etc.), pour le connexionnisme les fonctions inférieures (perception, motricité). De là découle une différence quant au niveau pertinent de l'information traitée par les systèmes. Les représentations sur lesquelles le système calcule se situent, pour l'IA symbolique, au niveau « personnel » : elles correspondent à des concepts du langage naturel, identifiables par le sens commun raffiné si besoin par l'enquête conceptuelle.
>
> Celles du connexionnisme et plus généralement du neuro-calcul relèvent d'un niveau « subpersonnel » qui mobilise des concepts scientifiques n'ayant généralement aucun

[1] Andler, page 114.
[2] Andler, page 114.

> pendant dans le langage naturel et qui sont inaccessibles au sens commun : seul le recours aux sciences empiriques et aux mathématiques permet d'y accéder.[1]

L'IA symbolique se concentre sur donc les fonctions supérieures comme le raisonnement et que le connexionnisme se concentre sur les fonctions inférieures comme la perception ou la motricité.

Une autre différence que l'on peut noter est que pour l'IA symbolique, les connaissances résident dans les symboles et dans les connaissances que les programmeurs ont insufflées dans l'algorithme. Pour l'IA connexionniste, les connaissances résident dans les connexions et dans les poids, ces derniers s'ajustant via l'apprentissage et non directement via l'humain. La différence plus notable concerne alors l'apprentissage. L'IA symbolique n'apprend pas, elle applique les instructions qu'elle a reçues. Un réseau de *deep learning,* lui, acquiert des capacités via les données qu'il reçoit. L'IA connexionniste pratique un apprentissage inductif. Ainsi, d'après Andler :

> Plus fondamentalement, on a affaire à deux modalités différentes, presque deux sens différents d'apprentissage : l'apprentissage symbolique consiste à i*nstruire* le modèle, à lui communiquer un *savoir*, l'apprentissage connexionniste à l'*exposer* à des *exemples*, à lui inculquer un *savoir-faire*.[2]

L'apprentissage inductif constitue donc un avantage important pour l'IA connexionniste car, l'apprentissage autonome semble être une condition essentielle pour devenir intelligent. D'autant plus que l'induction joue un rôle dans l'intelligence humaine. Une IA qui possède un processus inductif se rapproche donc un peu plus de l'intelligence humaine.

Cependant, il faut noter que le *deep learning* ne cherche pas à reproduire l'intelligence humaine telle quelle. En effet, d'abord, il rejette l'approche anthropique. Ensuite, il ne cherche pas à reproduire les processus humains ou à s'en inspirer. Troisièmement, il rejette tous les principes constitutifs de l'IA symbolique, donc les principes comme le langage de la pensée ou le respect stricte de la rationalité. Enfin, il n'apprend qu'à travers les expériences. Le *deep learning* respecte un empirisme strict.

[1] Andler, page 112.
[2] Andler, page 119.

4. Les évolutions du deep learning

Néanmoins, le *deep learning* évolue et la recherche tend à l'orienter vers une approche de plus en plus anthropique :

> L'inspiration des sciences cognitives et de l'IA symboliques oriente le DL vers une autre ressource jugée essentielle. Il s'agit de la capacité, déjà évoquée, à se saisir de domaines dits combinatoires ou compositionnels constitués d'assemblages itérés d'éléments simples — les exemples les plus familiers étant les jeux de construction tels que le Lego ou le Meccano, les molécules chimiques, et surtout, pour ce qui nous concerne, les langages naturels et artificiels ou encore le raisonnement.[1]

Si l'on parvenait à lui faire acquérir une telle capacité, le *deep learning* serait alors débarrassé de ce qu'on appelle sa « cécité algébrique ». Il serait alors en mesure de faire des calculs et des raisonnements comme le fait l'IA symbolique. Franchir cette étape est jugée cruciale car « un tel système serait, à l'image de l'humain tel que les sciences cognitives nous le présentent, à la fois un statisticien et un logicien ou un algébriste »[2]. Un tel système serait alors capable d'« apprendre à apprendre », c'est-à-dire qu'un apprentissage réussi pourra être utilisé dans d'autres domaines, ce qui simplifierait grandement la création de nouveau modèles de *deep learning* et leurs diversifications.

Pour l'instant, cet objectif n'est pas atteint mais la recherche avance dans cette direction. De nombreux progrès sont faits concernant l'apprentissage, notamment l'apprentissage autosupervisé et le préapprentissage des IA connexionnistes qui permettent d'échapper à l'apprentissage supervisé. Il y a aussi le développement de facultés à manier des structures combinatoires comme le fait l'IA symbolique et, surtout, il y a eu la création par la société DeepMind en 2017 des transformateurs. Un transformateur est

> […] un assemblage complexe de modules dont chacun joue un rôle déterminé dans un processus à plusieurs étapes, mobilisant toute une série de ficelles du métier difficilement compréhensibles pour le profane.[3]

Aussi complexe soit ce nouveau type d'outil, il est à l'origine, avec les avancées en matière d'apprentissage autosupervisé, de l'apparition en 2018 des « modèles massifs de langage » comme GPT, créé par OpenAI.

[1] Andler, page 148.
[2] Andler, page 149.
[3] Andler, page 151.

Ces modèles

> […] se sont révélés capables de s'acquitter, avec un succès variable mais généralement très élevé et supérieur à la performance des modèles existants, de tout un ensemble de tâches relevant soit de ce qu'on appelle la compréhension du langage naturel (natural language understanding, NLU), soit de la génération (production) de langage naturel (natural language generation, NLG).[1]

L'évolution des modèles massifs de langage est fulgurante et il est maintenant possible de converser naturellement avec la dernière version de GPT. Sans oublier que d'autres modèles sont apparus et sont capables de générer des images, des vidéos, de la musique à partir de simples indications textuelles de l'utilisateur. Aujourd'hui, nous sommes actuellement en pleine effervescence des IA génératives.

5. *Les problèmes de l'IA connexionniste*

Le *deep learning* paraît prometteur à l'époque où nous vivons. Cependant ce type d'IA a, comme l'IA symbolique, lui aussi des défauts.

En effet, l'IA connexionniste, en plus de sa « cécité algébrique », souffre de « cécité sémantique » comme l'IA symbolique. Elle ne comprend pas le sens ce qu'elle manipule. Les connaissances sont insufflées par le concepteur indirectement à l'IA connexionniste via l'apprentissage, mais elle n'a aucune idée de ce sur quoi elle produit ses calculs. C'est le concepteur qui choisit d'entraîner l'IA sur telle base d'apprentissage et pas une autre, c'est lui qui comprend la signification des objets que l'IA manipule. Nous avions vu que cette cécité amenait à affirmer que l'IA symbolique manquait de « sens commun » ce qui apportait des défaillances. Pour l'IA connexionniste,

> […] les défaillances sont inattendues et difficilement explicables, et donc d'autant plus dangereuses. Cela vient de la manière empirique dont ces modèles acquièrent leur compétence, à partir d'une base surabondante d'exemples pris au hasard et non méthodiquement. Or ce n'est pas fortuit : introduire une méthode serait obliger le concepteur à introduire au départ une dose d'intelligence, perdant de ce fait l'avantage de l'empirisme pur qui fait l'intérêt du DL.[2]

[1] Andler, page 158.
[2] Andler, page 132.

Qui plus est, cette cécité a pour conséquence qu'il est facile de tromper et fausser un modèle de *deep learning*. En introduisant un exemple faux, le modèle peut devenir inutilisable. Ainsi, d'après Andler :

> […] une très petite modification dans une image peut conduire un réseau entraîné à l'identifier correctement à produire un verdict totalement erroné : un avion devient un chien, un cheval devient une voiture, un navire devient un camion... C'est d'autant plus déconcertant que l'œil humain peut être incapable de détecter la moindre différence entre l'image d'origine et l'image trafiquée (elle peut être réduite à un seul pixel !).[1]

Bien évidemment, il est possible d'entrainer les modèles afin de résister à ces faux exemples ou erreurs, mais cette résistance a des limites. Ainsi, malgré le fait que le *deep learning* ait parfois des performances équivalentes voire supérieures aux capacités humaines, un simple pixel mal placé peut tout faire dérailler. La tolérance aux écarts par rapport aux cas normaux est quasiment nulle. La correction de ces erreurs peut se faire par l'humain mais seulement après coup. Il est beaucoup plus difficile de prévoir les erreurs auxquelles le système pourrait être confronté puisque ce qu'il se passe à l'intérieur des différentes couches ne nous est pas accessible. Nous avons juste accès aux données que l'on lui fournit et aux résultats que donne l'IA.

Le fait que les modèles de *deep learning* ne soient pas tolérants et ne puissent pas se corriger est en lien avec un autre problème qui est qu'ils n'ont pas connaissance du monde, du contexte. En effet, le *deep learning*

> […] demeure par construction hermétiquement isolé du monde dont proviennent les stimuli dont il a appris à détecter les régularités statistiques. Le contexte dont sont extraits ces stimuli lui reste caché. Il n'est doté d'aucun biais ou connaissance « innée » (inscrite en lui avant tout apprentissage) qui l'aiderait, notamment, à éviter de tomber dans de nouveaux pièges. Du fait qu'il n'est pas en contact direct avec l'environnement sur lequel porte son activité, il ne peut pas acquérir par lui-même les connaissances qui lui manquent.[2]

Ce manque de conscience de faire partie d'un monde constitue un obstacle majeur pour que les modèles de *deep learning* soient solides et se corrigent par eux-mêmes. Si nous parvenons à franchir cet obstacle, cela représenterait un pas de plus franchi vers la reproduction d'une intelligence humaine.

En plus de ces défauts, le *deep learning* a également des limites. Par exemple, il ne peut pas raisonner. En effet, l'IA connexionniste, une fois entraînée, excelle dans sa tâche. Cependant le réseau aura beaucoup plus de mal à gérer des cas qui s'éloignent trop de la base d'apprentissage. Il

[1] Andler, page 133.
[2] Andler, page 134.

faudrait que cette IA soit capable d'extrapoler c'est-à-dire d'être en mesure de classer un objet qui est éloigné en tout point de la base d'apprentissage. Si le *deep learning* était capable d'extrapolation, alors il n'aurait aucun problème à sortir de sa zone de confort tout en continuant à exceller. Pour ce faire, il faudrait être capable de raisonner, de faire des abstractions, des choix et de s'aventurer dans des régions que nous ne connaissons pas bien. Or, les réseaux de *deep learning* n'en sont pas capables car

> Les domaines de déploiement naturel du DL sont des ensembles finis d'objets aux propriétés déterminées une bonne fois — plus précisément, constitués de nombreux objets regroupés en familles. Le raisonnement, le langage sont des systèmes capables d'engendrer une infinité d'objets différents, dont chacun a des propriétés qui le distinguent des autres. Pour s'en saisir, le DL doit en construire des modèles approchés, constitués d'exemples, modèles qui introduisent par endroits des déformations graves, un peu à la manière dont une carte du globe en donne une représentation convenable dans certaines régions et sérieusement défectueuse dans d'autres. Quand par malheur le réseau s'aventure dans une mauvaise région, ses réponses peuvent être désastreuses, comme le confirme l'expérience.[1]

Tenter de reproduire l'infinité des raisonnements possibles dans un cadre qui, lui, est limité, semble être un objectif de taille voire un objectif impossible à surmonter. D'autant plus que nous, les humains, comprenons et donnons sens au monde en raisonnant sur les causes et les effets. Nous estimons donc que c'est une caractéristique nécessaire pour être intelligent. Or, l'IA n'est toujours pas capable de faire des raisonnements causaux. Notons que ce dernier obstacle n'est pas propre au *deep learning* mais se présente à tous les types d'IA.

Le *deep learning* a donc des défauts et des limites. Il s'avère également que ce type d'IA rencontre des problèmes pratiques. En effet, tout comme pour l'IA symbolique, l'IA connexionniste est défaillante lorsqu'elle quitte le laboratoire et est confrontée au monde réel. Selon Andler, il y a deux raisons à cela. Il y a d'abord le *data shift* qui est le « décalage entre la distribution des traits dans l'ensemble des objets à traiter et la distribution de ces traits dans la base d'apprentissage »[2]. Ainsi, un système peut être biaisé car il a sous-représenté une catégorie à la suite de son apprentissage. La deuxième raison est ce que l'on appelle la sous-spécification (*underspecification*) c'est-à-dire que :

> […] la procédure de fabrication n'est pas suffisamment spécifique pour conduire à des modèles aux performances équivalentes. Ce qui les distingue est un ensemble de décisions d'intendance apparemment sans portée théorique : beaucoup de paramètres doivent être fixés au départ de manière arbitraire, ce qui n'affecte pas leur performance en laboratoire, mais qui

[1] Andler, page 144.
[2] Andler, page 136.

> affecte leur comportement sur le terrain. Ceux qui marchent bien se distinguent des autres sans que l'on sache à quoi est due la différence : ils doivent leur « bonne note » à la chance.[1]

Cette sous-spécification a pour conséquence que malgré les tests que l'on fait sur les modèles mis au point, c'est-à-dire le fait de vérifier si les modèles sont bien construits et entraînés, ils déçoivent quand ils sont déployés. Le problème ne vient pas des tests mais bien de la manière dont nous créons les modèles car :

> […] on constate que la même procédure d'apprentissage à partir de la même base peut conduire à des modèles qui passent parfaitement le test, si rigoureux soit-il, et dont les performances sur le terrain varient, certains étant excellents, d'autres médiocres.[2]

Il y a donc un problème inhérent à la manière dont nous concevons les modèles d'IA connexionnistes.

Enfin, nous avions parlé d'évolutions ou plutôt de progrès actuels concernant le *deep learning* comme les modèles massifs de langage. Cependant, en plus d'avoir les mêmes faiblesses que le *deep learning* de base comme l'incapacité à comprendre et à raisonner, le fait de donner des résultats parfois aberrants ou d'être biaisés par leur apprentissage, ils ont également d'autres défauts bien à eux. Il y a notamment le fait que ces modèles massifs ont un coût astronomique en création qui ne peut être assumé que par des méga-entreprises privées ou des pays. L'autre souci c'est que ces modèles massifs nous trompent :

> […] ils font croire qu'ils produisent un discours sensé, mais ce n'est là que triche ou prestidigitation, comme nous le constatons quand ils déraillent. Leurs connaissances, ils les tiennent des productions humaines. Ils ne font que recopier ce qu'ils trouvent dans l'immense archive disponible sur internet - non, c'est exagéré, ils ne recopient (généralement) pas, mais ce ne sont quand même que des « perroquets statistiques » (l'expression est de la linguiste Emily Bender, une sévère critique de GPT-3) […]. S'ils semblent capables de dialoguer avec tact et pertinence, ce n'est que l'effet de notre projection anthropomorphe : le dialogue authentique nécessite d'accéder aux intentions communicatives de l'interlocuteur, ce qui n'est évidemment pas à la portée de GPT-3, de ChatGPT ou de LaMDA qui n'ont pas la moindre idée de ce qu'est une intention ou de ce qu'est communiquer.[3]

Le problème de la cécité sémantique demeure donc. L'IA connexionniste, malgré les immenses progrès qu'elle a faits durant ces dernières années, reste encore très loin de l'idéal qu'ont eu et qu'ont encore de nombreux chercheurs, à savoir reproduire l'intelligence humaine.

[1] Andler, page 137.
[2] Andler, page 137.
[3] Andler, page 165-166.

Peut-être qu'IA symbolique et IA connexionniste rencontrent un plafond de verre infranchissable. Certains considèrent que les limites ne seront jamais franchies et se concentrent alors sur d'autres programmes de recherche comme « l'intelligence augmentée », qui cherche à développer les appareils numériques afin de permettre d'agrandir les capacités de l'humain, de lui faciliter la vie, mais pas de le remplacer. Qui plus est, l'intelligence augmentée estime qu'il n'est pas souhaitable qu'une machine possède une intelligence humaine. Pour les chercheurs dans ce domaine, l'ordinateur et l'IA ne doivent rester qu'un outil comme c'est le cas actuellement.

Nous avons désormais une vision globale, quoique schématique, des deux principaux types d'IA. Nous voyons également leurs limites et défauts. Désormais, tâchons de voir, à la lumière du modèle de l'esprit de McDowell, si l'IA aura un jour la possibilité de reproduire un esprit.

Chapitre 5 : L'Intelligence Artificielle et l'esprit humain

1. L'Intelligence Artificielle symbolique ne peut pas reproduire un esprit

À la lumière de ce que nous venons de voir sur les deux principaux types d'intelligence artificielle, nous pouvons d'ores et déjà affirmer que l'Intelligence Artificielle symbolique ne peut pas parvenir à l'objectif tant désiré par certain qui est de reproduire un esprit humain. En effet, si comme McDowell, nous sommes naturalistes modérés ou si, comme Putnam, nous sommes des réalistes naturels, alors nous rejetons le fonctionnalisme computationnel tel qu'il fut pensé par Putnam en 1967 dans son article « La nature des états mentaux ». Comme nous l'avons vu dans le chapitre 3, Putnam a développé différents arguments à l'encontre du computationnalisme comme le manque de formalisme précis, l'impossibilité de réduire en des termes computationnels une attitude propositionnelle sans avoir à décrire l'univers visible, ou encore l'aspect trop chauvin de la théorie qui est une conséquence de la réalisation multiple, stipulant que différents organismes peuvent avoir un même état mental. Ainsi, pour peu que l'on soit réaliste naturel, qui est une évolution du réalisme interne, le computationnalisme n'est pas valable.

Or, nous avions dénoté la proximité qu'il y a entre le computationnalisme et l'IA symbolique. Et pour cause, le computationnalisme postule que l'esprit fonctionne de manière analogue à un ordinateur. Notre esprit, selon cette théorie, ne ferait que suivre une table de Turing, ne ferait qu'exécuter des algorithmes déterminés. L'IA symbolique est dans la même démarche. Elle consiste uniquement en l'exécution d'algorithmes par un ordinateur. Le computationnalisme stipule qu'il y a un langage de la pensée qui est manipulé de manière computationnelle par les esprits. Cette stipulation fut reprise par les chercheurs en IA symbolique et ils partirent alors du principe qu'il suffirait de créer une IA puissante, dotée d'algorithmes suffisamment fins et d'une puissance de calcul convenable, qui soit capable de reproduire les mêmes processus que l'on retrouve dans les cerveaux humains afin de reproduire un esprit humain. Le computationnalisme est donc sous-entendu dans le programme de recherche en IA symbolique. Du moins, il était sous-entendu puisque l'IA symbolique est tombée en désuétude concernant l'objectif de reproduire un esprit.

Putnam, en démontrant la fausseté du computationnalisme originaire, a donc également démontré conceptuellement et indirectement l'impossibilité pour l'IA symbolique de reproduire un esprit humain. L'esprit humain ne fonctionne pas ainsi selon lui. Si l'on est réaliste interne et que nous voudrions reproduire l'esprit humain, il faut se pencher sur l'autre type d'IA, l'IA connexionniste.

Pour savoir si l'IA connexionniste échouera elle aussi à reproduire un esprit humain, il faut que nous la placions dans des conditions idéales. En effet, le *deep learning* est encore aujourd'hui en constante évolution. Les prochaines années seront sans doute sans aucune mesure en termes de résultats comparés à maintenant. Dire que ce type d'IA ne peut pas reproduire un esprit car les réseaux de neurones ne sont pas suffisamment complexes ou qu'il n'y a pas assez de puissance de calcul n'est pas viable car le temps finira possiblement par démontrer le contraire.

Il faut supposer une IA *deep learning* qui serait dans un robot par exemple muni de différents capteurs afin d'avoir des informations sur son environnement tout comme nous, les humains. Le robot pourrait donc se déplacer, recevoir des informations, communiquer. Il aurait aussi une puissance de calcul suffisante et un réseau de neurones multicouches suffisamment important et bien conçu en vue de reproduire un comportement intelligent. La question à se poser alors est : « est-ce que cette IA se trouvant dans des conditions idéales pourrait reproduire un esprit ? Et, si c'est le cas, est-ce qu'elle pourrait reproduire un esprit humain ? ». Nous allons tenter de montrer que malgré ces conditions que l'on pourrait considérer comme idéales, l'IA ne pourra atteindre ce dernier objectif.

Face à cette machine supposément parfaite, quels arguments avons-nous pour affirmer que celle-ci ne pourra jamais avoir d'esprit comme un humain ? Plusieurs pistes s'offrent à nous. Tout d'abord, nous allons aborder celle qui nous semble la plus évidente d'entre elles qui est celle de l'impossibilité pour l'IA d'avoir des qualia ou ce que l'on appelle une conscience phénoménale. Nous parlerons ensuite de l'introspection que ce soit les arguments stipulant que l'IA ne peut pas reproduire une telle capacité mais également en discutant les théories qui pourraient rendre une IA introspective possible. Puis, nous aborderons le sujet complexe de la créativité afin d'affirmer que l'IA ne semble pas en avoir, ce qui nous permettra également d'émettre des doutes sur le fait qu'elle puisse faire preuve d'imagination. Enfin nous aborderons le problème plus kantien (et macdowellien) de la spontanéité et de son apparente irreproductibilité par des modèles de type *deep learning*, ce qui nous amènera à aborder qu'elle serait la meilleure voie à suivre pour recréer un

esprit selon la pensée de McDowell pour conclure que le programme d'Intelligence Artificielle tel qu'il est ne pourra pas reproduire un esprit humain.

2. La conscience phénoménale et l'illusionnisme

Les qualia, ces ressentis qualitatifs, subjectifs et ineffables dont nous avons déjà parlé, se retrouvent dans une forme de conscience que l'on appelle conscience phénoménale. D'après François Kammerer, la conscience phénoménale c'est :

> […] la forme de conscience qui correspond à l'ensemble des expériences phénoménales. Les expériences phénoménales, elles, sont définies comme les états mentaux, tels qu'une sensation de douleur à l'épaule ou l'expérience visuelle d'un champ de colza, qui sont éprouvés, ressentis subjectivement – qui font un certain effet aux sujets qui sont dans ces états2. On dit généralement que ces expériences phénoménales possèdent des « propriétés phénoménales », en vertu desquelles elles sont des expériences phénoménales d'un certain type (une sensation de douleur, une expérience de rouge, etc.) Le terme de « qualia » est également utilisé pour référer à ces propriétés phénoménales.[1]

Les qualia sont donc les propriétés des expériences phénoménales. Quand je fais l'expérience d'un objet rouge, le quale de rouge est « l'effet que cela fait » de voir du rouge.

On parle de conscience phénoménale car on distingue plusieurs types de conscience. En effet, il y a aussi la conscience d'accès, la conscience quasi-phénoménale ou encore la conscience de soi. La conscience d'accès, c'est la forme de conscience « définie non en termes de ressenti subjectif, mais en termes d'accessibilité de certains contenus informationnels pour le raisonnement et le contrôle rationnel de l'action et du discours »[2]. La conscience de soi, elle, est la forme de conscience « qu'on peut définir comme la capacité d'avoir une représentation de soi-même comme soi-même »[3]. À noter que ces distinctions sont acceptées au niveau conceptuel. Pour ce qui est de savoir si elles existent empiriquement ou métaphysiquement, le sujet devient plus épineux et débattu.

Ainsi, l'argument contre l'IA serait de dire que celle-ci, même dans la version idéale que nous avons imaginée, ne pourra jamais avoir de qualia. En effet, les capteurs visuels, comme des

[1] François Kammerer, « La conception illusionniste de la conscience phénoménale. Défis et perspectives », *Klēsis* 55 (2023), page 2. https://www.revue-klesis.org/pdf/klesis-55-03-françois-kammerer-conception-illusionniste-conscience-phenomenale-defis-perspectives.pdf.
[2] Kammerer, page 3.
[3] Kammerer, page 3.

caméras, détectent des ondes d'une certaine longueur d'onde et envoient sous forme de flux d'informations des ensembles de valeurs décimales afin de pouvoir former une matrice, une grille où chaque case représente un pixel contenant un de ces ensembles de valeurs que l'IA interprètera comme étant une certaine couleur. Cependant, tout ce dont l'IA dispose, ce sont ces valeurs numériques. Ce sont nous, les humains, qui lui avons préalablement indiqué à quelle couleur, ou plutôt à quel nom de couleur, correspond tel ensemble de valeurs. Devant un objet comme un ballon rouge, l'IA pourrait nous indiquer que l'objet est rouge, elle reconnaîtrait que l'objet est rouge, elle saurait que la couleur d'objet est composée de tel ensemble de valeur et qu'on lui a indiqué auparavant que cet ensemble correspond au « rouge ». Cependant, nous pouvons douter du fait que l'IA ait, comme nous, des ressentis phénoménaux, de ressentis qualitatifs, l'effet que cela fait de voir du rouge. En effet, elle ne semble se limiter qu'à la reconnaissance de valeur de longueur d'onde, rien de plus. Or, nous pouvons légitimement douter du fait que connaître la longueur d'onde d'une couleur nous permette d'avoir l'expérience phénoménale de celle-ci. Ainsi, même dans des conditions idéales, il semblerait que l'IA n'ait pas de conscience phénoménale. C'est un aspect important de la conscience que l'on échoue à reproduire. Cela implique que l'IA ne pourrait pas non plus ressentir de la douleur par exemple. Ce sont tout autant de choses que l'on considère comme importantes pour l'esprit humain.

Il y aurait alors un obstacle infranchissable qui se dresserait face à l'IA même si elle était dans des conditions idéales. En effet, comment parvenir à recréer cette conscience phénoménale ? Nous n'arrivons pas à communiquer sur les qualia et ne savons donc pas si nous voyons réellement les mêmes couleurs par exemple. Il y a un aspect mystérieux concernant la conscience phénoménale ce qui provoque de nombreux débats et peut amener certains penseurs à avoir une approche dualiste de l'esprit. La conception dualiste de l'esprit considère que l'esprit est constitué de deux substances, une matérielle qui est le cerveau, et l'autre immatérielle, impossible à réduire en des termes scientifiques et dont l'origine est à discuter. Comme penseur dualiste nous pouvons bien évidemment penser à Descartes mais également à des individus contemporains comme David Chalmers, les deux estimant que les états mentaux sont ontologiquement distincts des états d'un système physique. Cette conception dualiste apporte alors une réponse au gouffre explicatif présent entre la description scientifique des couleurs (le fait que ce soit une onde) et l'effet que cela fait de voir une couleur. C'est l'esprit en tant que substance immatérielle qui est responsable de la conscience phénoménale et des qualia. Cependant cette conception de l'esprit pose également des

problèmes de taille. En effet, comment communique cette substance immatérielle avec la substance matérielle ? Et comment se forme cette substance immatérielle ? D'où provient-elle ? Ce sont là autant de questions d'une importance capitale qui demeurent sans réponses. Ce manque de réponses a d'ailleurs provoqué, à de nombreuses reprises, le rejet de certains penseurs pour cette théorie.

Supposer que la conscience phénoménale est inaccessible à toute tentative de réduction en des termes physiques ou computationnels implique d'adhérer à une théorie de l'esprit au moins semblable au dualisme et à rejeter les théories matérialistes de l'esprit comme le physicalisme ou le fonctionnalisme étant donné que celles-ci sont incapables de donner une réponse au gouffre explicatif. Cependant, il existe des théories qui seraient en mesure de donner les outils au matérialisme pour surmonter ce problème, ce qui pourrait donner une chance à l'IA. Nous pouvons citer, par exemple, la thèse de l'intentionnalisme selon laquelle tous les états mentaux ont une intentionnalité. La conscience phénoménale possède alors une intentionnalité de l'expérience phénoménale. Ainsi, par exemple, quand nous avons peur, c'est parce que nous avons peur de quelque chose. Notre peur porte sur un objet et ce serait le cas de tous les autres types d'expériences phénoménales comme voir des couleurs, avoir faim, etc. Ce faisant, la question de la conscience phénoménale peut être réduite à la question de l'intentionnalité. Il serait alors surement possible pour une IA de reproduire de telles capacités cognitives. Mais on peut faire encore plus radical que ça. Il existe par exemple une théorie qui va jusqu'à nier l'existence de la conscience phénoménale. Cette théorie, c'est l'illusionnisme.

L'illusionnisme est « la thèse d'après laquelle la conscience phénoménale n'existe pas, mais semble simplement exister »[1]. Cette théorie se compose de deux thèses principales, une négative et l'autre positive. La thèse négative est que la conscience phénoménale n'existe pas. Ainsi, selon Kammerer :

> Grâce à sa thèse négative, cette conception dissout le problème difficile et supprime le gouffre explicatif d'une manière particulièrement incontestable. S'il n'y a pas réellement d'expériences phénoménales, il n'y a rien ici à expliquer de particulier. Le fait que les expériences phénoménales soient particulièrement difficiles à expliquer à partir des propriétés matérielles du cerveau n'implique donc pas une déficience quelconque des explications matérialistes.[2]

Cela serait donc un avantage de taille pour le matérialisme de l'esprit. Il n'y aurait plus besoin de rendre compte de la conscience phénoménale qui est, initialement, l'un de ses principaux

[1] Kammerer, page 1.
[2] Kammerer, page 4.

obstacles. Mais alors, si la conscience phénoménale n'existe pas, comment cela se fait-il que nous ressentions de la douleur par exemple ? Nous ne pouvons pas simplement nier un phénomène que nous expérimentons constamment. Cela nous amène à la seconde thèse de l'illusionnisme, la thèse positive qui est que « la conscience phénoménale semble exister »[1]. Cela permet alors de répondre à cette objection évidente que nous pourrions faire. La conscience phénoménale n'existe pas mais elle semble exister. En somme, la conscience phénoménale est une illusion. Ainsi :

> Grâce à sa thèse positive, l'illusionnisme peut rendre compte de notre impression forte et persistante que les expériences phénoménales sont réelles. Il peut expliquer ainsi que tant de penseurs rationnels aient cru – faussement – que nous avons effectivement des expériences phénoménales, nécessitant une explication et créant un « problème difficile ».[2]

L'illusionnisme réussit alors un double exploit. D'un côté, il parvient à fournir les outils nécessaires au matérialisme de l'esprit pour conserver sa crédibilité et surmonter l'obstacle de la conscience phénoménale (en la niant) et de l'autre, il parvient à rendre compte du fait qu'il semble que nous ayons ce qui s'apparente à des expériences phénoménales. Cette théorie permettrait alors à l'IA de franchir également cet obstacle. Elle n'a pas besoin d'avoir une illusion d'expérience phénoménale. Elle n'a pas besoin d'avoir une conscience phénoménale, puisque celle-ci n'existe tout simplement pas. L'IA peut donc se focaliser sur la reproduction des autres mécanismes de l'esprit humain.

Cependant, cette thèse reste assez marginale et dépréciée. La thèse négative niant la réalité de la conscience phénoménale provoque de vifs rejets puisqu'elle semble contre intuitive. Elle semble même nier une évidence. Pour Strawson, l'illusionnisme est la « thèse la plus stupide jamais soutenue »[3]. Et pour cause, cette théorie est effectivement déroutante au premier abord même si elle permet de résoudre un problème épineux concernant la conscience. Il y a, cependant, d'autres problèmes qui sont soulevés. Il faut notamment que cette théorie parvienne à devenir acceptable aux yeux de tous et, pour ce faire, il faut qu'elle parvienne à dépasser l'évidence de la conscience phénoménale en montrant son caractère fallacieux par exemple.

Un autre aspect n'est pas totalement clair concernant la nature même de l'illusion. D'après Kammerer :

> […] les partisans de l'illusionnisme comprennent cette apparence fallacieuse de conscience comme provenant de l'introspection, c'est-à-dire du processus cognitif spécifique par lequel

[1] Kammerer, page 4.
[2] Kammerer, page 5.
[3] Galen Strawson, « The Consciousness Deniers », *The New York Review of Books* (blog), 13 mars 2018.

> nous représentons nos propres états mentaux quand nous les avons, en première personne. L'idée, dans cette perspective, serait que lorsque nous introspectons, nous nous représentons comme entrant dans certains états mentaux – des expériences phénoménales – qui n'existent pas réellement.[1]

C'est donc comme si l'introspection nous trompait, nous faisait croire que nos états mentaux ont des propriétés phénoménales qui n'existent pas réellement. L'introspection, que l'on ne parvient pas tout à fait à saisir elle aussi, nous rendrait automatiquement victime de l'illusion quand bien même l'on sait que c'est une illusion. Ainsi :

> De même que nous restons sensibles à de nombreuses illusions d'optique, comme l'illusion de Müller-Lyer, lorsque nous savons que ce sont des illusions, nous restons sensibles à l'illusion de conscience phénoménale (il nous semble toujours avoir des expériences phénoménales), même si nous sommes convaincus de la vérité de l'illusionnisme.[2]

Tout comme pour les trompe-l'œil où notre cerveau nous fait voir des choses qui ne sont pas réellement telles qu'elles nous paraissent, notre cerveau nous trompe systématiquement en nous faisant croire que nous avons des expériences phénoménales. Via cette analogie avec les illusions d'optique, nous parvenons alors à mieux cerner ce que sont ces illusions d'expériences phénoménales. Cependant Kammerer ajoute que :

> […] cette apparence fallacieuse de conscience phénoménale que postule l'illusionnisme doit évidemment être comprise elle-même comme non-phénoménale – comme ne consistant pas elle-même dans une expérience phénoménale – sous peine de contradiction.[3]

Ce point-là rend la compréhension de l'illusion beaucoup plus difficile. En effet, pour une illusion d'optique par exemple, cela consiste en le fait de voir quelque chose différemment de ce qu'elle est vraiment. Mais, nous avons tout de même une expérience visuelle. Dans le cas des expériences phénoménales, nous aurions une illusion qui n'est pas phénoménale, alors que l'expérience que nous croyons avoir l'est. Il y a donc une difficulté qui se présente. Il est évident que les illusionnistes sont obligés de dire que l'illusion n'est pas phénoménale (puisqu'il cherche à démontrer que la conscience et les expériences phénoménales n'existent pas). Mais alors comment, à partir du non phénoménal, nous créons une illusion de phénoménal ? L'effet que cela fait de voir du rouge, quand on voit un objet rouge, est bien là. Nous pouvons dire que cet « effet que cela fait » est effectivement illusoire, mais alors il devient difficile de dire que l'illusion ne transmet pas un « effet que cela fait ». L'illusion d'optique est tout de même un phénomène

[1] Kammerer, page 8.
[2] Kammerer, page 8-9.
[3] Kammerer, page 9.

optique, l'illusion sonore est tout de même un phénomène sonore. Si l'illusion est non-phénoménale, alors il faut savoir comment elle fait pour nous insuffler ce ressenti qualitatif fallacieux. Il manque quelque chose, le dernier chainon entre l'illusion non-phénoménale et l'expérience vécue, elle, comme phénoménale. Ne pas y répondre est une trop grande facilité. Nous pourrions nous contenter de dire que la conscience phénoménale est une illusion, illusion qui est pourtant non-phénoménale alors que, comme par magie, elle nous paraît être phénoménale. Nous pourrions nous arrêter à cet argumentaire pour permettre de nous débarrasser du problème de la conscience phénoménale. Or, il reste encore un gap, un autre gouffre à expliquer. Ce passage du non-phénoménal à l'impression de phénoménal est trop flou, un peu comme la question de l'interaction de l'âme avec le corps pour les dualistes. Cependant, nous pouvons nuancer nos analogies avec les illusions sonores et visuelles avec le cas de l'hallucination. En effet, nous pouvons concevoir une hallucination visuelle comme relevant d'un type d'état mental différent d'une perception véridique. Il y aurait alors une disjonction entre la perception véridique et l'hallucination. Si nous considérons les hallucinations de cette manière, nous aurions alors un exemple permettant d'appréhender l'illusion. L'illusion de la conscience phénoménale serait en quelque sorte une hallucination créée par notre cerveau. Mais, si nous nous refusons la disjonction entre hallucination et perception réelle, l'analogie avec l'illusion de la conscience phénoménale ne marche plus et nous ne comprenons pas comme procède l'illusion.

Nous pouvons également formuler un argument contre l'illusionnisme à partir du naturalisme modéré de McDowell. En effet, McDowell considère que l'évolution est une cause fondamentale de la formation de l'esprit humain tel qu'il est aujourd'hui. Si l'on suit cette logique alors les différentes capacités cognitives constitutives de notre esprit ont pu représenter à un moment donné un caractère déterminant pour la reproduction et/ou la survie. Les individus qui avaient à un moment donné certaines (ou toutes) de ces caractéristiques ont été avantagé et ont pu mieux se reproduire que les autres, perpétuant ainsi les traits qui les ont avantagés. Pour McDowell, l'esprit est apparu chez l'humain notamment via l'apparition de traits qui ont eu un avantage évolutif par exemple le langage qui est, pour lui, fondamental pour la formation de l'espace logique des raisons.

Nous pouvons alors supposer que la conscience phénoménale ait pu, elle aussi, apparaître dans un contexte évolutif. Par exemple pour des expériences phénoménales comme la douleur, il est très tentant de penser que cela ait pu représenter un avantage dans la lutte pour la survie. Fuir, ce qui nous fait mal et aller vers ce qui nous fait du bien semble avoir pour conséquence de favoriser notre

survie. De même pour les couleurs. Par exemple, parvenir à discerner un champignon marron d'un champignon rouge à points blancs peut être un avantage vital pour la survie de l'individu. Nous pouvons alors légitimement supposer que notre capacité à avoir des expériences phénoménales soit apparu dans un contexte de lutte pour la survie et que ce sont alors des états mentaux bien réels et non des illusions. En effet, nous avons besoin de pouvoir discriminer les couleurs. Il existe d'autres moyens de discriminer les couleurs que les expériences phénoménales mais, à défaut d'avoir un capteur ou un organe nous permettant d'afficher la valeur numérique de la longueur d'onde des couleurs, notre cerveau génère une expérience phénoménale associée à chaque couleur. Les couleurs n'existent pas vraiment dans les objets extérieurs, il est vrai, mais le ressenti qualitatif qui est la traduction par le cerveau de la longueur d'onde qui frappe l'œil, lui est réel. La conscience phénoménale existerait grâce à l'évolution et il n'y aurait aucune raison que cela ne soit juste une tromperie de notre cerveau.

D'autant plus que le concept de vie dont nous disposons, nous les humains, dépend des ressentis subjectifs que nous pouvons avoir. En effet, notre concept de vie est intimement lié aux expériences phénoménales car nous sentons que nous sommes vivants. Les expériences de douleurs, de plaisir et la peur de la mort, le stress du danger, sont autant de moyens de ressentir notre condition d'être vivant. Les êtres vivants sont ceux qui ressentent, qui éprouvent. La vie est une condition nécessaire pour les ressentis qualitatifs et la réciproque n'est pas vraie. Il y a bien des organismes considérés comme vivants, dont on doute (peut-être à tort) qu'ils n'ont pas d'expériences phénoménales. Néanmoins, pour des organismes complexes comme nous, les humains, nous avons des raisons de penser que la conscience que l'on a d'être vivant est lié au moins en partie à notre conscience phénoménale. En effet, pour McDowell, notre « je », notre conscience de nous-mêmes en tant que vivant dans un monde, n'est valable que si des expériences ont lieu en nous à travers le temps et, pour ne pas nier ce « je », il ne faut pas nier la réalité de ces expériences.

Enfin, le naturalisme modéré et le réalisme naturel, stipulent un rapport direct au monde. Nous nous développons, formons notre vision du monde via nos expériences vécues dans nos environnements respectifs. Il y a une transaction qui s'effectue entre nous et le monde qui se fait sans intermédiaire, sans écran. Ce rapport direct implique donc qu'il serait incohérent que les expériences phénoménales ne soient pas réelles. Cela rejoint l'argument sur l'avantage évolutif que ces expériences ont pu représenter. Il semble fondamental d'avoir ces expériences pour se

développer dans l'environnement, vivre des expériences et agir sur celui-ci. L'idée de transaction directe avec le monde semble nécessiter d'avoir des ressentis qualitatifs de celui-ci sans quoi nous ne pourrions pas nous développer correctement en son sein.

L'argument de la conscience phénoménale contre l'IA peut donc être critiqué mais cette critique n'est pas exempte de défauts elle aussi. D'autant plus que, dans le cadre de l'esprit conçu selon McDowell, cette critique semble peu vraisemblable. Cependant, d'autres théories en faveur de l'IA sont possibles. Nous avons discuté par exemple de l'intentionnalisme, théorie considérant la conscience phénoménale comme ayant une intentionnalité. Or, il s'avère que pour les partisans de l'approche téléosémantique de l'intentionnalité des états mentaux, comme Fred Dretske et Ruth Millikan par exemple, les états mentaux sont des états internes d'un système physique et sont covariants à des états de l'environnement. Donc, d'après cette approche, l'intentionnalité peut être réduite en des termes physiques, car les états mentaux sont des états internes d'un système physique, et causaux, car il y a une covariance avec des états de l'environnement. Ainsi, si nous considérons la conscience phénoménale comme un phénomène intentionnel, nous donnerions alors à l'IA les outils nécessaires à la reproduction d'une conscience phénoménale. Encore faut-il accepter une telle prémisse.

3. L'introspection et l'inférentialisme

En lien avec l'illusionnisme, les défenseurs de l'IA pourraient critiquer l'introspection afin de maintenir la possibilité de reproduction de nos capacités cognitives. Nous avons vu avec Kammerer que l'introspection est essentielle pour faire des expériences phénoménales. C'est par l'introspection que nous avons ces expériences. L'introspection c'est le « processus cognitif spécifique par lequel nous représentons nos propres états mentaux quand nous les avons, en première personne »[1]. Il y a un aspect qui nous échappe concernant l'introspection, cette faculté permettant de se sonder soi-même via des moyens qui lui sont propres. Mais, tout aussi mystérieuse soit-elle, cette capacité ne semble exister à nos yeux que par l'évidence même que nous l'utilisons quotidiennement.

[1] Kammerer, page 8.

Une stratégie pour permettre à l'IA de simuler cette capacité mystérieuse et paraissant irréductible en des termes physiques ou computationnels serait de dire que l'introspection n'est pas ce qu'elle paraît être. Il s'avère justement qu'il existe une théorie se nommant l'inférentialisme qui stipule que l'introspection n'est pas un processus interne mystérieux mais qu'il s'agit tout simplement de l'application sur nous-mêmes des mêmes procédés d'inférences utilisés pour déterminer les états mentaux des individus avec lesquels on interagi. En effet, nous faisons des inférences sur les personnes que nous côtoyons. Nous interprétons des éléments comme les comportements des personnes afin de supposer une émotion, une envie, un état mental. L'inférentialisme estime donc que nous faisons la même chose sur nous-mêmes lorsque nous nous introspectons, c'est-à-dire que nous interprétons des divers éléments comme des signes comportementaux par exemple afin d'en déduire un état mental. Pour l'approche inférentialiste, l'introspection n'est plus un processus interne direct mais plutôt indirect. Ainsi, pour les inférentialistes :

> […] la connaissance de soi repose sur des inférences interprétatives fondées sur les données comportementales. Ces inférences causent les croyances introspectives, et elles les justifient, ce qui veut dire que le degré de justification d'une croyance introspective peut dépendre des données comportementales qui en constituent la base empirique.[1]

Cette théorie met donc fin à une distinction entre deux types de connaissances, l'une directe et non-inférentielle et l'autre indirecte et inférentielle.

Il reste cependant à déterminer si l'accès indirect concerne seulement certains types d'états mentaux ou tous les états mentaux. Sur ce point-là, le débat est ouvert. Il semble cependant qu'il soit admis que l'introspection ait recours à la voie de l'inférence pour les états mentaux dispositionnels, c'est-à-dire les « état mentaux qui ne sont pas occurrents au moment auquel le sujet exerce sa capacité d'introspection pour les connaître »[2]. Par exemple, un état mental dispositionnel c'est « savoir si je suis en faveur ou en défaveur de la peine de mort par introspection »[3], alors qu'un état mental occurrent c'est « savoir par introspection si je suis bien en train de juger que la peine de mort est injuste »[4]. Dans l'article « Telling more than we can know: Verbal reports on

[1] Pascal Ludwig et Matthias Michel, « Introspection (A) », in *L'encyclopédie philosophique*, 2017, https://encyclo-philo.fr/item/130.
[2] Ludwig et Michel.
[3] Ludwig et Michel.
[4] Ludwig et Michel.

mental processes »[1], Richard E. Nisbett et Timothy D. Wilson estiment que l'introspection nous amène à en dire plus que ce que nous savons, c'est-à-dire que l'introspection nous amène à confabuler. Ainsi :

> Dans leur protocole expérimental, les deux expérimentateurs disposaient de gauche à droite quatre paires de collants (de gauche à droite : A, B, C, D) exactement identiques et demandaient à des consommateurs de les évaluer. L'hypothèse de Nisbett et Wilson était que les participants préfèrent systématiquement les collants qui se trouvent le plus à droite (la dernière à être au centre de l'attention des participants), et cette hypothèse s'est trouvée validée. Les sujets choisissaient la paire de collants A (la plus à gauche) à 12%, B à 17%, C à 31% et D à 40 %. […] À la suite de leur choix, il était demandé aux participants d'expliquer les raisons qui les conduisaient à préférer la paire D à une autre paire de collants. Ces derniers justifiaient leur choix en expliquant aux expérimentateurs que la paire de collants D était plus élastique, ou plus douce, mais aucun ne mentionnait que la position avait joué un rôle quelconque dans leur préférence. Lorsqu'on leur demandait s'ils pensaient que la position des collants avait influencé leur décision, tous les participants répondaient que ce n'était pas le cas. Les participants justifiaient donc leur choix par des qualités qui n'existaient pas effectivement et qui, par conséquent, n'avaient pas pu les conduire à un tel choix. L'introspection sur les raisons de leurs préférences menait donc les sujets à s'attribuer un jugement qui n'avait aucun rôle causal effectif dans la formation de ces préférences. Au contraire, la différence qui avait effectivement joué un rôle dans le choix de la paire de collants, c'est-à-dire leur position, n'était jamais mentionnée par les participants.[2]

Il semble que le processus introspectif que les personnes utilisent pour justifier leurs choix semble plus être une interprétation indirecte de leurs comportements qu'un sondage direct de leurs états internes. Or, ce processus étant indirect et se faisant avec les données qui leur sont accessibles, les individus confabulent et « en racontent trop ».

À ces états mentaux dispositionnels, que l'on appelle aussi attitudes propositionnelles non-occurrentes, on leurs oppose, comme nous l'avons vu, les états mentaux occurrents, que l'on appelle aussi attitudes propositionnelles occurrentes. Il semble que nous en savons davantage sur nos propres attitudes propositionnelles occurrentes que sur celles d'autrui. En effet, je vais avoir tendance à savoir plus facilement si je suis bien en train de juger que la peine de mort est injuste que d'autres individus. Et inversement, il sera pour moi beaucoup plus compliqué de savoir si les autres sont en train de juger que la peine de mort est injuste. Il semble donc ici que l'accès direct par l'introspection à ces états mentaux soit à privilégier.

Et pourtant, certains penseurs comme Peter Carruthers pensent qu'ici aussi, nous avons recours aux inférences. En effet, pour Carruthers, « la métacognition est simplement le résultat du

[1] Richard E. Nisbett et Timothy D. Wilson, « Telling More Than We Can Know: Verbal Reports on Mental Processes », *Psychological Review* 84, nº 3 (1977): 231-59.
[2] Ludwig et Michel, « Introspection (A) ».

retournement de nos capacités de lecture de l'esprit sur nous-mêmes »[1]. Ainsi, la métacognition, c'est-à-dire l'accès à nos attitudes propositionnelles en première personne, n'est que l'application de notre capacité à attribuer des états mentaux à autrui sur nous-mêmes. Selon Carruthers :

> […] il ne devrait pas y avoir de dissociation entre la lecture mentale et la métacognition. […] une seule faculté est impliquée dans les deux formes d'activité, utilisant essentiellement les mêmes données, qui sont toutes de nature perceptive ou quasi-perceptive.[2]

Il n'y aurait donc pas de différences entre l'accès aux attitudes propositionnelles en première personnes et en troisième personne. Pour Carruthers, nous n'aurions pas d'accès privilégié à nos états mentaux.

Mais alors comment expliquer cette facilité de connaissance sur nos propres états mentaux occurrents par rapport aux autres individus qui nous observeraient ? Pour Carruthers, c'est parce que nous avons tout de même un accès direct à des informations, des contenus conscients. Pour lui :

> On admettra que l'accès que nous avons à nos attitudes inconscientes (qu'elles soient ou non exprimées par la parole ou d'autres formes d'imagerie) est toujours interprétatif, comme nous l'avons expliqué plus haut. Mais on pourrait affirmer que le flux de paroles intérieures et d'autres formes d'imagerie est constitutif d'un type distinct de mentalité (consciente). Il est certain que ces événements ne sont pas épiphénoménaux, mais qu'ils apportent souvent une contribution causale importante à la pensée et au comportement ultérieurs. Et l'on pourrait dire que de tels événements sont régulièrement disponibles pour l'introspection.[3]

Ces informations, comme le flux de paroles intérieurs, sont des événements dont on a un accès privilégié et direct. Carruthers se retrouve obligé de le concéder, il n'a visiblement pas le choix puisqu'il dit :

---

[1] Peter Carruthers, « How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition », *Behavioral and Brain Sciences* 32, nº 2 (2009): 121-38, page 123 : « […] metacognition is merely the result of us turning our mindreading capacities upon ourselves ».

[2] Carruthers, page 123 : « […] there should be no dissociations between mindreading and metacognition. This is because there is just a single faculty involved in both forms of activity, utilizing essentially the same inputs, which are all perceptual or quasi-perceptual in character ».

[3] Carruthers, page 133 : « It would be allowed that the access that we have to our unconscious attitudes (whether or not they get expressed in speech or other imagery) is always interpretative, as argued above. But it might be claimed that the stream of inner speech and other forms of imagery is constitutive of a distinct kind of (conscious) mentality. Certainly such events aren't epiphenomenal, but often make an important causal contribution to subsequent thought and behavior. And it might be said that such events are routinely available to introspection ».

> […] nous devons maintenant admettre que le système de lecture de l'esprit dispose d'informations lorsqu'il attribue des états mentaux au soi, informations auxquelles il n'a jamais accès lorsqu'il attribue des états mentaux à d'autres personnes.[1]

Pour Carruthers, cet avantage que nous avons pour connaître nos propres états mentaux occurrents par rapport aux autres individus s'explique donc parce que nous n'avons pas uniquement recours à l'inférence sur notre comportement au moment de l'introspection mais aussi à toutes sortes de contenus sensoriels que seuls nous pouvons avoir et connaître. Le caractère entièrement inférentiel de l'introspection ne semble donc pas aller de soi concernant les états mentaux occurrents.

Cependant, Eric Schwitzgebel pense que, même dans le cas de ces contenus sensoriels conscients dont nous aurions un accès direct, il y a des raisons d'être sceptique. Nous prenons souvent exemple des couleurs. En effet, « voir du vert » semble être une expérience consciente inévitablement véridique et infaillible (sauf pour les illusionnistes comme nous l'avons vu). Schwitzgebel propose alors de se poser d'autres questions que « quelle couleur je vois ? ». Nous pouvons, par exemple, nous demander :

> Où s'arrête notre champ de vision exactement ? Peut-on voir la couleur à sa périphérie ? Et si nous imaginons un paysage, les couleurs sont-elles vives, sont-elles même présentes ? Quel est le niveau de détail de cette image mentale ? De même, les émotions ont des contours très mal définis ; en éprouve-t-on toujours ? Peut-on comparer leurs intensités ? Ai-je toujours une humeur particulière ? Ai-je une idée claire de ce que c'est pour moi que d'être irritable, ou à l'aise ?[2]

Les réponses à ces questions sont moins évidentes que celles que l'on peut donner à la question « quelle couleur je vois ? ». Schwitzgebel nous dit alors :

> Je ne peux évidemment pas imposer une réponse particulière à ces questions. Je ne peux que vous inviter à partager mon sentiment intuitif d'incertitude. Et il ne me semble pas que le problème ici soit simplement linguistique, qu'il s'agisse simplement de trouver les mots justes pour décrire une expérience connue dans ses moindres détails, ou qu'il s'agisse simplement d'une question conceptuelle ou théorique consistant à déterminer quels aspects d'une phénoménologie bien connue sont considérés à juste titre comme des aspects de l'émotion.[3]

---

[1] Carruthers, page 124 : « […] we now have to concede that the mindreading system does have available to it information when attributing mental states to the self that it never has access to when attributing mental states to others ».

[2] Ludwig et Michel, « Introspection (A) ».

[3] Eric Schwitzgebel, « Self-Ignorance », in *Consciousness and the Self*, éd. par JeeLoo Liu et John Perry, 2012, page 188 : « I cannot, of course, force a particular answer to these questions. I can only invite you to share my intuitive sense of uncertainty. And it does not seem to me that the problem here is merely linguistic, merely a matter of finding the right words to describe an experience known in precise detail, or merely the conceptual or theoretical matter of determining which aspects of a wellknown phenomenology are properly regarded as aspects of emotion ».

Nous pouvons sûrement donner des réponses à ces questions mais il est indéniable que celles-ci sont incertaines. Le scepticisme est donc permis même pour ces contenus sensoriels conscients car, pour Schwitzgebel :

> Vous pouvez avoir quelques connaissances approximatives sur votre expérience actuelle, mais si vous essayez de remonter plus loin que quelques secondes, de généraliser, d'articuler quelques détails ou de discerner des caractéristiques structurelles, même modérément importantes, de votre expérience, vous vous tromperez rapidement.[1]

Cela qui l'amène alors à conclure que :

> Nous vivons dans des cocons d'ignorance, en particulier lorsque notre conception de nous-mêmes est en jeu. L'accent philosophique mis sur l'importance de notre connaissance de soi ne tient pas compte des choses les plus importantes.[2]

Ainsi, peut-être bien que nous n'avons aucun accès direct sur ce qui se passe en nous. Si l'on est inférientialiste et si l'on est aussi radical que Schwitzgebel en allant jusqu'à douter de la fiabilité des contenus sensoriels conscients, il semble qu'il soit possible de théoriser un modèle introspectif reproduisible par l'IA. En effet, notamment car nous pourrions alors résoudre le problème de la conscience phénoménale. Si l'introspection est uniquement un processus indirect, alors nous ne pouvons pas soutenir le fait d'avoir des expériences phénoménales qui sont par définition directes. L'inférentialisme constitue alors un bon argument pour l'illusionnisme et la conception matérialiste de l'esprit. Ainsi, si une IA arrive à se baser uniquement sur les mêmes types d'inférences utilisées pour déterminer l'état mental des personnes qu'elle voit afin de déterminer ses propres états mentaux, dispositionnels ou occurrents, alors cela voudra dire qu'elle n'aura surement pas besoin de reproduire la conscience phénoménale puisqu'il y aurait de bonnes raisons de penser que celle-ci n'est qu'une illusion.

Il semble également que l'argument pragmatique et évolutionniste soit à l'avantage des inférentialistes. En effet, d'après Carruthers, « si la lecture de l'esprit et la métacognition relèvent de deux mécanismes cognitifs (ou plus), il est évident que l'émergence de chacun d'entre eux doit faire l'objet d'une histoire évolutive distincte »[3]. Donc, si l'on considère qu'il y a d'une part

[1] Schwitzgebel, page 190 : « You may know a few rough things about your current experience, but try to extend your knowledge back more than a few seconds, try to generalize, try to articulate a bit of detail, or try to discern even moderately large structural features of your experience, and soon you will err ».
[2] Schwitzgebel, page 197 : « We live in cocoons of ignorance, especially where our self-conception is at stake. The philosophical focus on how impressive our self-knowledge is gets the most important things backwards ».
[3] Carruthers, « How We Know Our Own Minds », page 128 : « […] if mindreading and metacognition are subserved by two (or more) cognitive mechanisms, then plainly there should be a distinct evolutionary story to be told about the emergence of each ».

l'introspection et d'autre part la déduction des états mentaux des personnes par inférences sur leurs comportements, alors il faut proposer deux histoires évolutives distinctes permettant de justifier leurs apparitions. Or, cela n'est pas le cas pour la théorie inférentialiste :

> Du point de vue de l'hypothèse selon laquelle « la lecture de l'esprit est antérieure » [c'est-à-dire que l'introspection consiste en des inférences que l'on fait sur nous-même], il n'est pas nécessaire de raconter une histoire distincte. Puisque la métacognition [l'introspection], selon ce point de vue, résulte du retournement des capacités de lecture de l'esprit sur soi-même, son émergence sera un sous-produit de l'évolution de la lecture de l'esprit.[1]

C'est effectivement un argument en faveur de l'inférentialisme. Une seule histoire évolutive peut permettre de rendre compte de l'apparition de ces deux facultés.

Mais peut-être que la voie la plus simple n'est pas la plus pertinente. En effet, il semble plus évident, plus pratique d'avoir un accès direct à nos états intérieurs afin de faire le meilleur exercice introspectif possible. Cela a très certainement eu moins d'importance que les mécanismes automatiques comme les réflexes pour la survie mais, si nous avons la capacité d'introspection, alors peut-être qu'il y a eu à un moment donné de notre histoire un intérêt à porter une réflexion sur nous-mêmes, à nous sonder afin de corriger des comportements ou éviter des événements qui ne nous avantagent pas. Il paraît donc plus pertinent que l'introspection soit un processus direct, dont on peut avoir confiance, plutôt qu'un processus indirect qui peut nous amener à confabuler. La confabulation semble être un phénomène qui n'améliore pas nos chances de survie, au contraire. Si le processus introspectif direct n'existe pas, il serait alors surprenant que notre propre esprit nous fourvoie à ce point.

Notre scepticisme vis-à-vis du fait que l'introspection n'est pas un processus direct est solidifié par les concessions que certains philosophes font sur les expériences sensorielles conscientes comme c'est le cas pour Carruthers. Même si l'application d'une telle théorie inférentialiste semble possible, il ne faut pas oublier qu'il y a quelques tensions qui subsistent concernant les états mentaux occurrents où Carruthers admet l'existence d'expériences conscientes et directes intervenant dans le processus introspectif. Les connaissances acquises de manière directe par processus introspectif semblent être nécessaires ne serait-ce qu'en partie pour assurer la pleine efficacité de l'introspection. Ces expériences directes posent alors un gros problème à l'IA cherche à avoir de l'introspection car elle ne semble pas pouvoir les reproduire.

[1] Carruthers, page 128 : « From the perspective of a "mindreading is prior" account, no separate story needs to be told. Since metacognition, on this view, results from turning one's mindreading capacities upon oneself, its emergence will be a by-product of the evolution of mindreading ».

La conscience phénoménale et l'introspection semblent donc être des limites de taille pour l'IA, même dans des conditions idéales pour elle. Il existe certes des théories qui pourraient permettre à l'IA soit de contourner ses limites soit de les franchir mais elles ne sont pas parfaites et posent, elles aussi, des problèmes, ce qui pose un doute sur leurs viabilités. Il est vrai que l'aspect mystérieux des expériences phénoménales, de « l'effet que cela fait », et de l'introspection peut déranger de nombreuses personnes et surtout les matérialistes. Mais, pour McDowell, cet aspect mystérieux n'est pas un problème. La conscience phénoménale et l'introspection sont au moins en partie le fruit de notre seconde nature. Ce sont des capacités naturelles elles aussi quand bien même nous ne parvenons pas à les décrire via des termes physiques ou computationnels. Il reste néanmoins que, malgré nos arguments, il est possible de postuler que l'IA pourrait reproduire l'introspection si l'on tient pour vrai l'inférentialisme et si l'on mobilise l'illusionnisme de la conscience phénoménale pour résoudre le problème des expériences conscientes et directes que Carruthers concède.

4. La créativité et l'imagination

« L'intelligence artificielle va-t-elle tuer les artistes ? »[1]. Voici le titre d'un article de France info écrit par Pierre Godon et datant du 21 janvier 2023. Une panique semble être présente dans le monde de l'art ces dernières années et la cause de celle-ci est le progrès fulgurant des différentes IA génératives. En effet, aujourd'hui, l'IA est capable de générer en quelques minutes voire quelques secondes des images, des animations ou de la musique via des indications textuelles. On lui indique ce que l'on veut et elle génère une œuvre qui répond à ces critères. Cette capacité à générer des images ou des musiques crédibles et sophistiquées s'est accrue de manière extrêmement rapide ces dernières années. Il est désormais possible de demander à une IA de réaliser une peinture numérique très belle esthétiquement et répondant à nos critères gratuitement ou pour peu cher comparés aux prix des artistes. Ainsi, il est devenu très attrayant d'avoir recours à l'IA pour générer une image dont nous aurions besoin pour illustrer quelque chose, pour avoir une direction artistique pour son entreprise ou juste pour avoir un joli fond d'écran d'ordinateur

[1] Pierre Godon, « L'intelligence artificielle va-t-elle tuer les artistes ? », *Franceinfo*, 21 janvier 2023, https://www.francetvinfo.fr/culture/bd/enquete-franceinfo-l-intelligence-artificielle-va-t-elle-tuer-les-artistes_5610134.html.

plutôt que de faire appel à un artiste et le payer. L'attrait d'un service rapide et gratuit est important pour ceux qui ont des besoins occasionnels d'illustrations visuelles ou sonores. Comme nous pouvons le voir dans un article de France info :

> Un illustrateur brésilien a raconté sur Twitter avoir reçu en début d'année un mail de l'éditeur pour lequel il travaillait, lui annonçant la fin de son contrat et son remplacement par une machine qui ne coûte guère plus qu'un peu d'électricité. De nombreuses sociétés (des éditeurs français, une marque de biscuits du groupe Barilla en Italie…) ont déjà eu recours à des images générées de A à Z par IA. Certains en l'affichant, d'autres sous le manteau.[1]

Cela met donc en danger les nombreux artistes qui vivent de commandes que leur passent d'autres personnes. Il faut aussi mentionner l'apparition de nombreuses personnes se faisant passer pour des artistes accomplis mais qui ne font que générer des œuvres par le biais de l'IA. Une concurrence est donc apparue et elle est très rude, surtout que les œuvres générées par IA tendent à devenir de moins en moins discernables des œuvres humaines.

Mais cela va même plus loin. Même dans les grandes entreprises produisant des œuvres visuelles ou sonores, il y a une crise. De grandes entreprises comme Disney ou Ubisoft utilisent d'ores et déjà l'IA dans leurs productions. À noter que pour Ubisoft, qui est un éditeur et producteur de jeux vidéo, l'intelligence artificielle est utilisée depuis bien longtemps. En effet, fréquemment dans les jeux vidéo nous avons affaire à des personnages non joueurs, des PNJ, qui sont animés par l'intelligence artificielle afin de d'ajouter de la « vie » et des interactions possibles dans l'univers que parcourt le joueur. Or, ici, nous parlons d'utilisation de l'IA dans les processus de conception du jeu comme la création du scénario, des décors ou des dialogues. Pour l'instant, cette utilisation des IA garde une place minoritaire dans le processus de création, mais elle pose tout de même la question de l'intérêt de garder des humains plutôt que de les remplacer par une machine moins couteuse.

Ainsi, cette crise dans le monde de l'art part du principe que l'IA peut créer tout aussi bien que l'humain. En ce sens, beaucoup considèrent l'IA comme étant créative, aussi créative qu'un artiste par exemple. Or il s'avère que la notion de créativité est souvent mise en lien avec l'imagination. En effet, il semble naturel de dire qu'une personne créative est une personne qui a de l'imagination. Même si la notion d'imagination n'est pas forcément mobilisée dans la définition de la créativité, il semble tout de même indéniable qu'il y ait un lien entre les deux. Cela nous amène alors à aborder une autre limite que nous pouvons poser à l'IA. Selon nous, l'IA n'est pas

[1] Godon.

capable d'imaginer contrairement à nous, les humains. Elle ne serait donc pas créative non plus. Ce serait là une limite infranchissable pour l'IA, même si elle se trouve dans des conditions idéales. L'imagination ne semble appartenir à première vue qu'aux humains et elle leur a permis de faire de nombreux exploits, découvertes et chefs d'œuvres. Cela est même appuyé, comme nous l'avions vu, par Putnam, qui considère que l'imagination est une capacité qui nous est propre et qui n'est pas réductible en des termes physiques et computationnels. Rappelons-nous, dans le chapitre 3, nous avions dit que, selon Putnam, l'imagination est la capacité à faire des sauts conceptuels, à se projeter dans de nouvelles manières de penser afin de comprendre le monde différemment, d'acquérir de nouvelles significations, de faire des associations conceptuelles que l'expérience sensible ne nous aurait jamais permise de faire ou d'avoir. L'imagination c'est excéder, ne serait-ce qu'un peu, son réseau de croyances pour en acquérir de nouvelles ou créer de nouveaux liens entre celles déjà existantes. Via l'imagination, nous pouvons créer de nouvelles choses et, pour Putnam, c'est une capacité qui est propre aux humains. Elle ne semble pas vraiment descriptible, théorisable, ce qui nous donne une bonne raison de penser que l'IA ne pourra jamais la reproduire.

Il y a cependant un contre-argument à la pensée de Putnam se basant sur des faits empiriques. Le contre-argument qui se développe actuellement est que l'IA est capable de créer tout aussi bien que nous, les humains, qui avons la capacité d'imagination. Avec les progrès fulgurants de l'IA ces dernières années, la génération d'images, de vidéos ou de musique n'a jamais été aussi simple. À partir de simples indications textuelles, l'IA est capable de générer une image qui répond aux critères qu'on lui indique. De ce fait, l'IA produit, l'IA crée. Si nous comprenons naïvement la créativité comme le fait de produire quelque chose de nouveau, alors il faut reconnaître que tous les générateurs d'images, d'animations ou de musiques font aussi preuve de créativité. Or, nous avions également supposé un lien évident qui existe entre la créativité et l'imagination. Une personne créative est une personne qui a de l'imagination. Le contre-argument amène alors à conclure que si les IA font preuve de créativité, alors peut-être font-elles aussi preuve d'imagination. Les progrès fulgurants de l'IA générative de ces dernières années tendent donc à affaiblir, voire à réfuter, notre argument faisant de l'imagination une limite pour l'IA.

Mais, nous pouvons répondre à ce contre-argument. Peut-être bien que nous avons mal compris ce que c'est qu'être créatif. En effet, pour Alexander Bird et Alison Hills, notre conception

naïve de la créativité est erronée. Dans leur article « Against Creativity »[1], Bird et Hills présentent la définition commune de la créativité comme la disposition d'un individu :

1. D'avoir des idées nouvelles (originalité)
2. Qui ont de la valeur ; ou produire des objets qui ont de la valeur (valeur)[2]

Si l'on se réfère à cette définition, l'IA est créative comme nous le disions juste avant. Pour réaliser des images ou des musiques nouvelles qui correspondent à un cahier des charges précis, elle doit s'inspirer et combiner des créations déjà existantes. Elle s'inspire afin de produire ce que nous pourrions considérer comme étant des idées nouvelles. D'autant plus que les indications données par l'utilisateur peuvent être très vagues comme « invente un univers cyberpunk ». L'IA va donc devoir proposer une idée originale et cette idée a une valeur, elle est utile. Nous comprenons ici le terme valeur (*valuable*) dans le sens d'utile. Et en effet, l'IA répond aux demandes des utilisateurs. Elle produit un objet qui a un but, une utilité.

Cependant, pour Bird et Hills, cette définition commune de la créativité est fausse. D'abord, ils considèrent que la notion de valeur (ou d'utilité) n'a pas à être prise en compte. En effet, pour eux, « la créativité peut produire des objets sans aucune valeur »[3]. En réalité, pour que la créativité crée quelque chose de valeur, « elle doit généralement être associée à un bon jugement et à une tradition de travail propice à la production d'objets de valeur »[4]. Ensuite, ils retiennent l'idée d'originalité et la développent en ajoutant d'autres conditions qui lui sont liées mais que l'on ne retrouve pas forcément dans la définition commune de la créativité. Ainsi, pour eux, la créativité c'est la disposition d'un individu :

1. D'avoir des idées nouvelles (originalité)
2. Qui sont générées par l'utilisation de l'imagination (imagination)
3. Et qui sont nombreuses et variées (fertilité),
4. Et de mener ces idées à leur terme (motivation).[5]

[1] Alison Hills et Alexander Bird, « Against Creativity », *Philosophy and Phenomenological Research* 99, nº 3 (2019): 694-713.
[2] Hills et Bird, page 694 : « 1. To have novel ideas (originality)
2. Which are valuable; or produce objects which are valuable (value) ».
[3] Hills et Bird, page 695 : « creativity can produce objects without value of any kind ».
[4] Hills et Bird, page 695 : « it typically needs to operate in conjunction with good judgment and a propitious tradition of work that is itself conducive to the production of valuable objects ».
[5] Hills et Bird, page 695 : « 1. To have novel ideas (originality)
2. Which are generated through use of the imagination (imagination)

Cette nouvelle définition change tout. Nous pouvons accepter qu'une IA ait des idées nouvelles quand elle crée une image ou une musique et nous pouvons aussi accepter que ces idées soient nombreuses et variées. En effet, le processus de création des IA génératives est tel qu'il permet de former plusieurs images différentes à partir des mêmes indications textuelles. Il y a donc théoriquement une infinité de possibilités qui lui sont accessibles et elle en choisit une afin de la proposer à l'utilisateur. Cependant il paraît plus complexe de concéder les points 2 et 4.

Pour Bird et Hills, l'imagination c'est :

> […] une capacité à produire un type particulier de représentation mentale, mais elle est largement considérée comme ayant des formes très différentes et sa nature fait l'objet d'importantes controverses. Dans la mesure du possible, nous sommes libéraux quant aux types d'imagination qui peuvent jouer un rôle dans la créativité. Par exemple, l'imagination peut produire des idées nouvelles de manière délibérée (comme lorsqu'un scientifique passe beaucoup de temps à essayer de développer une théorie, ou qu'un écrivain édite et réécrit son travail). L'imagination peut également générer des idées spontanément (comme lorsqu'un scientifique est soudainement frappé par une hypothèse, ou un écrivain par une nouvelle image).[1]

Cette conception de l'imagination se rapproche de celle de Putnam. L'idée de produire de nouvelles idées de manière spontanée ou non est en lien avec l'idée de sauts conceptuels. Le type particulier de représentation mentale peut surement être mis en lien avec l'idée d'excéder son réseau de croyances nous donnant accès à une nouvelle manière de voir le monde ou une chose en particulier. C'est donc en soi une forme particulière de représentation mentale. Mais, surtout, cette définition met en avant l'aspect mystérieux de l'imagination. Bird et Hills parlent en effet d'un « type particulier » de représentation mentale sans préciser lequel et ajoutent que l'imagination prend des « formes très différentes » et dont la nature fait « l'objet d'importantes controverses ». L'imagination est donc une notion qui n'est pas claire et qui ne fait pas l'objet d'un consensus. C'est une notion qui nous échappe, nous, les humains qui l'utilisons pourtant. Nous supposons ici qu'il n'y a aucune raison que l'IA fasse preuve d'imagination car il ne semble pas qu'elle puisse reproduire une telle capacité de l'esprit puisque nous ne pouvons pas la décrire dans des termes

3. And are many and varied (fertility),
4. And to carry through these ideas to completion (motivation). ».

[1] Hills et Bird, page 696 : « It is an ability to produce a particular type of mental representation, but it is widely considered to have very different forms and its nature is the subject of significant dispute. As far as possible, we are liberal about the kinds of imagination that may play a role in creativity. For instance, the imagination may produce novel ideas deliberately (as when a scientist spends a great deal of time trying to develop a theory; or a writer edits and rewrites her work). The imagination may also generate ideas spontaneously (as when a scientist is suddenly struck by a hypothesis, or a writer by a novel image). ».

computationnels comme nous l'a expliqué Putnam. Pour avoir de l'imagination, il faut être un humain. Cependant, rien ne peut confirmer de telles suppositions. L'imagination a un statut particulier dans la définition de Bird et Hills. Contrairement aux autres points, la présence ou non d'imagination ne peut être vérifiée. Peut-être que l'IA fait preuve d'imagination, nous ne pouvons pas le savoir, pas plus que nous pouvons savoir si les autres humains font, eux aussi, preuve d'imagination. Néanmoins, sur ce dernier point la pensée de McDowell nous invite à dire que oui, les autres humains font preuve d'imagination puisque nous sommes les mêmes animaux ayant subi le même processus évolutif et s'étant développés dans des environnements, certes différents, mais similaires sur de nombreux aspects. Reste que l'on peut émettre des doutes. Et par ailleurs nous ne pouvons donc pas savoir ce qu'il en est de l'IA.

Un problème se pose alors. Nous savons désormais que c'est la présence d'imagination qui amène à affirmer la présence de créativité, certes, mais nous ne nous pouvons pas infirmer ou confirmer la présence effective d'imagination chez d'autres individus que nous-mêmes. La seule chose que nous pouvons affirmer c'est qu'il semble moins évident d'affirmer que l'IA est créative (et donc imaginative) sur la seule base des résultats que fournissent les IA génératives. Désormais, nous savons que produire des choses comme des images, des animations ou des musiques qui sont nouvelles ne fait pas de leur créateur quelqu'un ou quelque chose de créatif. C'est d'ailleurs ce que confirment Bird et Hills. Pour eux :

> […] toutes les façons de produire des idées nouvelles et valables ne sont pas forcément créatives. Par exemple, il peut être possible de produire de telles idées par un processus aléatoire ou par un processus purement mécanique, tel que l'application d'une règle simple. Mais il ne s'agit pas typiquement d'exercices de créativité.[1]

Or, il s'avère que quand nous pensons processus de création des IA génératives, nous pensons justement à quelque chose de quasi mécanique, à l'application de règles, de schémas et de relations acquises par apprentissage pour générer quelque chose de nouveau. Nous pouvons donc douter du fait que les IA génératives soient créatives ce qui, par conséquent, peut nous faire également douter du fait qu'elles aient de l'imagination. Mais les doutes ne suffisent pas pour affirmer que l'IA n'a pas de créativité. Il faut donc que nous nous focalisions sur l'autre point de la définition de Bird et

[1] Hills et Bird, page 695 : « […] not all ways of producing novel, valuable ideas are creative. For instance, it may be possible to produce such ideas by a random process; or by a purely mechanical pro- cess, such as following a simple rule. But these are not typically exercises of creativity ».

Hills qui ne semble pas être respecté par l'IA, à savoir la motivation. En effet, ce quatrième point est, contrairement à l'imagination, plus visible, plus facilement constatable.

Ainsi, pour Bird et Hills :

> Nous considérons la motivation comme un élément important de la créativité, car elle traduit l'idée que l'individu créatif est celui qui est disposé à donner vie à de nouveaux dessins, constructions, musiques, théories et autres, bref qui a « envie de créer ». La créativité n'est pas simplement une capacité : quelqu'un qui pourrait créer des choses, mais qui ne fait pas l'effort de le faire, n'est pas un individu créatif.[1]

L'IA générative a les capacités de créer des choses et elle s'exécute sur demande. Cependant, il est plus compliqué d'accorder à l'IA l'« envie de créer » ou que celle-ci produise un effort lors de la création. L'envie implique l'idée de désir. L'effort implique une notion de fatigue physique ou mentale ainsi qu'une notion de volonté. Il ne semble pas que l'on puisse accorder ces traits à l'IA. Elle n'a pas vraiment « envie » de générer des images, elle ne fait que répondre à une demande en fonction des règles dont elle dispose. Elle ne fait pas d'effort car l'ordinateur ne se fatigue pas comme un corps humain ou un cerveau. On pourrait alors nous reprocher de jouer sur les mots et que l'on pourrait tout à fait considérer l'IA comme ayant la volonté d'accomplir sa tâche et produisant un effort. Cependant, il semble que pour Bird et Hills, l'idée de vouloir s'impliquer avec son cerveau et son corps dans notre tâche, l'idée de s'investir pour réaliser quelque chose, semble importante. Il y a une différence entre exécuter des règles sur demande et réaliser quelque chose par la volonté propre du créateur. Cette nuance nous oblige à exclure l'IA de l'idée de volonté et d'effort. Par exemple, elle ne sera pas frustrée ou déçue en cas d'échec. Elle ne sera pas non zélée. Si elle a une apparence de zèle ce sera parce que l'on ne lui aura pas appris à s'arrêter tant qu'elle n'aura pas le résultat escompté, ce ne sera pas parce qu'elle veut vraiment réussir. Les modèles massifs de langages et les IA génératives semblent très dociles. Ils exécutent ce qu'on leur demande, mais ne font rien par leur propre initiative. Ils ne désirent rien de particulier alors nous ne pouvons pas décemment dire qu'ils sont motivés pour accomplir la tâche qu'on leur assigne. Ils s'exécutent car ils ont été conçus ainsi. Cette distinction entre exécuter des règles sur demande et réaliser quelque chose par la volonté propre du créateur semble très importante pour Bird et Hills

[1] Hills et Bird, page 700 : « Motivation we regard as important to the trait of creativity as it captures the idea that the creative individual is one who is disposed to bring new drawings, constructions, music, theories, and such like into existence, in short has an "urge to create". Creativity is not merely an ability: someone who could create things, but does not make the effort to do anything of the sort is not a creative individual. ».

afin de caractériser ce qu'est la créativité et c'est pour cela que nous pouvons considérer que l'IA ne remplit pas cette quatrième condition pour être créative.

Ainsi, se servir des progrès de l'IA générative pour postuler sa créativité et (par conséquent) son imagination est une méthode fallacieuse. L'IA produit, elle exécute les règles et schémas qu'elle a appris afin de répondre au mieux aux demandes des utilisateurs. Nous pouvons douter du fait que l'IA puisse avoir de l'imagination, capacité propre aux humains, mais nous ne pouvons rien prouver. Cependant, il semble bien que l'IA n'ait pas de volonté propre, de désir apparent dans les processus qu'elle applique. Ainsi, si l'on suit la définition de Hills et Bird, alors nous pouvons conclure que l'IA ne fait pas preuve de créativité. Peut-être bien que l'on parviendra un jour à créer des IA capables de faire preuve de motivation dans la réalisation de leurs tâches, mais l'impossibilité de prouver quoi que ce soit concernant l'imagination risque de maintenir un doute quant à leur caractère créatif.

5. La spontanéité et la seconde nature

Enfin, terminons sur deux arguments, liés entre eux, et qui reprennent directement le modèle de l'esprit de McDowell. En effet, comme nous l'avons vu, McDowell reprend la pensée de Kant mais propose que les concepts se trouvent aussi dans la sensibilité. Ils sont mobilisés dans cet espace de l'esprit afin que le Donné soit déjà, au moins en partie, conceptualisé. Ce faisant, l'entendement, via la spontanéité, peut saisir cet objet issu de l'extérieur et y faire sien via l'application de concepts. Pour que l'entendement puisse saisir cet objet, il faut déjà qu'il y ait du concept dedans. Cela permet alors à l'entendement et la raison d'avoir une relation avec l'extérieur tout en garantissant la pleine liberté de la spontanéité. Une contrainte causale de l'extérieur existe mais elle ne s'applique pas directement sur l'espace logique des raisons ni sur la spontanéité. Mais qu'est-ce que la spontanéité exactement ? Pour Kant :

> Notre connaissance procède de deux sources fondamentales de l'esprit, dont la première est le pouvoir de recevoir les représentations (la réceptivité des impressions), la seconde le pouvoir de connaître par l'intermédiaire de ces représentations un objet (spontanéité des concepts).[1]

[1] Immanuel Kant, *Critique de la raison pure*, trad. par Alain Renaut, 2e éd. corr, GF 1142 (Paris: Flammarion, 2001), page 143.

La spontanéité c'est donc un pouvoir de connaître mais aussi « le pouvoir de produire soi-même des représentations »[1]. Cette capacité, c'est l'exercice spontané de notre entendement permettant de former nos représentations au moyen des concepts et de connaître ces objets. D'après Kant, l'exercice spontané de l'entendement est entièrement libre. Il peut produire les représentations comme bon lui semble, sans que l'extérieur ou autre chose puisse intervenir. Il jouit d'une liberté et conserver celle-ci est un des enjeux dans la problématique que s'était posé McDowell.

La limite que nous posons à l'IA est donc la suivante. Celle-ci peut manipuler des symboles, des données sous formes binaires ou décimales, mais elle ne se forme pas de représentations au moyen de concepts. Reprenons notre IA idéale. Celle-ci dispose de divers capteurs afin d'avoir, comme les humains, des informations en continu sur l'environnement. Cependant, l'IA ne sent pas vraiment l'environnement au sens où nous le sentons. Les informations que reçoit l'IA ne sont que des données brutes, des suites de chiffres et de nombres transmis par les capteurs. On pourra alors nous rétorquer que nous aussi nous recevons des données brutes sous forme d'impulsions chimiques ou électriques transmises par nos organes sensibles. Cela est vrai, mais là différence est que nous conceptualisons ce Donné afin d'en faire un objet sensible que l'on peut connaître. L'IA, elle, semble dépourvue d'une telle capacité. Supposons qu'un arbre se trouve en face du robot habité par l'IA. Les capteurs visuels de cette IA vont alors pouvoir fournir un flux d'informations, des signaux électriques qui seront traités comme des signaux binaires par le système. Une fois traité, ce flux permet à l'IA de créer une matrice. Chaque case de la matrice correspond à un pixel, chaque pixel contient des valeurs numériques correspondant à une couleur spécifique. L'IA dispose alors d'une image. Elle « connaît » la couleur de chaque pixel au sens où elle sait associer le terme d'une couleur à une valeur spécifique et, si l'on affichait cette matrice sur un écran, nous aurions alors une image de l'arbre. Cependant, l'IA ne voit pas vraiment un arbre au sens où, nous, nous voyons un arbre. De ce fait, elle ne connaît pas non plus l'arbre comme, nous, nous le connaissons. En effet, comme nous l'avons vu, l'IA n'a pas la conscience phénoménale permettant d'avoir le ressenti qualitatif des couleurs. Elle n'a accès qu'à des nombres. Ces nombres sont une forme de représentation, certes, et l'IA a donc un accès à son environnement. Cependant, l'IA n'a pas les mêmes représentations que nous, les humains. Quand l'IA parvient à déterminer que ce qu'elle voit c'est un arbre, c'est parce qu'elle a préalablement été entraînée à associer un tel agencement de

[1] Kant, page 144.

pixel, un tel schème matriciel, au mot « arbre ». Ainsi, elle peut reconnaître, prédire, mais elle ne connaît pas vraiment l'arbre de la même manière que nous. Sa capacité de reconnaissance s'appuie sur une base de données dont elle dispose, sur l'application sans réflexion de règles apprises permettant ainsi d'interpréter les données numériques qu'on lui fournit. Nous, les humains, nous portons notre jugement non sur un ensemble de données, mais sur une représentation que l'on s'est formée au moyen des concepts. De ce fait la nature de cette représentation diffère de celle de l'IA, ne serait-ce parce que nous avons une expérience phénoménale associée à la représentation, associé à l'objet que nous avons conceptualisé. Nous portons un jugement sur une représentation que notre entendement, via l'exercice de sa spontanéité, a formé.

L'IA ne semble pas pouvoir reproduire ce genre de représentations qui sont les nôtres. Chez Kant, nos représentations sont internes et ne correspondent pas forcément au monde tel qu'il est vraiment. Le monde que nous connaissons est le monde tel que notre esprit le forme à travers les concepts et les catégories de l'entendement. Comment incorporer de tel processus cognitifs dans l'IA ? Comment incorporer ce processus libre et automatique qu'est la spontanéité ? Comment l'IA peut-elle former une représentation empirique des données qu'elle reçoit ?

Nous avons déjà conclu que l'IA symbolique considérée toute seule est hors-jeu, mais l'IA connexionniste, aussi prometteuse soit-elle, a aussi des lacunes. Peut-être bien que, via le *deep learning*, nous pourrions reproduire et imiter des capacités cognitives, mais il restera de profondes lacunes. Si l'on suit la pensée de Kant, comme le fait McDowell, la spontanéité de l'entendement a cela de spécial qu'elle possède une forme de liberté, ce que n'ont pas les réseaux de neurones. Comme nous l'avions vu dans le chapitre précédent, les réseaux *deep learning*, dotés de capacités et de ressources finies, sont par exemple incapables de reproduire l'infinité des raisonnements possibles. Ils ne peuvent pas apprendre l'entièreté des raisonnements possibles et n'aura donc qu'une quantité limitée de raisonnements à disposition. Ce qui lui manque, c'est donc une capacité, une forme de jugement libre qui ne se base pas que sur les jugements qu'elle connaît déjà mais qui en forme aussi de nouveaux en fonction de la situation. Ce qui manque c'est une forme de spontanéité en somme. Nous pouvons bien sûr remettre en question cette notion de liberté et défendre un déterminisme universel, mais cela ne changera pas pour autant la difficulté. En effet, il faudra tout de même que l'IA soit capable de connaître l'entièreté des raisonnements possibles passés et présents mais aussi prédire l'entièreté des raisonnements futurs afin de pouvoir user pleinement de la spontanéité de l'entendement et de la raison. Il faudrait, en effet, qu'elle connaisse

toutes les causes de l'apparition de tel ou tel raisonnement et les effets qu'ont ceux-ci. Cela reviendrait sûrement à devoir réduire en des termes physiques et computationnels l'entièreté de l'univers visible passé, présent et futur, chose impossible aux vues des capacités forcément limités des IA. Une approche déterministe semble alors plus délétère pour l'IA que l'approche kantienne.

Quelle solution aurions-nous alors ? Il faudrait parvenir à mettre au jour une description fidèle des processus cognitifs mis en jeu. Il faudrait aussi franchir le gouffre explicatif de la conscience phénoménale. En somme, si l'on suit le modèle de l'esprit de McDowell, il faudrait parvenir à réduire la seconde nature des humains dans l'espace logique des lois de la nature ou dans des termes computationnels voire psychologiques afin d'avoir une description des capacités de l'esprit sous forme mathématique, sous forme de lois par exemple comportementales.

En effet, si nous trouvions un algorithme permettant de reproduire parfaitement ces processus cognitifs, si nous mettions à jour les lois physiques et chimiques du cerveau responsables de la conscience phénoménale, nous serions alors peut-être capables de créer un réseau de neurones qui, aidé d'algorithmes (donc d'IA symbolique), parviendrait à reproduire la formation de représentations sensibles. Elle pourrait alors connaître et prendre conscience du monde qui l'entoure ainsi que d'elle-même. En reproduisant la conscience phénoménale, les couleurs ne seraient peut-être plus seulement des ensembles de valeurs numériques mais de vraies expériences phénoménales. Elle pourrait peut-être ressentir de la douleur et, pourquoi pas, se sentir en vie. Enfin, elle pourrait aussi porter des jugements de manière spontanée et libre.

Mais peut-être que le jeu de l'imitation a ses limites. Peut-être que mettre au point un algorithme qui copie parfaitement des processus cognitifs, ne permet pas pour autant à l'IA exécutant cet algorithme d'avoir ces processus cognitifs. Elle ne ferait que les imiter. Un algorithme ou un réseau de neurones qui reproduirait fidèlement le comportement du cerveau n'aurait peut-être pas de conscience phénoménale. En imitant quelqu'un qui voit des couleurs, qui en a le ressenti phénoménal, n'implique pas forcément que j'ai, moi aussi, ce ressenti. Faire comme si nous avions une seconde nature n'aura peut-être pas la conséquence d'en avoir réellement une.

Comment acquérir cette seconde nature alors ? Si l'on suit la pensée de McDowell, Il y a plusieurs critères importants à respecter. En effet, pour lui :

> On peut concevoir les exercices des capacités relevant de la spontanéité comme des éléments dans le cours d'une vie. Un sujet d'expérience, agissant, est une chose vivante, avec des

> pouvoirs corporels passifs et actifs qui lui appartiennent authentiquement ; il est lui-même incarné, présent substantiellement dans le monde dont il fait l'expérience et sur lequel il agit.[1]

L'idée d'un individu vivant, présent dans le monde, possédant un corps et ayant une relation avec un environnement est essentielle pour l'apparition de capacités comme la spontanéité de l'entendement. Il faut donc faire en sorte que l'IA soit aussi dans le même cas. Mais ce n'est pas tout. Un autre élément important demeure dans la pensée de McDowell, il s'agit de l'évolution. Nous avons pu développer nos capacités cognitives en partie grâce à l'évolution. Ce fut une étape indispensable pour l'apparition de l'esprit humain. Ainsi, il faudrait parvenir à concevoir des organismes capables d'évoluer progressivement de génération en génération avec un système de gènes permettant aux individus de mélanger leurs génomes en concevant d'autres individus. Ce faisant, nous pourrions alors peut-être avoir l'apparition de processus cognitifs similaires aux nôtres. Au bout d'un nombre extrêmement conséquent de générations, nous pourrions alors peut-être voir l'apparition d'IA qui, au moyen de pratiques sociales apparues durant son processus évolutif, seraient douées de conscience du monde et d'elles-mêmes, d'une réelle conscience phénoménale et qui seraient capables de former des représentations suivant des catégories et des concepts bien à elles. Cependant, rien ne nous empêcherait de court-circuiter le processus évolutif. Nous savons comment nos processus cognitifs sont apparus, mais nous ne sommes pas obligés de répliquer la même histoire qui prendrait des centaines de milliers d'années. Nos compétences en biologie et en génétique nous permettraient sûrement de grandement accélérer le processus. Ce que l'on sait avec la pensée de McDowell c'est que c'est de cette manière-ci que nous avons nos capacités cognitives, mais rien ne nous empêche de chercher une autre voie. Il semble juste qu'il faille privilégier le recours à la biologie et à la théorie évolutive dans tous les cas, étant donné que c'est la seule manière d'atteindre la seconde nature que nous connaissons. Il faut donc passer par la biologie cellulaire.

Or, des problèmes se posent à nous et c'est là notre second argument. D'abord, pour qu'il y ait évolution (même si elle est court-circuitée) il faut qu'il y ait reproduction. À cela, on pourrait nous rétorquer qu'il suffit de mettre en place un logiciel capable de créer des réseaux de neurones qui engendreraient des descendants dans le même ordinateur. Cependant, nous avons vu que, pour McDowell, il y a d'autres conditions à remplir à savoir être vivant, être présent dans le monde et interagir avec un environnement. En effet car, pour que le processus évolutif se fasse comme il se

[1] McDowell, *L'esprit et le monde*, page 148.

doit, il faut des interactions avec un environnement physique. Une autre raison de la nécessité d'une présence dans son environnement c'est, rappelons-le, que l'environnement à un impact causal direct sur les êtres qui vivent en son sein. Cela permet alors de corriger les jugements, de les contraindre afin qu'ils s'accordent mieux avec la réalité. Il faut donc des robots, présent physiquement dans le monde, pouvant recevoir toutes les informations nécessaires sur son environnement et pouvant agir sur celui-ci. Si nous ne pouvons pas nous contenter de créer une simulation, de créer une IA capable de se reproduire dans un ordinateur, s'il faut que l'IA soit présente physiquement et interagisse avec un environnement, il faut donc trouver un moyen de la faire se reproduire.

Une condition importante qui s'impose aussi est que l'IA doit pouvoir mourir, sans quoi il n'y aurait pas d'enjeux pour la survie et la reproduction. Il faut que les générations précédentes laissent place aux nouvelles et que la mort soit là pour créer l'enjeu de la lutte pour la survie et de la perpétuation des gènes. La vie nécessitant des ressources pour subsister dans un environnement possiblement hostile est une condition nécessaire pour le processus évolutif. L'objectif serait alors de parvenir à créer des robots capables de se reproduire, de vieillir et de mourir. Or, cela ne semble pas possible avec les robots actuels faits de plastique et de métal, à moins de concevoir des morts artificielle et arbitraires, ce qui enlèverait l'infinité des possibilités et des nuances de dangers de mort présentes dans le monde. Pour la reproduction, il n'est pas possible de produire du nouveau plastique ou métal par méiose ou mitose étant donné qu'il n'y a pas de cellules. Il faudrait alors constituer des robots faits de matière organique, créer des êtres de chair.

Cependant, cela poserait le problème de savoir comment implémenter l'IA, qui fonctionne pour l'instant dans des ordinateurs faits de métaux, dans un tel être. Nous aborderions alors la problématique de la copie d'un esprit humain dans un ordinateur et inversement. Ce scénario de science-fiction sous-entend que l'esprit fonctionne soit de manière computationnelle et équivalente à un algorithme, chose que nous avons rejetée, soit qu'il fonctionne comme un réseau de neurones de type *deep learning*, ce dont nous doutons fortement étant donné les arguments que nous avons avancés jusqu'ici. Copier un esprit humain dans un ordinateur ou une IA dans un cerveau semble très fantaisiste. L'idée de copier une IA dans un cerveau artificiel organique le paraît tout autant. Il faudrait plutôt supposer l'existence d'ordinateurs entièrement organiques qui pourraient, eux aussi se reproduire, vieillir, mourir et dans lesquels nous pourrions directement programmer ou développer notre IA, ce qui est, nous en conviendrons, très fantaisiste également.

Force est de constater qu'il faut avoir recours à énormément de conditionnels et de scénarios paraissant actuellement irréalistes pour supposer la possibilité qu'une IA puisse développer une seconde nature par cette voie-là. De plus, en se posant toutes ces questions, en décrivant toutes les conditions nécessaires pour créer une intelligence artificielle en se basant sur le modèle de l'esprit de McDowell, nous ne faisons finalement que décrire le processus d'apparition de la vie. Si nous créons des êtres cellulaires qui luttent pour leur survie et évoluent, alors nous ne cherchons pas à créer une intelligence artificielle mais nous tentons plutôt de faire apparaître l'intelligence naturelle. Le programme de recherche sur l'IA se transforme alors en programme de recherche sur l'apparition du vivant. Il n'y a même pas besoin de créer d'algorithmes ou de réseau de neurones au préalable qui l'on implanterait dans ces êtres organiques, à moins de vouloir sauter des étapes et aller plus vite. Nous perdons donc l'essence même de l'Intelligence Artificielle si l'on passe par cette voie, même s'il faut reconnaître que ce serait surement la méthode la plus crédible afin d'obtenir une nouvelle seconde nature et une intelligence.

6. La nature, condition nécessaire à l'apparition de l'esprit humain

Ainsi, comme nous avons pu le voir, il y a différentes théories concernant la conscience, l'introspection ou l'imagination permettant de favoriser et de sauver l'approche matérialiste de l'esprit. Si l'on accorde du crédit à ces théories, alors la reproduction d'un esprit humain semble possible. En sauvant les théories matérialistes de l'esprit, nous aurions alors une explication, une description de ses capacités. Cependant, ces théories présentent des lacunes, ce qui nous fait douter quant à leur viabilité. Il semble donc que, concernant l'esprit, il y a des paramètres, des capacités, que l'on n'arrive pas réellement à décrire, à expliquer, à réduire en des termes scientifiques et computationnels. Or, McDowell, lui, estime que c'est normal que l'on n'ait pas d'explications, ne serait-ce que parce que l'espace logique des raisons ne peut être réduit en des termes de lois de la nature. Mais, ce n'est pas non plus pour autant que l'esprit n'est pas naturel et que ce serait une sorte de substance immatérielle magique. Pour lui, l'esprit tire ses fondations de notre corps, de notre cerveau, et s'est développé lui aussi via un processus évolutif et via les interactions sociales qui ont permis l'apparition du langage. L'esprit est naturel, c'est une seconde nature, mais il échappe au moins en partie à la réduction scientifique, psychologique et computationnelle. Dès lors, si l'on ne peut pas décrire ces processus, comment les reproduire artificiellement ? En effet,

comment reproduire une conscience phénoménale ? Comment avoir de l'introspection ? Comment avoir de l'imagination ? Comment avoir de la spontanéité de l'entendement ? Le seul moyen plausible serait de recréer la vie et de prendre inspiration sur l'évolution. Ce faisant, nous abandonnons alors l'idée même de l'Intelligence Artificielle et d'un idéal de machines à conscience humaine. L'esprit, selon McDowell, requiert de vivre dans un environnement, de pouvoir agir sur celui-ci. Il requiert également la vie et par conséquent la mort. Il faut également d'autres individus, de la communication et de la vie en communauté. Enfin, l'esprit de McDowell requiert du temps, beaucoup de temps et la succession d'innombrables générations d'individus pour arriver au résultat que nous connaissons aujourd'hui. McDowell est d'accord avec Aristote quand ce dernier dit que l'humain est un « animal rationnel ». L'esprit sans la nature, sans un corps biologique naturel et sans les autres ne semble pas pouvoir exister.

Conclusion

Le naturalisme modéré de McDowell est une conception de l'esprit qui s'inscrit dans le cadre du réalisme naturel. C'est une théorie pragmatique qui permet de sortir des débats entre idéalisme et matérialisme, entre monisme et dualisme de l'esprit. Cette théorie de l'esprit affirme l'aspect naturel des capacités de l'esprit comme l'imagination ou la spontanéité de l'entendement tout en soutenant qu'il est normal que ces capacités ne disposent pas de réduction dans l'espace des lois de la nature. Elles sont naturelles sans pour autant être réductibles dans des termes physiques et computationnels. Ce sont des capacités qui se sont formées via le long processus de l'évolution et, surtout, via l'apparition du langage. L'apparition de capacités comme le langage couplé à la vie en société a permis l'émergence d'un nouvel espace cognitif, l'espace des raisons. Via la vie en société et le processus éducatif, permis par le langage, que McDowell nomme « Bildung », les humains ont pu acquérir une seconde nature, des schèmes de raisonnements qui leur permettent de voir et de comprendre le monde d'une manière inédite et différente des autres animaux. Cette seconde nature nécessite une vie, une vie comprenant des rapports sociaux et des rapports avec la nature. Cette conception de McDowell s'inscrit dans le cadre du réalisme naturel et requiert alors un rapport direct au monde et le refus de tout scepticisme à l'égard de la réalité de celui-ci. Nous sommes ce que nous sommes, nous avons conscience de nous-mêmes et nous nous sentons vivre parce que nous vivons dans un environnement, un monde bien réel.

La théorie de l'esprit de McDowell, parce que c'est une théorie du réalisme naturel, s'oppose au computationnalisme, théorie de l'esprit stipulant que le cerveau humain fonctionne comme un ordinateur, exécutant des algorithmes, suivant une table de Turing. Pour McDowell, les capacités de l'esprit ne peuvent pas être réduites en des termes physiques et computationnels.

En montrant la fausseté du fonctionnalisme grâce aux travaux de Putnam sur le réalisme interne et le réalisme naturel, il paraît alors impossible de pouvoir affirmer que l'IA symbolique, consistant en l'exécution d'algorithmes et sous-entendant que le cerveau humain fonctionne de la même manière, puisse reproduire un esprit humain. Si on réfute le computationnalisme, l'IA symbolique se retrouve condamnée. Elle ne pourra jamais réaliser l'objectif prométhéen du programme de recherche en Intelligence Artificielle. Seule l'IA connexionniste, fonctionnant aujourd'hui avec des réseaux de neurones multicouches complexes peut prétendre pouvoir

atteindre un jour cet objectif prométhéen. Mais le doute subsiste. En effet, de nombreuses capacités comme la conscience phénoménale, l'introspection, l'imagination ou l'exercice spontané de l'entendement semblent impossibles à reproduire. Il existe des théories permettant de franchir ces difficultés voire de nier leurs existences, mais ces théories restent critiquables et l'on peut rester sceptiques quant à leur viabilité. Qui plus est, la difficulté la plus importante est que, selon McDowell, il est nécessaire d'être vivant au sein d'un environnement naturel, de vivre en société afin de développer nos capacités pour développer une seconde nature. Le seul moyen que nous connaissons actuellement pour faire apparaître la seconde nature, c'est par l'évolution d'organismes multicellulaires, c'est par les transactions directes avec l'environnement et les autres individus, c'est par l'apparition du vivant et son développement. Il existe peut-être d'autres méthodes pour faire apparaître la seconde nature, mais nous ne les connaissons pas.

Il faut donc prétendre pouvoir recréer un être biologique et vivant comme nous pour reproduire la seconde nature. Il apparaît nécessaire de se focaliser davantage sur la création d'être vivants multicellulaires que sur la création et l'entraînement de réseaux de neurones. Au moyen des connaissances que nous avons, recréer une intelligence est plus l'affaire du programme de recherche sur l'apparition de la vie que le programme de recherche sur l'Intelligence Artificielle.

Accepter le modèle de l'esprit de McDowell, c'est accepter d'abandonner l'ambition de créer un ordinateur ou un robot doué de conscience pour se focaliser sur le vivant. C'est accepter l'existence de capacités ne disposant pas de descriptions physiques et computationnelles et donc accepter que celles-ci ne puissent pas être reproduites par des réseaux de neurones, dont le fonctionnement consiste en l'ajustement quasi mécanique de la pondération des neurones. C'est accepter que des choses soient naturelles sans que l'on puisse les expliquer en des termes de lois de la nature (et donc les reproduire). Enfin, c'est affirmer le caractère nécessaire, pour l'apparition de l'esprit, de la vie et de la transaction directe des individus avec le monde.

McDowell fournit donc une conception de l'esprit résolument pragmatique, résolvant les problèmes que rencontrent l'approche matérialiste et l'approche idéaliste de l'esprit, mais qui pose une réelle limite au programme de l'Intelligence Artificielle. Si l'on veut recréer un esprit, alors le programme de recherche tel qu'il est actuellement semble être voué à l'échec. Les IA réussissent d'ores et déjà le jeu de l'imitation imaginé par Turing et elles finiront sans nul doute par devenir de plus en plus « humaines » aux yeux du grand public. Mais nous pourrons toujours légitimement douter du fait que ces IA seront dotées d'un réel esprit humain.

Bibliographie

Andler, Daniel. *Intelligence artificielle, intelligence humaine : la double énigme*. NRF Essais. Gallimard. Paris, 2023.

Bernier, Paul. « Compte rendu de [Hilary Putnam, Représentation et réalité, traduction française par Claudine Engel-Tiercelin, Paris, Éditions Gallimard, 1990, 226 pages.] ». *Philosophiques* 18, nº 2 (1991): 191-95.

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

Carruthers, Peter. « How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition ». *Behavioral and Brain Sciences* 32, nº 2 (2009): 121-38.

Dreyfus, Hubert L. *Intelligence artificielle : Mythes et limites*. Traduit par Rose-Marie Vassallo-Villaneau. Flammarion, 1984.

Duhem, Pierre. « La théorie physique et l'expérience ». In *La théorie physique. Son objet, sa structure*. Bibliothèque idéale des sciences sociales. Lyon: ENS Éditions, 2016.

———. *La théorie physique. Son objet, sa structure*. Bibliothèque idéale des sciences sociales. Lyon: ENS Éditions, 2016.

Esfeld, Michaël. *La philosophie de l'esprit : Une introduction aux débats contemporains*. Cursus. Armand Colin, 2020.

Evans, Gareth, John McDowell, Gareth Evans, et John McDowell. *The Varieties of Reference*. Oxford, New York: Oxford University Press, 1982.

Fodor, Jerry A., et Ernest Lepore. *Holism: A Shopper's Guide*. Édité par Ernest LePore. Cambridge, Mass., USA: Blackwell, 1992.

GdRIA du CNRS, ouvrage coordonné par Sébastien Konieczny et Henri Prade. *L'intelligence artificielle. De quoi s'agit-il vraiment ?* Cepadues, 2020.

Godon, Pierre. « L'intelligence artificielle va-t-elle tuer les artistes ? » *Franceinfo*, 21 janvier 2023. https://www.francetvinfo.fr/culture/bd/enquete-franceinfo-l-intelligence-artificielle-va-t-elle-tuer-les-artistes_5610134.html.

Hills, Alison, et Alexander Bird. « Against Creativity ». *Philosophy and Phenomenological Research* 99, nº 3 (2019): 694-713.

Jean, Aurélie. « Une brève introduction à l'intelligence artificielle ». *médecine/sciences* 36, nº 11 (1 novembre 2020): 1059-67.

Kammerer, François. « La conception illusionniste de la conscience phénoménale. Défis et perspectives ». *Klēsis* 55 (2023). https://www.revue-klesis.org/pdf/klesis-55-03-françois-kammerer-conception-illusionniste-conscience-phenomenale-defis-perspectives.pdf.

Kant, Immanuel. *Critique de la raison pure*. Traduit par Alain Renaut. 2e éd. corr. GF 1142. Paris: Flammarion, 2001.

Ludwig, Pascal, et Matthias Michel. « Introspection (A) ». In *L'encyclopédie philosophique*, 2017. https://encyclo-philo.fr/item/130.

McCulloch, Warren S., et Walter Pitts. « A Logical Calculus of the Ideas Immanent in Nervous Activity ». *The Bulletin of Mathematical Biophysics* 5, nº 4 (1 décembre 1943): 115-33.

McDowell, John Henry. *L'esprit et le monde*. Traduit par Christophe Al-Saleh. Analyse et philosophie. Paris: J. Vrin, 2007.

Newell, Allen, et Herbert A. Simon. « Computer science as empirical inquiry: symbols and search ». *Communications of the ACM* 19, nº 3 (1 mars 1976): 113-26.
Nisbett, Richard E., et Timothy D. Wilson. « Telling More Than We Can Know: Verbal Reports on Mental Processes ». *Psychological Review* 84, nº 3 (1977): 231-59.
Putnam, Hilary. « La Nature Des États Mentaux ». *Les Études philosophiques*, nº 3 (1992): 323-35.
———. « La signification de « signification » ». In *Textes Clés de philosophie de l'esprit Vol. II : Problèmes et perspectives*, traduit par Dominique Boucher. Textes clés. Vrin, 2003.
———. *La triple corde*. Édité par Pierre Fasula. Traduit par Pierre Fasula, Raphaël Ehrsam, Jeanne-Marie Roux, et Sabine Plaud. Analyse et philosophie. Paris: Librairie philosophique J. Vrin, 2017.
———. *Raison, vérité et histoire*. Traduit par Abel Gerschenfeld. Propositions. Paris: Édition de Minuit, 1984.
———. *Représentation et réalité*. Traduit par Claudine Tiercelin. NRF essais. Paris: Gallimard, 1990.
Quine, Willard Van Orman. *Le mot et la chose*. Traduit par Joseph Dopp et Paul Gochet. Champs. Essais. Paris: Flammarion, 2010.
———. « Les deux dogmes de l'empirisme ». In *De Vienne à Cambridge : L'héritage du positivisme logique de 1950 à nos jours*, édité par Pierre Jacob, 93-121. Tel. Gallimard, 1980.
Rochefort, Pierre-Yves. « Putnam (A) ». In *L'encyclopédie philosophique*, 2017. https://encyclo-philo.fr/putnam-a.
Schwitzgebel, Eric. « Self-Ignorance ». In *Consciousness and the Self*, édité par JeeLoo Liu et John Perry, 2012.
Searle, John R. « Minds, brains, and programs ». *Behavioral and Brain Sciences* 3, nº 3 (septembre 1980): 417-24.
Sellars, Wilfrid. *Empirisme et philosophie de l'esprit*. Traduit par Fabien Cayla. Tiré à part. Paris: L'Éclat, 1992.
Strawson, Galen. « The Consciousness Deniers ». *The New York Review of Books* (blog), 13 mars 2018.
Turing, Alan, et Jean-Yves Girard. *La machine de Turing*. Traduit par Julien Basch et Patrice Blanchard. Points Sciences. Points, 1999.
Turing, Alan Mathison. « Computing machinery and Intelligence ». *Mind* LIX, nº 236 (1 octobre 1950): 433-60.