**UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES**

# LE RÔLE DU LANGAGE DANS L'ATTRIBUTION DE CAPACITÉS COGNITIVES ET D'ÉTATS MENTAUX AUX MACHINES « INTELLIGENTES »

**MÉMOIRE PRÉSENTÉ**
**COMME EXIGENCE PARTIELLE DE LA**

**MAÎTRISE EN PHILOSOPHIE**

**PAR**
**JÉRÉMIE GARCEAU**

**JANVIER 2024**

# REMERCIEMENTS

Je souhaite d'abord remercier mon directeur de recherche, le professeur Jimmy Plourde. Son accueil chaleureux dans le département de philosophie, son ouverture, sa disponibilité, ainsi que la justesse de ses conseils et de ses enseignements ont grandement contribué à alimenter ma curiosité intellectuelle et mon attention aux nuances qui font de la vie un objet d'étude aussi merveilleux. Merci de m'avoir aidé à acquérir la confiance nécessaire à la réalisation de ce projet et de m'avoir guidé, du premier au deuxième cycle, dans ce parcours dont je suis extrêmement fier.

Je remercie bien évidemment le département de philosophie de l'UQTR et la Fondation de l'UQTR pour le soutien financier qui m'a été apporté tout au long de mes études.

Je souhaite aussi remercier les membres de ma famille, en particulier, mes parents qui m'ont encouragé, depuis l'enfance, à poser toutes les questions du monde. Merci aussi pour la confiance que vous m'avez accordée dans la poursuite de mes projets. Votre support et votre confiance ont contribué à mon développement et à la réussite de ce travail.

Je remercie aussi toutes les personnes qui ont croisé mon chemin durant ces années d'études. De nombreuses conversations ont longuement occupé mes réflexions et font maintenant de ce travail ce qu'il est.

Finalement, je réserve ces dernières lignes pour mon amoureuse, ma douce Rosie. Tu m'as aidé et tu m'as accompagné durant les neuf dernières années. Nous avons eu la chance de grandir ensemble et je ne saurais pointer des moments précis où tu as contribué à la réalisation de ce projet, tellement chacune de ses pages est influencée par ta présence.
Merci pour toutes les belles journées. Je t'aime !

# Table des matières

« *A stupid man's report of what a clever man says is never accurate,*
*because he unconsciously translates what he hears*
*into something that he can understand.* »

- Bertrand Russell. *A history of western philosophy*.

# INTRODUCTION

Le développement rapide de l'intelligence artificielle (IA) amène aujourd'hui un nombre important de chercheuses et de chercheurs à s'intéresser aux interactions entre humains et machines. Les machines dotées d'IA peuvent effectivement simuler de plus en plus fidèlement l'intelligence humaine, et ce, notamment en communiquant grâce aux langues naturelles. Plusieurs ont d'ailleurs observé un phénomène qui renverse nos intuitions : l'attribution, par des êtres humains raisonnables, d'états mentaux et de capacités mentales/cognitives à des machines dotées d'IA. Sachant que ces machines sont dépourvues de toute vie mentale, les études qui rapportent ce phénomène frappent l'imaginaire.

En effet, l'utilisation de l'IA permet de créer des machines phénoménologiquement complexes qui nous sont présentées comme pouvant prendre des décisions, remplir certains rôles et tâches, pouvant nous fournir des informations, etc. Elles sont, ainsi, intégrées dans nos vies sociales avec des caractéristiques utiles pour interagir avec nous. La capacité à traiter les langues naturelles en est une immensément importante. En effet, plusieurs machines et systèmes qui font maintenant partie intégrante de notre quotidien peuvent traiter des énoncés formulés verbalement ou textuellement et fournir des sorties sous la forme de mots, de phrases et d'expressions. On peut donc maintenant créer des machines qui semblent « dire » certaines choses pour impacter le monde autour d'elles et qui semblent comprendre ce que nous leur disons à notre tour. De là, nous proposons de nous intéresser au rôle du langage dans l'attribution de capacités mentales/cognitives et d'états mentaux aux machines dotées d'IA. La capacité à comprendre est, en effet, une capacité

que nous attribuons fréquemment aux implémentations de l'IA. En posant des questions à l'assistante vocale Siri ou au robot conversationnel *ChatGPT*, par exemple, nous considérons que ces assistants peuvent « interpréter » et « répondre » adéquatement à nos demandes. Cependant, il y a des limites à cette compréhension. Une machine peut, par exemple, être perçue comme incapable de comprendre certaines choses, incapable de comprendre des demandes, des informations implicites, etc. De là, nos interactions avec ces machines se trouvent modifiées de façon importante si nous leur attribuons ou non la capacité à comprendre le langage (si nous considérons, en interagissant avec elles, qu'elles peuvent comprendre). Nous nous concentrerons sur l'étude de la notion de compréhension, car c'est une capacité mentale/cognitive qui paraît plus « fondamentale » que les états mentaux particuliers que nous attribuons aux machines (comme avoir des croyances et des désirs, par exemple) et parce que son attribution aux machines dotées d'IA est très fréquente, vu la nature de ces machines qui sont conçues comme des assistants qui peuvent répondre à nos besoins et réaliser des tâches pour nous. Ce sont donc les questions qui motivent la rédaction de ce mémoire : en quel sens est-ce qu'une machine peut « comprendre » ? Qu'est-ce qu'elle doit pouvoir faire pour que l'on considère qu'elle « comprend » le langage ? Quel est le rôle du traitement artificiel du langage dans l'attribution, par des humains, de capacités mentales et d'états mentaux à des machines dotées d'intelligence artificielle ?

Ces questions sont immensément importantes, de nos jours, car nous nous trouvons aujourd'hui à un moment clé dans le développement de l'intelligence artificielle. Les robots conversationnels et les assistants vocaux comme *ChatGPT*, Siri ou Alexa sont de plus en plus puissants et simulent mieux que jamais l'intelligence humaine. Le langage est l'outil au cœur du développement de ces machines et c'est à travers lui qu'elles peuvent réaliser les tâches nommées plus haut. En ce sens, les réponses que nous donnerons aux questions de savoir comment se définit la compréhension artificielle du langage, comment elle est comparable à la compréhension

naturelle (des êtres humains) et comment les machines réussissent à manifester des capacités mentales équivalentes aux nôtres (comme la compréhension linguistique) peuvent jouer des rôles décisifs dans le développement des systèmes d'IA. En effet, il faut considérer qu'une machine comprend (au moins en un sens minimal) les informations qu'elle fournit pour avoir confiance en la véracité de celles-ci, comme il faut considérer qu'elles peuvent comprendre les états psychologiques que nous leur communiquons pour considérer qu'elles peuvent être empathiques, etc. Suivant ces idées, il est fort plausible que la caractérisation de la capacité à comprendre que nous attribuons aux machines dotées d'IA ait le pouvoir d'influencer notre propension à attribuer à ces machines des états mentaux, mais aussi d'influencer les domaines de nos vies dans lesquels nous accepterons, ou non, d'intégrer ces nouvelles technologies.

C'est à la décennie de 1950 que l'on attribue la naissance de l'IA. C'est au même moment que la question de savoir si une machine peut « penser » et donc « comprendre » ce qu'elle fait prit une place d'envergure dans les travaux d'importants penseurs et penseuses en philosophie de l'esprit et du langage. Les travaux de nombreux auteurs et autrices issus de la tradition analytique sont alors immensément intéressants pour faire sens du phénomène d'attribution de capacités mentales et d'états mentaux aux machines dotées d'IA. Parmi les outils que nous ferons nôtres dans la cadre de ce mémoire, on compte notamment l'approche fonctionnaliste du mental. Ainsi, ultimement, ce mémoire s'appuiera sur l'élaboration d'une définition fonctionnelle de la compréhension qui représentera un point de départ pour rendre compte de nos tendances à attribuer cette capacité mentale aux machines. Cette définition ouvrira donc la voie à une nouvelle façon de réfléchir l'attribution d'états mentaux et de capacités mentales à l'IA, en partant d'une particularité saillante de ces machines : la capacité à traiter les langues naturelles.

D'un point de vue méthodologique, nous nous questionnerons sur la nature de la compréhension et sur les conditions nécessaires à l'attribution de la capacité à comprendre aux

machines intelligentes. En donnant une définition fonctionnelle de la compréhension, réalisable par des machines, nous entendons donner une définition différente de celle que l'on donnerait de la compréhension réalisable par des humains. Nous partons donc en présupposant que la compréhension dont peut faire preuve l'IA n'est pas forcément celle dont fait preuve l'humain, mais que l'IA peut réaliser un niveau de manipulation de la langue assez élevé pour qu'un humain lui attribue, en contexte d'interaction communicationnelle, la capacité à comprendre d'une façon équivalente à un humain. De là, nous analyserons des études qui portent sur les interactions humains-machines et les mettrons en relation avec des travaux importants en philosophie de l'esprit et du langage, pour définir les modalités nécessaires à l'attribution de cette capacité. Nous étudierons des éléments théoriques portant sur les trois dimensions du langage qui nous permettront de mieux comprendre ce qu'une machine doit faire pour que nous soyons portés à considérer qu'elle comprend le langage : la manipulation syntaxique, sémantique et pragmatique.

Le premier chapitre, intitulé *L'attribution d'états mentaux, la compréhension et l'IA* visera à situer notre problématique au sein de la littérature philosophique et scientifique. Ce faisant, nous présenterons les concepts philosophiques et théoriques utiles pour bien entamer ce travail de recherche. Nous présenterons donc le concept d'état mental en précisant deux types d'états mentaux attribuables aux machines dotées d'IA qui nous intéresseront particulièrement dans le cadre de ce travail : les représentations mentales et les attitudes propositionnelles. Ensuite, nous présenterons l'approche fonctionnaliste du mental et préciserons les raisons derrière notre adoption de cette approche. Nous présenterons ensuite, à partir d'exemples concrets tirés d'études empiriques, le phénomène qui motive la rédaction de ce mémoire : l'attribution d'états mentaux et de capacités mentales à l'IA. Cette présentation sera suivie d'une explication des raisons derrière notre choix d'étudier en particulier l'attribution de la capacité à comprendre le langage en montrant que le traitement des langues naturelles est une capacité non négligeable dans l'étude des

interactions entre les êtres humains et les machines intelligentes contemporaines. Ainsi, nous présenterons les tentatives d'explication du phénomène d'attribution d'états mentaux et de capacités mentales aux machines à partir du concept d'anthropomorphisme et montrerons qu'il y a, dans ces explications, un manque de considération pour le traitement du langage. Nous poursuivrons en précisant comment cette capacité rend les interactions avec ces machines uniques en montrant qu'elles sont des objets et des outils drastiquement différents de ceux avec lesquels nous sommes habitués d'interagir, parce qu'elles sont « interprétables ». Enfin, la dernière partie de ce chapitre servira à mettre en lumière la relation entre la compréhension et les états mentaux, en précisant en quel sens l'explication des attributions de la capacité à comprendre (à partir d'une définition fonctionnelle adéquate de la compréhension) peut nous fournir une piste pour expliquer l'attribution de certains états mentaux aux machines.

Dans le deuxième chapitre, intitulé *L'analyse syntaxique et la compréhension*, nous commencerons à présenter les éléments théoriques utiles à l'élaboration d'une définition fonctionnelle de la compréhension en nous concentrant sur le traitement de la syntaxe. Se déployant en trois temps, le chapitre défendra l'idée voulant que la capacité à traiter et à respecter les règles syntaxiques d'une langue par une machine dotée d'IA puisse manifester une capacité à comprendre des relations exprimées dans le langage. Pour en arriver là, nous présenterons d'abord différentes approches philosophiques du traitement de la syntaxe et de son apport à ce que l'on appelle la compréhension : l'approche de John R. Searle, les principes de compositionnalité et de contexte, le concept de formes logiques et le rapport entre le langage et le monde, puis les niveaux de structures syntaxiques, présentés notamment par Noam Chomsky. Nous clorons enfin cette première partie en présentant comment certains systèmes d'IA traitent les structures syntaxiques et montrerons en quel sens ces méthodes permettent le traitement artificiel de relations entre les mots. La deuxième partie du chapitre, quant à elle, illustrera ce qui aura été présenté précédemment à

l'aide d'exemples concrets de machines capables d'analyser les structures syntaxiques. De là, nous présenterons différents types de relations pour lesquelles une machine peut manifester de la compréhension : les relations spatiales et temporelles, agentielles, puis les intentions de communication. Enfin, la dernière partie de chapitre servira à montrer que la maîtrise syntaxique peut se montrer suffisante pour manifester un certain niveau de compréhension, mais qu'elle peut aussi s'avérer insuffisante selon le contexte conversationnel et que sa maîtrise complète n'est pas nécessaire pour que l'on attribue la capacité de comprendre à une machine. Nous présenterons alors les résultats d'un test où un système est confronté à des énoncés et des questions ambiguës pour lesquelles le traitement syntaxique à lui seul s'avère insuffisant pour manifester de la compréhension. Les notions d'ambiguïté référentielle et de connaissances d'arrière-plan nous permettront donc d'ouvrir le troisième chapitre qui fournira, quant à lui, des outils complémentaires à notre définition fonctionnelle de la compréhension. Le chapitre 2 se conclura par la formulation d'une première partie de définition.

Dans le troisième chapitre, intitulé *Les représentations sémantiques et la compréhension*, nous nous concentrerons sur le traitement artificiel des données sémantiques. Nous y soutiendrons l'idée voulant que certaines machines dotées d'IA puissent manifester de la compréhension à travers une capacité à montrer qu'elles sont en mesure de former et d'entretenir des représentations des significations des mots/expressions fonctionnellement équivalentes à nos représentations mentales. Pour en arriver là, nous explorerons d'abord les types d'expressions pour lesquelles les machines dotées d'IA peuvent manifester de la compréhension en les groupant sous deux grandes catégories : les représentations de la réalité et les expressions d'états internes. Nous préciserons ensuite ce que nous entendons par le concept de « représentation sémantique » et évaluerons, à partir des critiques philosophiques de celle-ci, l'idée voulant qu'une machine puisse entretenir des représentations de la sorte. Ensuite, nous montrerons en quoi peut consister une représentation

sémantique fonctionnellement équivalente à une représentation mentale de la signification d'une expression/d'un mot. De là, nous élargirons notre discussion sur la représentation des significations en analysant la notion de « concept ». Ultimement, cela servira à préciser, à partir d'une étude des approches les plus importantes en philosophie des concepts, comment une représentation sémantique peut être fonctionnellement équivalente à une représentation mentale comme un concept. De là, nous ferons ressortir différents contenus que les représentations sémantiques doivent avoir et différentes capacités qu'elles doivent permettre à l'agent qui les possède, pour que nous soyons portés à considérer que celui-ci comprend des significations équivalentes à nos concepts. Ultimement, cette étude nous permettra de formuler une seconde partie pour notre définition fonctionnelle de la compréhension qui précisera comment une machine peut manifester une capacité à comprendre des significations à partir de la nature des représentations et de ce qu'elles permettent normalement de faire.

Enfin, le quatrième chapitre de ce mémoire, intitulé *Le langage pragmatique et les intentions de communication*, proposera une analyse de la compréhension fondée dans une étude du traitement du langage pragmatique. Ainsi, c'est l'étude de la notion philosophique d'intention de communication qui guidera la rédaction de ce chapitre. Cette étude sera réalisée en trois moments, chacun d'eux présentant des types de situations où une machine peut manifester une capacité à comprendre des intentions de communications à travers le traitement du langage. Nous présenterons alors en premier lieu le traitement d'informations implicites, en deuxième lieu, le traitement et la formulation d'énoncés qui suivent des règles d'usages et, enfin, le traitement et la formulation d'énoncés exprimant des actes de langage. Ultimement, notre recherche nous permettra de formuler la dernière partie de notre définition fonctionnelle de la compréhension attribuable à l'IA.

Finalement, pour conclure ce mémoire, nous reviendrons sur la question du caractère fondamental de la compréhension par rapport aux états mentaux comme les attitudes propositionnelles. Nous utiliserons donc les dernières lignes de ce mémoire pour montrer comment notre définition fonctionnelle de la compréhension permet de rendre compte, non seulement de l'attribution de la capacité à comprendre le langage aux machines intelligentes, mais aussi de l'attribution d'autres états mentaux. Pour ce faire, nous reprendrons les résultats généraux des chapitres 2, 3 et 4 et montrerons en quoi les différentes façons par lesquelles une machine peut manifester de la compréhension démontrent aussi une capacité à comprendre et à avoir d'autres états mentaux. C'est d'ailleurs dans cette partie du mémoire que nous présenterons, sous une forme claire et concise, notre définition fonctionnelle de la compréhension applicable à l'IA, en formulant les conditions nécessaires à respecter par une machine/un système pour que nous soyons portés à lui attribuer la capacité à comprendre.

# CHAPITRE 1 — L'ATTRIBUTION D'ÉTATS MENTAUX, LA COMPRÉHENSION ET L'IA

## 1.1 Introduction

L'objectif de ce chapitre est de situer notre problématique et notre question de recherche au sein de la littérature qui traite de l'attribution d'états mentaux aux machines dotées d'intelligence artificielle (IA). Notre question de recherche est la suivante : quel est le rôle du traitement artificiel du langage dans l'attribution, par des humains, de capacités cognitives et d'états mentaux à des machines dotées d'intelligence artificielle ? Notre volonté d'étudier l'hypothèse selon laquelle l'utilisation d'un langage partagé entre les humains et l'IA joue un rôle dans l'attribution d'états mentaux et d'autres capacités mentales aux machines dotées d'IA vient, notamment, d'un intérêt actuel dans la recherche en robotique sociale et en psychologie pour ce phénomène. En effet, plusieurs chercheuses et chercheurs ont montré que les humains sont portés à attribuer des états mentaux et des capacités mentales/cognitives à ces nouvelles machines pour faire sens de leurs interactions avec elles, comme on le fait régulièrement en interagissant avec d'autres êtres humains (Horstmann et al., 2018; Lee et al., 2020; Appel et al., 2012; Spatola et al., 2021; Epley et al., 2007). Cependant, le rôle du langage dans ce phénomène reste aujourd'hui sous-étudié.

Pour introduire adéquatement notre problématique de recherche, nous débuterons par une revue de la littérature philosophique traitant de la nature des états mentaux. Nous expliquerons en quoi ils consistent et ce qu'ils impliquent pour les agents qui les ont. Nous nous concentrerons sur

les représentations mentales et les attitudes propositionnelles, parce que ces états mentaux nous permettront d'introduire des notions philosophiques fondamentales pour notre travail : l'intentionnalité, le contenu, l'expérience subjective, les propositions et l'utilité des concepts mentaux dans l'explication comportementale. Après avoir défini les états mentaux, nous introduirons l'approche que nous utiliserons pour répondre à notre question de recherche : l'approche fonctionnaliste. En contrastant cette approche aux approches physicalistes, nous montrerons que l'approche fonctionnaliste du mental est à favoriser pour identifier et définir adéquatement quels types d'états et de processus mentaux peuvent être attribués à des machines pourtant dépourvues de vie mentale. Ensuite, nous poursuivrons en présentant des tentatives d'explications des mécanismes sous-jacents de ce phénomène tirées de la littérature scientifique sur les interactions humains-machines et montrerons que ces tentatives ne tiennent pas suffisamment compte du rôle du langage dans l'attribution d'états mentaux et de capacités mentales. Ce faisant, dans la dernière partie de ce chapitre, nous introduirons l'hypothèse que nous étudierons tout au long de notre projet : celle voulant que l'utilisation d'un langage partagé entre l'humain et l'IA puisse nous amener à attribuer à ces machines une capacité mentale/cognitive : la capacité à « comprendre » le langage. Nous proposerons, d'ailleurs que celle-ci soit plus fondamentale que d'autres états mentaux, comme les états mentaux représentationnels et les attitudes propositionnelles. Ultimement, de ce travail de recherche et d'analyse philosophique résultera une définition fonctionnelle de la compréhension attribuable aux machines dotées d'intelligence artificielle[1].

[1] Le fonctionnalisme caractérise les états mentaux à partir de ce qu'ils font et non partir du réalisateur physique duquel ils surviennent. Cette approche est donc attractive pour expliquer la perception d'états mentaux chez des machines inconscientes. Elle est aussi particulièrement intéressante puisqu'elle permet de définir clairement la relation entre les états mentaux et le comportement, ce qui lui prodigue un pouvoir d'application important dans l'explication des interactions humains-machines.

## 1.2 La vie mentale : ce que sont les états mentaux

### 1.2.1 Les représentations mentales

Pour débuter, nous montrerons ce qui est entendu par « états mentaux », afin de mieux comprendre ce que l'on veut dire lorsque nous parlons d'attribution d'états mentaux à l'IA. Pour ce faire, il est utile de nous intéresser, comme point de départ, aux états mentaux représentationnels. Cette approche propose que certains états mentaux que nous avons consistent en des représentations mentales d'états de choses du monde. L'approche est, entre autres, utile pour montrer en quel sens les états mentaux sont intimement liés à la notion d'intentionnalité que l'on a utilisée, historiquement, pour qualifier la conscience en philosophie de l'esprit (la conscience est à propos d'autre chose : conscience *de quelque chose*). Dans le même sens, les états mentaux représentationnels sont « à propos d'autre chose »[2]. Ainsi, le concept d'intentionnalité est utile pour comprendre le caractère extrinsèque du contenu de certains de nos états mentaux (Kim, 2008 : 225-229).

Si le rôle du monde extérieur dans la formation du contenu des états mentaux fut relativement bien défendu et reconnu en philosophie de l'esprit, il est toutefois plus difficile d'accepter que l'entièreté du contenu de nos états mentaux provienne de l'extérieur. Jerry Fodor fut l'un des tenants de cette position et étudia la possibilité qu'il y ait des différences entre les types de contenus mentaux, en introduisant les notions de contenu étroit (*narrow content*) et de contenu large (*broad content*). Pour Fodor, c'est en développant une distinction entre les conditions de vérité d'un état mental et de son contenu qu'on parvient à comprendre la pertinence de faire une

[2] C'est d'ailleurs ce qui est entendu par le terme « représentation » : ces états *représentent* des choses du monde.

distinction entre ces deux types de contenu (Fodor, 1991). La réflexion de Fodor est la suivante : l'identité des contenus des états mentaux n'assure pas l'identité de leurs extensions (Fodor, 1987 : 45-46). Pour en arriver à cette conclusion, Fodor utilise l'expérience de pensée des terres jumelles élaborée par Hilary Putnam qui va comme suit :

> *[…] we shall suppose that somewhere in the galaxy there is a planet we shall call Twin Earth. Twin Earth is very much like Earth; in fact, people on Twin Earth even speak English. In fact, apart from the differences we shall specify in our science-fiction examples, the reader may suppose that Twin Earth is exactly like Earth. He may even suppose that he has a Doppelgänger - an identical copy - on Twin Earth […] One of the peculiarities of Twin Earth is that the liquid called "water" is not $H_20$ but a different liquid whose chemical formula is very long and complicated. I shall abbreviate this chemical formula simply as XYZ. I shall suppose that XYZ is indistinguishable from water at normal temperatures and pressures. In particular, it tastes like water and it quenches thirst like water. Also, I shall suppose that the oceans and lakes and seas of Twin Earth contain XYZ and not water, that it rains XYZ on Twin Earth and not water, etc. […] If a spaceship from Earth ever visits Twin Earth, then the supposition at first will be that "water" has the same meaning on Earth and on Twin Earth. This supposition will be corrected when it is discovered that "water" on Twin Earth is XYZ, and the Earthian spaceship will report somewhat as follows: "On Twin Earth the word 'water' means XYZ."* (Putnam, 1975 : 223)

Fodor, en reprenant cette expérience de pensée, propose qu'il existe une certaine relation d'identité entre le contenu des états mentaux de la personne qui habite la terre (Oscar) et ceux de la personne qui habite la terre jumelle (Oscar2) (Fodor, 1991). Cette relation est à découvrir dans le contenu étroit des croyances des deux personnages. Selon Fodor, les croyances de ces derniers ont le même contenu étroit, parce qu'une partie du contenu de leur état mental est indépendante de leurs contextes de vie respectifs (Fodor, 1987 : 50-54). Dans les deux cas, ils croient que la chose dont ils parlent (un liquide ayant certaines propriétés perceptuelles) est mouillée et transparente (Fodor, 1987 : 50-54). En ce sens, ce n'est qu'une fois implantées dans un certain contexte que leurs croyances en viennent à avoir certaines conditions de satisfaction/conditions de vérité et ce sont ces conditions qui forment le contenu large de leurs états mentaux. Autrement dit, si Oscar2 visitait

la Terre, il dirait la même chose pour parler de l'eau et utiliserait le terme *water* de la même façon que lorsqu'il en parle et l'utilise chez lui, et ce, parce qu'au moins une partie de sa croyance est identique à celle d'Oscar. Néanmoins, il n'y a pas de doute que si Oscar ou Oscar2 découvrait que le mot *water* ne réfère pas à la même chose sur les deux planètes et qu'on leur demandait ce qu'ils veulent dire lorsqu'ils parlent de *water*, ils poseraient probablement la question de savoir quelle signification de *water* nous cherchons à connaître : de quelle matière nous parlons (à laquelle des deux nous référons). Ainsi, selon Fodor, le contenu d'un état mental détermine son extension (la signification de *water* détermine sa référence), mais il le fait relativement à un contexte (Fodor, 1987 : 47, 48). Suivant ces idées, ce qui diffère entre les croyances d'Oscar et d'Oscar2 à propos de l'eau, ce sont leurs contenus larges : leurs conditions de satisfaction/conditions de vérité qui fait que leurs croyances sont réellement à propos de l'eau avec laquelle ils sont en relation ($H_20$ ou XYZ) (Fodor, 1987 : 47, 48). Fodor, en développant cette position, arrive à concilier, d'une certaine façon, les intuitions internalistes et externalistes à propos du contenu des états mentaux.

Pour compléter cette introduction aux états mentaux représentationnels, nous devons distinguer les représentations mentales conceptuelles des non conceptuelles. Jerry Fodor a proposé, par exemple, qu'il existe un langage de la pensée (un ensemble de symboles et de règles syntaxiques) qui nous permet d'agencer certaines représentations (certains symboles qui représentent des objets externes) (Fodor, 1975). Ce genre d'approche implique que les représentations mentales soient formées de concepts, qui prennent la forme de symboles utilisés par la pensée et qui réfèrent à des objets en tant que catégories ayant des conditions de satisfaction (Fodor, 1998 : 24, 25). Ce type de représentations mentales fut historiquement différencié du contenu mental phénoménologique. Pour parler des représentations qui ont pour contenu des propriétés phénoménologiques, on peut, entre autres, utiliser les termes « représentations non

conceptuelles » ou *qualias*[3]. Pensons à l'expérience de la sensation de chaleur, par exemple : nous pouvons avoir une représentation de ce que cela fait de ressentir de la chaleur dans un des membres de notre corps et nous nous représentons certains objets comme ayant la propriété d'être chauds. En ce sens, les contenus de nos états mentaux peuvent *représenter* des effets qu'ont les objets sur nous (des représentations non-conceptuelles), comme ils peuvent *représenter* les objets eux-mêmes (des représentations conceptuelles).

Précisons maintenant les différences et les similarités entre les types de représentations mentales abordés. D'abord, ces représentations sont toutes à propos de choses extérieures à celui qui les a. Les représentations non-conceptuelles représentent les effets qu'ont les choses sur nous, ce qui nous poussera fort probablement à interagir avec les objets qui ont des effets positifs davantage qu'avec ceux qui ont des effets négatifs, alors que les représentations conceptuelles nous permettent d'organiser notre pensée à propos du monde en agissant comme des outils pour y réfléchir et pour interagir avec lui. Il est effectivement difficile d'imaginer une machine qui aurait des représentations sous la forme de *qualias* (non-conceptuelles), mais beaucoup plus aisé de voir que certaines machines sont capables, jusque dans une certaine mesure, de catégoriser des objets et donc de démontrer une capacité à entretenir et à communiquer des représentations conceptuelles des choses.

Il est important de garder cette distinction en tête, car elle pointe vers l'idée que les représentations mentales peuvent avoir différentes fonctions. Intéressons-nous justement aux représentations que l'on qualifie d'attitudes propositionnelles et à leurs fonctions. L'analyse des attitudes propositionnelles nous permettra de mieux comprendre comment des machines peuvent

---

[3] Cela dit, il faut noter que les phénomènes mentaux dont les contenus sont des *qualias* sont fréquemment considérés non pas comme des représentations, mais plutôt comme des sensations et sont regroupés sous ce que l'on appelle la « conscience phénoménale » (Kim, 2008a : 17).

« exprimer » des états mentaux et comment nous pouvons expliquer les comportements de certaines machines en faisant appel à ce type d'états mentaux. Les attitudes propositionnelles constituent le type d'états mentaux sur lesquels nous nous concentrerons dans le cadre de ce mémoire.

### 1.2.2 Les attitudes propositionnelles

Dans le cadre de l'approche représentationnelle de l'esprit (*representational theory of* mind), on nomme « attitudes propositionnelles » les états mentaux qui consistent en des attitudes qu'a une personne à l'égard de propositions (Richard, 2013 : 28; Pitt, 2022 : 1). Ces attitudes s'expriment par un verbe souvent suivi de la conjonction de subordination « que », puis d'une proposition P (Kim, 2008 : 17, 18). Par exemple, nous pouvons *croire que P*, *espérer que P*, *désirer que P*, etc. Nous pouvons aussi attribuer ces états mentaux aux autres, de la même manière que nous exprimons les nôtres : nous pouvons dire ou penser qu'une personne X *croit que* P. À ce moment-là, l'attitude propositionnelle que nous lui attribuons est : *croire que.* X pourrait aussi *désirer que* les choses soient d'une certaine manière : X, en regardant par la fenêtre de son appartement, pourrait *désirer qu'*il cesse de pleuvoir. L'état mental que nous lui attribuons est alors celui du désir et son contenu est la proposition « qu'il ne pleuve plus ».

Ces états mentaux sont effectivement à propos d'autre chose qu'eux-mêmes. Dans l'exemple précédent, l'état mental exprimé est à propos du temps qu'il fait à l'extérieur. La proposition désirée (la proposition qui est l'objet de l'attitude de X) exprime donc une représentation. L'analyse à faire est ainsi la suivante : X a un état mental que l'on peut caractériser par une attitude (*désirer*) à l'égard d'une proposition qui exprime une représentation du monde[4] (Pitt, 2022 : 1). Partant de

[4] Nous identifions le contenu des propositions étant objets d'attitudes au concept de « représentations » parce que nous prenons ce dernier concept dans un sens très large, tel qu'il est employé dans les approches représentationnalistes de

cette notion, nous aborderons deux sujets primordiaux pour la suite de notre projet. D'abord, nous parlerons des conditions de satisfaction des attitudes propositionnelles pour ensuite nous diriger vers les pouvoirs explicatifs et prédictifs de celles-ci.

#### 1.2.2.1 Les conditions de satisfaction des attitudes

Lorsque l'on parle de condition de satisfaction des attitudes propositionnelles, il est fondamental de s'intéresser à la notion de vérité. Nous savons qu'une proposition peut être vraie ou fausse[5]. Par conséquent, si une proposition qui est l'objet d'une attitude (comme une croyance) est vraie, alors le contenu de la croyance sera dit « vrai ». En ce sens, nous dirons ordinairement qu'une croyance est « vraie » si son contenu est vrai (Richard, 2013 : 7, 8). Par exemple, en disant « je crois que les chats roux sont plus aimables que les chats blancs », nous exprimons une croyance qui peut être vérifiée et dont on peut juger de la valeur de vérité (en cherchant à savoir s'il est effectivement le cas que les premiers sont plus aimables que les derniers). Il faut alors voir que les conditions de vérité du contenu d'une attitude propositionnelle viennent de la correspondance entre son contenu et les faits (Richard, 2013 : 7, 8). Toutefois, il faut comprendre qu'une croyance n'est pas réellement « fausse » même si son contenu (la proposition qui est l'objet de la croyance) est faux. Il va de soi qu'une personne peut véridiquement croire quelque chose de faux. En ce sens, la croyance elle-même (en tant qu'attitude) est dite vraie, même si ce qui est cru est faux. Elle correspond à l'état de croyance de la personne qui l'a. De plus, parler de la « vérité » de certaines attitudes propositionnelles n'est pas toujours adéquat. Prenons la seconde catégorie paradigmatique

---

l'esprit. Nous prenons les propositions comme exprimant des représentations parce que ce sont des objets ayant certaines propriétés sémantiques évaluables (comme une valeur de vérité et des conditions de vérité) qui peuvent être l'objet d'états mentaux (Fodor, 1993 : 273; Georgi, 2019).

[5] Pour un panorama des différentes conceptions philosophiques de la vérité, voir : Glanzberg, 2021.

d'attitudes propositionnelles : les désirs. Il semble étrange de dire que la condition de satisfaction d'un désir soit sa condition de vérité. Si on désire quelque chose, ce désir est toujours vrai. Pour cette raison, Jerry Fodor parle plutôt de *fullfilment* et de *frustration* (que nous pouvons traduire librement par accomplissement/satisfaction et frustration/refoulement) des attitudes comme les désirs (Fodor, 1987 : 10, 11). Suivant cette idée, avoir une croyance vraie ou accomplir/satisfaire un désir, c'est avoir une croyance rendue vraie ou un désir qui est considéré comme satisfait/accompli en vertu d'un état de la réalité.

Maintenant, si le contenu des états mentaux peut être vrai ou faux dû à un état de correspondance ou de non-correspondance à la réalité, il faut remarquer que nous pouvons attribuer des attitudes propositionnelles aux autres, qu'ils aient vraiment ou non ces attitudes (nous pouvons, par exemple, nous tromper en expliquant les comportements des autres en faisant appel à des croyances ou des désirs qu'ils n'ont pas réellement) ou que leurs contenus soient vrais ou faux[6]. Ce qui rend les attitudes propositionnelles si fascinantes, c'est effectivement la capacité qu'elles nous donnent d'expliquer les comportements des autres à partir de leur vie mentale. Nous attribuons des états mentaux aux autres pour décrire ce qu'ils font et les raisons derrière leurs actes en utilisant les attitudes propositionnelles. Nous opérons de la même manière aussi pour prédire ce qu'ils feront dans le futur : en leur attribuant des états mentaux comme des désirs, des croyances, des intentions, etc. (Taschek, 1995 : 275; Kim, 2008 : 16-18; Churchland, 2002 : 119-121) En regardant quelqu'un ouvrir successivement plusieurs portes d'armoires, par exemple, nous sommes

[6] Dans la même optique, plusieurs autrices et auteurs en philosophie de l'esprit ont pris position à propos de la nature de ces états mentaux soit en tant que réalistes, soit en tant qu'éliminativistes, les premiers soutenant l'existence réelle de ces attitudes dans le cerveau des gens (Fodor, 1978; Baker, 1995), les derniers soutenant leur inexistence ou leur caractère purement hypothétique et instrumental (Churchland, 2002; Dennett, 1987). Quoiqu'il en soit, la présupposition de l'existence de ce type d'états mentaux chez chacun de nous se trouve donc à la base de la psychologie populaire/du sens commun (*folk psychology*).

en mesure d'en déduire des énoncés comme « elle *voudrait que* l'une de ces armoires contienne un verre », « elle *croit qu*'il y a des verres dans l'armoire », etc.

#### 1.2.2.2 Les schémas d'attribution des attitudes

Pour comprendre comment s'exécutent ces capacités explicatives et prédictives que nous partageons, plusieurs philosophes de la tradition analytique se sont intéressés aux mécanismes d'attributions et d'expressions d'attitudes propositionnelles. Paul Churchland, par exemple, soutient que la capacité à utiliser la psychologie du sens commun vient de notre connaissance partagée de relations nomologiques entre des circonstances externes (on peut penser à des situations, par exemple), des états internes que sont censés avoir les agents et des comportements qu'ils ont (Churchland, 2002 : 119-120). Notre connaissance de ces relations nous permet alors d'élaborer ce que Churchland nomme des hypothèses explicatives fondées dans un appareillage de concepts qui sont censés référer à des états mentaux réels (Churchland, 2002 : 119-121).

Chez Daniel Dennett, on trouve une approche similaire : l'attribution d'états mentaux sous la forme d'attitudes propositionnelles permet l'interprétation de *patterns* comportementaux. Voici comment Dennett formule le processus d'attribution d'attitudes propositionnelles selon sa théorie de la posture intentionnelle (*intentional stance*):

> *[…] first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many—but not all—instances yield a decision about what the agent ought to do; that is what you predict the agent will do.* (Dennett, 1983 : 17)

Dennett, accorde une importance fondamentale aux concepts de croyance et de désir comme outil de prédiction efficace[7], parce qu'ils nous permettent de construire, dans notre propre esprit, l'agentivité rationnelle et intentionnelle d'autrui. De plus, il faut savoir que la théorie de la posture intentionnelle de Dennett, est aussi prévue pour rendre compte des attributions d'états mentaux à des entités non humaines comme des objets et des machines.

Dans la philosophie de l'esprit de Jerry A. Fodor, on trouve aussi un schéma d'attribution des attitudes propositionnelles dans lequel, cette fois, l'emphase est mise sur la notion de pouvoirs causaux. De là, l'essence de la position de Fodor sur les attitudes propositionnelles se résume effectivement en trois caractéristiques fondamentales : elles sont évaluables sur le plan sémantique, elles attribuent des contenus et des pouvoirs causaux aux « choses mentales » et elles sont des généralisations du sens commun qui réfèrent à des relations réelles entre des contenus d'états mentaux et des comportements volontaires/intentionnels (Fodor, 1987 : 10-16). Les deux premières caractéristiques sont celles qui nous intéressent particulièrement. Fodor nous dit d'abord que les attitudes propositionnelles sont évaluables sémantiquement ce qui signifie que lorsque nous attribuons des attitudes propositionnelles aux autres, nous pouvons juger de leur valeur de vérité et de leur accomplissement en nous intéressant aux relations que leurs contenus entretiennent avec le monde[8]. Nécessairement, dans le processus d'attribution d'attitudes, cette caractéristique est fondamentale, car elle permet d'expliquer pourquoi une personne agit de telle ou telle façon : on peut dire qu'une personne a laissé son partenaire pour accomplir son désir de vivre une relation

[7] L'importance de la force prédictive de la psychologie populaire est aussi au cœur du texte « De l'existence des patterns » (Dennett, 2002). Dennett tente d'expliquer cette force en montrant que si les processus d'attribution d'attitudes propositionnelles à des agents sont subjectifs, leurs résultats, quant à eux, sont objectifs. Plus précisément, grâce à la stratégie/posture intentionnelle, nous pouvons décrire différemment les mêmes *patterns* comportementaux qui, eux, sont bien réels. Attribuer certaines croyances et certains désirs à l'agent qui le réalise est une façon de le faire (Dennett, 2002 : 183-193).

[8] C'est l'idée que nous trouvons quelques lignes plus haut : les croyances peuvent être vraies ou fausses, et des désirs peuvent être comblés ou refoulés (Fodor, 1987 : 10, 11).

amoureuse plus satisfaisante, et qu'elle croyait, à raison ou à tort, qu'un autre partenaire pouvait lui fournir une telle satisfaction. De plus, d'un point de vue plus personnel, les gens rendent constamment compte de leurs comportements volontaires en énonçant des croyances et des désirs qu'ils entretiennent et justifient leurs actes par la rationalité des croyances et l'honnêteté ou la décence des désirs qui les poussent à agir (Fodor, 1993 : 272). Cela nous amène à la deuxième caractéristique proposée par Fodor. Il nous indique que les attitudes propositionnelles attribuent des pouvoirs causaux au mental. La causalité mentale est effectivement présupposée dans la psychologie du sens commun. Non seulement certains états mentaux comme les croyances et les désirs en causent d'autres, mais les événements qui ont lieu dans notre environnement causent chez nous des états mentaux et, enfin, certains états mentaux causent des comportements (Fodor, 1987 : 12). Ce qu'il y a là d'intéressant, c'est que l'explication comportementale de la psychologie du sens commun passe par l'attribution d'attitudes propositionnelles qui entretiennent entre elles des relations de contenus et des relations causales et que ces relations servent l'explication (Fodor, 1987 : 12-14). L'exemple suivant illustre cela : « *Hamlet believed that somebody had killed his father because he believed that Claudius had killed his father. His having the second belief explains his having the first.* » (Fodor, 1987 : 12)

L'approche de Churchland implique elle aussi une idée similaire, sans toutefois mettre l'emphase sur la causalité, mais plutôt sur les relations logiques. Il soutient que les différentes attitudes propositionnelles entretiennent entre elles des relations comme l'implication logique, l'équivalence logique, l'incompatibilité, etc. (Churchland, 2002 : 122, 123) Churchland en donne des exemples formels :

1. « $(x)(p)$ [($x$ craint que $p$) $\supset$ ($x$ désire que ~p)] » (Churchland, 2002 : 123).
2. « $(x)(p)$ [(($x$ espère que $p$) & ($x$ découvre que $p$)) $\supset$ ($x$ est heureux que $p$)] » (Churchland, 2002 : 123).

Il faut ici voir que ces relations sont intéressantes, parce qu'elles montrent comment nous pouvons attribuer différentes attitudes propositionnelles à un agent, à partir de l'attribution ou l'expression de l'une d'elles et donc comment nous pouvons en tirer des inférences.

Afin de conclure cette présentation des attitudes propositionnelles, revenons sur leurs principales caractéristiques. D'abord, ce sont des états mentaux que l'on peut exprimer ou attribuer aux autres en utilisant des verbes comme *croire, désirer, vouloir,* etc. suivie de la conjonction de subordination *que* et d'une proposition dont le contenu peut représenter le monde (tel qu'il est ou tel qu'il pourrait être). De là, les attitudes propositionnelles sont des états « intentionnels », au sens où ils sont à propos d'autre chose qu'eux-mêmes et leurs conditions de satisfaction dépendent des relations entre leurs contenus et les faits[9]. Nos usages quotidiens des attitudes propositionnelles servent à justifier, comprendre, expliquer et prédire des comportements. Diverses explications de ces capacités ont été données par les philosophes de la psychologie du sens commun : Dennett pense que les attitudes propositionnelles permettent d'interpréter des *patterns* comportementaux réels à travers une stratégie, Churchland ne croit pas que les attitudes propositionnelles comme les croyances et les désirs sont de réels états mentaux. Il croit plutôt que l'usage des concepts mentaux exprimés dans les attitudes propositionnelles permet de former des hypothèses explicatives fondées dans une connaissance de certaines lois qui relient comportement, vie mentale et circonstances. Fodor, quant à lui, croit que nos attributions et nos expressions d'attitudes propositionnelles permettent des explications et des prédictions efficaces des comportements, car elles expriment des relations (de contenu et causales) entre différents états mentaux, nous permettant alors de décrire

[9] L'état mental du désir est comblé s'il est effectivement le cas que le contenu de la proposition désirée est réalisé, alors qu'une croyance est vraie si son contenu s'accorde avec la réalité. Cela dit, nous avons bien noté que nous pouvons réellement croire des choses fausses : la vérité du contenu de la croyance et la vérité de la croyance elle-même sont alors à différencier.

les interactions entre des états internes et externes. Churchland, malgré son approche éliminativiste incompatible avec celle de Fodor, le rejoint sur le point des relations entre les attitudes[10].

Ces processus d'attribution d'états mentaux à autrui nous renseignent sur l'attribution d'états mentaux à des machines dotées d'IA sur deux points. En premier lieu, ils mettent en lumière les forces explicatives et prédictives, mais aussi les faiblesses des attributions d'états mentaux de type « attitudes propositionnelles » (on attribue parfois faussement, par exemple, des croyances à des gens que nous observons et nous attribuons aussi des états intentionnels à des machines inconscientes). En deuxième lieu, ces schémas soulignent l'existence de relations entre les attitudes propositionnelles. Ces différents états mentaux sont alors dans des relations d'implications logiques et donc les explications et prédictions comportementales qui sont basées sur ceux-ci utilisent ces relations pour expliquer et prédire adéquatement. On peut ainsi, inférer plusieurs états mentaux d'autrui à partir d'un état qu'il exprime ou que nous lui attribuons et il paraît plausible que nous le fassions aussi avec les machines qui peuvent se montrer capables d'en exprimer. C'est précisément l'attribution d'états mentaux ayant la forme d'attitudes propositionnelles aux machines dotées d'IA qui nous intéressera, car pour pouvoir avoir des états mentaux de ce type, il semble qu'une machine doive pouvoir faire preuve d'une certaine compréhension conceptuelle. Nous reviendrons sur ce point.

### 1.2.3 Étudier l'architecture du mental : le fonctionnalisme

Nous souhaitons maintenant introduire la lectrice et le lecteur à l'approche du mental que nous prendrons tout au long de ce mémoire : le fonctionnalisme. Pour comprendre la pertinence de

[10] Bien entendu, ces relations ne sont pas présentées comme réelles par Churchland, mais plutôt comme purement « logiques ».

celle-ci, il est d'abord utile de présenter brièvement les critiques adressées à la psychologie populaire dans l'histoire de la philosophie de l'esprit. Pour commencer, de nombreuses et nombreux philosophes ont remis en doute la pertinence des concepts mentaux que nous utilisons pour parler de l'interaction entre les actions et la vie mentale d'autrui et ont ainsi contribué à l'élaboration de plusieurs théories « physicalistes » ou « matérialistes » de l'esprit. Le physicalisme est une approche du mental voulant que tout phénomène mental dépende de phénomènes physiques et qu'il soit alors explicable à partir de ces phénomènes physiques. Herbert Feigl dans un article de 1958 synthétise les motivations des philosophes adoptant cette approche en proposant qu'ils en fassent leur cheval de bataille justement en bonne partie parce qu'elle permet d'évacuer les implications ontologiques de nos attributions quotidiennes d'états mentaux aux autres (Feigl, 1958 : 372). Pour ces auteurs (Feigl y compte, entre autres, Carnap, Quine, Ryle, Skinner et Sellars), le langage de la psychologie populaire peut s'avérer fautif et son contenu doit être, au mieux, relégué au statut d'ensemble de concepts intersubjectifs utiles pour parler des comportements (Feigl, 1958 : 372). Jaegwon Kim, quant à lui, donne une définition englobante du physicalisme à partir d'un exemple : « Si *x* possède la propriété mentale M (ou se trouve dans l'état mental M à l'instant *t*, alors *x* est une chose matérielle et *x* possède M à *t* en vertu du fait que *x* possède, à *t*, une certaine propriété physique P qui réalise M dans *x* à *t*. » (Kim, 2008a : 130) Le physicalisme ici décrit par Kim est dit « non réductionniste », notamment parce qu'il permet la « survenance » (*supervenience*) du mental sur le physique, par opposition au physicalisme réductionniste[11].

[11] L'exemple suivant permet d'illustrer la distinction : un physicaliste réductionniste soutiendrait que l'état mental croire que P est réductible à l'activation d'un ensemble de neurones N (Kim, 2008a : 100), tandis que le physicaliste non réductionniste soutiendrait que l'activation des mêmes neurones fait effectivement survenir des propriétés mentales, mais que ces propriétés supérieures ne sont pas réductibles aux propriétés physiques responsables de leur survenance (Kim, 2008a : 326-327, Kim, 2008b : 33, 34).

D'ailleurs, les réflexions qui ont émergées autour de ces différentes approches du mental ont menés à d'importantes discussions sur le problème de la réalisabilité multiple (Kim, 2008a : 130-135). Ce problème veut qu'un même événement mental puisse être réalisé par des réalisateurs physiques différents (Kim, 2008a : 130-135). Par exemple, un chien et un humain peuvent éprouver de la peur même si leurs neurones qui font survenir l'état mental « peur » sont différents. Ces discussions ont contribué à faire naître une approche novatrice des états mentaux qui reste encore aujourd'hui l'une des plus influentes pour adresser les questions entourant la nature des états mentaux et notre compréhension de ceux-ci : l'approche fonctionnaliste. L'approche fonctionnaliste nous sera utile, car elle permet d'expliquer l'équivalence entre des processus informatiques et des processus mentaux/cognitifs à partir de la notion de fonction. Pensons, par exemple, à un ordinateur et un humain qui peuvent tous deux réaliser des additions, et ce, même si le premier est un artéfact, alors que le deuxième est un organisme biologique (Kim, 2008a : 132, 133). La première chose à remarquer à propos des approches fonctionnalistes, lorsque nous les contrastons aux variétés de physicalismes, c'est qu'elles ne s'intéressent pas au mental à partir du même point de vue. Dans son livre, Kim parle d'un « niveau supérieur d'abstraction »[12] dans lequel les fonctionnalistes croient pouvoir trouver des propriétés communes aux différentes occurrences d'états mentaux (Kim, 2008a). Kim présente cette idée de la façon suivante : « Selon le fonctionnalisme, une espèce mentale est une espèce fonctionnelle, ou une espèce fonctionnelle causale, car la ‘‘fonction’’ en question consiste à remplir un rôle causal particulier. » (Kim, 2008a :

[12] L'idée d'un niveau supérieur d'abstraction d'analyse des états mentaux est aussi bien exemplifiée dans le chapitre *The nature of mental states* (1975), par Hilary Putnam. Il propose que les états mentaux soient fonctionnels au sens où ils sont caractérisables à partir de leurs fonctions. La présence d'un état fonctionnel (il faut noter Putnam définit un état fonctionnel comme l'état de réception d'entrées qui jouent un rôle dans l'organisation fonctionnelle d'un organisme) est la présence d'un état « total » : une sorte d'état global de l'organisme réalisé par plusieurs états partiels organisés d'une façon particulière à un temps donné. Le concept d'état total démontre un niveau supérieur d'abstraction (Putnam, 1975 : 433-435)

134, 135) La fonction d'un état mental est donc son rôle dans une chaine causale : une définition fonctionnelle est donc une description du travail que l'état mental accomplit pour un système ou un organisme[13] (Kim, 2008a : 131, 132, 135). On peut se demander, par exemple : « quel est le rôle de la douleur pour l'entité qui l'expérimente ? » (voir : Kim, 2008 : 135) De la même manière, on peut demander : « à quoi cela sert d'avoir une certaine croyance P ? »

Pour continuer, notre connaissance des états mentaux fonctionnels nous permet aussi d'expliquer les comportements. En effet, lorsqu'on sait ce que fait un état mental S, nous sommes en mesure d'inférer l'occurrence de cet état mental chez les autres, notamment lorsque nous reconnaissons des comportements propres à l'occurrence de l'état. Les entrées et les sorties reçues et générées par les états mentaux fonctionnels peuvent être multiples et de différentes natures. Les entrées dans la plupart des cas sont des entrées sensorielles ou d'autres états mentaux, tandis que les sorties peuvent être des comportements ou d'autres états mentaux (Kim, 2008a : 137). Maintenant, il faut noter que les états mentaux, selon l'approche fonctionnaliste, consistent en des états internes qu'on suppose comme réels qui relient entrées et sorties dans un système[14] (Kim, 2008a : 137-140). Donnons un exemple concret pour exemplifier ce qui vient d'être présenté. Puisqu'un état mental fonctionnel relie causalement certaines entrées à certaines sorties, une définition fonctionnelle de la douleur pourrait ressembler à cela : la douleur est un état interne S qui met causalement en relation les entrées X, Y et Z et les sorties A, B et C. On pourrait même être plus précis et dire que c'est un état mental qui lie causalement certaines entrées sensorielles

[13] La définition formelle de « propriété fonctionnelle (F) » est la suivante : « Pour une chose *x*, avoir F (ou être un F) = pour *x*, avoir une propriété P telle que C(P), où C(P) est une spécification de la tâche causale que P est censée remplir dans *x*. » (Kim, 2008a : 136)

[14] De là, les approches fonctionnalistes du mental ne contredisent pas nécessairement les approches physicalistes, parce qu'elles proposent simplement de ne pas seulement étudier le cerveau à partir d'un point de vue purement physique, mais d'aussi s'intéresser à l'architecture de la cognition et de l'esprit (grâce à son intérêt pour l'étude du mental à partir d'un niveau d'abstraction supérieur).

(un contact avec une surface coupante, brulante, etc.) aux états mentaux de la peur et de la tristesse, mais aussi à certains comportements de fuite, d'évitement, de défense, etc.

Puisque le fonctionnalisme permet de définir des états mentaux à partir de ce qu'ils font, plutôt qu'à partir du réalisateur physique duquel ils surviennent, cette approche est attractive pour expliquer notre capacité à faire sens des machines dotées d'intelligence artificielle en leur attribuant des états mentaux. L'approche fonctionnaliste permet effectivement de rendre compte de l'équivalence fonctionnelle entre des états d'une machine et les états mentaux biologiques. Elle est donc englobante et permet d'aborder l'hypothèse voulant que même sans expérience subjective, sans vie mentale comme celle de l'humain, des entités puissent entretenir des états équivalents à nos états mentaux. Il est alors plausible que nous percevions chez des machines qui imitent les facultés et les comportements humains certains états que nous considérons comme fonctionnellement équivalents aux nôtres. Enfin, l'approche fonctionnaliste est aussi particulièrement intéressante pour notre projet, parce qu'elle permet de définir clairement la relation entre les états mentaux et le comportement, et ce, en raison de l'attention qu'elle accorde à la description des relations causales entre les états internes et les actions, ce qui prodigue à cette approche un pouvoir d'application important dans l'explication des interactions humains-machines.

## 1.3 « Je sais ce qui se passe dans la machine » : l'attribution d'états mentaux

### 1.3.1 L'anthropomorphisme

Notre étude des états mentaux a montré que ces états, en plus d'être vécus, sont fréquemment utilisés pour faire sens des actions que nous observons chez les autres et même chez les machines. Cela va dans le même sens que les études qui tendent à montrer que les utilisateurs de machines forment fréquemment des modèles explicatifs personnels du fonctionnement de ces

dernières (Payne, 2009 : 40-43). On appelle ces modèles des modèles mentaux (*mental models*) que forment et qu'entretiennent les utilisateurs et les utilisatrices et ceux-ci influencent les comportements qu'ont les utilisateurs à l'égard des machines (Payne, 2009 : 40-43). De là, sachant que les modèles mentaux que nous entretenons à propos des machines dotées d'IA sont différents de ceux que nous formons pour des machines plus simples et qu'ils peuvent inclure l'attribution d'états mentaux aux machines, il est primordial de réfléchir aux modèles que nous avons tendance à former et à leurs conséquences. Une approche propose des explications intéressantes pour rendre compte des modèles mentaux que nous entretenons à l'égard de l'IA : l'anthropomorphisme.

L'anthropomorphisme est l'une des plus importantes tentatives d'explication de l'attribution de vie mentale aux entités non humaines. On le présente comme un phénomène ayant des origines à la fois cognitives et motivationnelles consistant à attribuer à des agents non humains des propriétés connues propres aux humains (Epley et al., 2007 : 865-866). Une façon claire d'imager le processus d'anthropomorphisme est le modèle de l'inférence que fournit Higgins (Epley et al., 2007 : 865 ; Higgins, 1996). Selon ce modèle, on anthropomorphise en effectuant des inférences à partir de connaissances préacquises emmagasinées, qui sont activées et appliquées à un agent cible (Epley et al., 2007 : 865 ; Higgins, 1996). C'est donc la grande accessibilité de nos connaissances[15] portant sur notre expérience phénoménologique du monde qui rendrait l'anthropomorphisme aussi fréquent, car elles nous permettent de les appliquer sur les animaux et les objets avec lesquels nous interagissons. Dans le même sens, on a aussi proposé que la formation

[15] Ceci fait écho aux théories de la simulation (Severson, Woodard, 2018; Harris, 2000; Harris, 2007). En effet, pour pouvoir appliquer de la connaissance à propos de la vie mentale à des machines, il semble qu'il faille d'abord être capable de simuler l'expérience de cette vie mentale. Autrement dit, pour anthropomorphiser une machine en lui attribuant des états mentaux, il faut avoir antérieurement fait l'expérience de ces états pour être en mesure de les simuler mentalement, afin de les attribuer à la machine (afin de considérer que la machine les expérimente aussi) (Harris, 2007 : 40-67).

de représentations alternatives à propos des agents non humains puisse contribuer à diminuer l'anthropomorphisation de ceux-ci (Epley et al., 2007 : 870).

Le débat sur les déterminants de l'anthropomorphisme devient stimulant pour nous, surtout quand on introduit l'idée qu'il existe des déterminants motivationnels derrière cette tendance. Epley et al. en étudient une : l'*effectance motivation.* En effet, l'anthropomorphisation viserait, selon eux, à tenter de diminuer les incertitudes et les ambiguïtés qui nous empêchent de progresser de façon efficace dans un contexte donné. L'anthropomorphisme représenterait donc une méthode intuitive et rapide pour réduire ces incertitudes et se sentir plus en contrôle et plus efficace dans son environnement, en plus de mieux pouvoir prédire ce qui se passera à l'avenir (Epley et al., 2007 : 871-75). Dans le contexte de notre travail, cela est intéressant parce que les machines dotées d'IA sont nouvelles et prennent diverses formes nous laissant face à de nombreuses incertitudes et ambiguïtés autant quant à leur fonctionnement qu'aux raisons de leurs actions.

Rappelons-nous maintenant la théorie de la posture intentionnelle de Daniel Dennett : il propose que l'on utilise des concepts de la psychologie populaire comme les croyances, les désirs et les intentions de façon instrumentale pour interpréter et prédire des comportements (Dennett, 1989 : 43-57). Cette théorie de la posture intentionnelle fut notamment reprise par des chercheurs dans le but de construire des tests servant à évaluer dans quelle mesure les gens sont effectivement portés à expliquer les comportements de robots à partir de concepts mentaux (Marchesi et al., 2019; Spatola et al., 2021). De plus, la psychologue Gabriela Airenti a elle aussi élaboré une théorie de l'anthropomorphisme assez sophistiquée, dans laquelle elle propose que l'anthropomorphisme consiste en un type particulier de relation que nous pouvons entretenir avec une entité non humaine, en nous adressant à cette entité comme s'il s'agissait d'un humain dans une situation communicationnelle (Airenti, 2018 : 8). Cette approche est particulièrement intéressante, car elle aborde le rôle de la communication dans le processus d'anthropomorphisation. Pour elle, c'est une

attitude naturelle que nous prenons pour établir une relation avec des entités non humaines, en faisant « comme si » elles étaient des interlocutrices à partir de deux modalités de base : celle de la compétition et celle de la coopération (Airenti, 2018 : 8). Ainsi, selon Airenti, dans une interaction où l'on observe de l'anthropomorphisme, les actions de l'entité (l'objet, la machine ou l'animal) apparaissent à l'humain comme adressées à lui (Airenti, 2018 : 9). Une dizaine d'années avant l'article d'Airenti, paraissait un papier d'Alexandra C. Horowitz (psychologue) et Marc Bekoff (biologiste) dans lequel on proposait quelque chose de similaire : l'autrice et l'auteur ont soulevé des corrélats comportementaux de l'anthropomorphisme dans les interactions entre humains et chiens. Plus précisément, quatre corrélats sous la forme de « catégories sociales » y sont présentés comme exposant des caractéristiques d'interactions sociales qui mènent à l'attribution d'états mentaux aux chiens par les humains, en contexte de jeu (Horowitz, Bekoff, 2007 : 26-28):

1. Le chien et l'humain utilisent des réponses dirigées pour montrer et vérifier leur participation mutuelle au « dialogue »;
2. le chien et l'humain indiquent leurs intentions à travers des comportements et des actes de communication (par exemple : montrer la balle au chien avant de la lancer);
3. les comportements entre les acteurs sont coordonnés de façon dynamique montrant un engagement mutuel dans l'activité qui rappelle l'engagement dans une conversation humaine;
4. il s'installe un rythme, une coordination entre les actions des deux acteurs, de sorte que le succès de l'activité est fondamentalement dépendant de la contribution de chacun et du fait que la contribution est apportée au bon moment.

En effet, Horowitz et Bekoff placent, comme Airenti, le concept de communication au cœur de leur théorie de l'anthropomorphisme. Mais il faut remarquer que dans les deux cas, ce concept est pris en un sens très large, notamment parce qu'il inclut la communication non verbale.

L'impact de l'utilisation d'un langage verbal partagé sur l'anthropomorphisme est en effet très peu étudié, notamment parce que l'anthropomorphisme est, par définition, l'acte de caractériser des comportements non humains ou des objets inanimés en utilisant des termes humains (Horowitz, Bekoff, 2007 : 23). Ainsi, puisque nous partageons le langage, normalement seulement avec des humains, l'idée semble paradoxale[16]. Pourtant le développement actuel de machines comme des robots conversationnels et des assistants vocaux ravive la pertinence d'aborder le phénomène de l'attribution d'états mentaux à des entités non humaines à partir de l'influence d'un langage verbal partagé. C'est maintenant d'autant plus pertinent, sachant que la plupart des machines intelligentes avec lesquelles nous interagissons ne présentent pas de caractéristiques physiques qui suggèrent l'anthropomorphisme. En ce sens, il paraît clair qu'une bonne analyse du phénomène de l'attribution de capacités cognitives et d'états mentaux à l'IA doit passer par une analyse plus complète des interactions entre humains et machines et ces interactions incluent des communications verbales.

### 1.3.2 L'IA comme phénomène interprétable

La question à laquelle nous accorderons toute notre attention dans ce mémoire est celle qui cherche à savoir comment nous pouvons être portés, en tant qu'humains rationnels, à attribuer la capacité à comprendre le langage à des systèmes qui fonctionnent grâce à l'intelligence artificielle. Cette réflexion s'impose à nous comme des plus importantes, principalement parce que plusieurs chercheurs et chercheuses ont pu observer le phénomène plus général d'attribution d'états mentaux et de processus cognitifs à certaines machines par des humains sans toutefois avoir pu l'expliquer

[16] Les modèles explicatifs que nous venons de présenter semblent effectivement surtout adéquats pour l'attribution d'états mentaux aux animaux, mais semblent quelque peu minimalistes pour traiter de l'IA, parce qu'ils font généralement abstraction de la capacité à utiliser un langage partagé ou n'y attribuent pas une importance signifiante.

d'une façon satisfaisante. Une raison fondamentale motive notre volonté de nous concentrer sur le processus mental de la compréhension : l'IA fournit, comme sorties, des phénomènes physiques et ceux-ci prennent souvent la forme de mots, de phrases et d'expressions. L'implémentation de l'IA permet donc de créer des machines qui semblent « dire » certaines choses et qui impactent ainsi le monde autour d'elles (Cappelen, Dever, 2021 : 10-20). Ces machines représentent, pour nous, des créatures complexes et intrigantes. Elles prennent diverses formes (des ordinateurs qui affichent des suites de lettres ou de chiffres sur des écrans, des assistants vocaux qui ressemblent à de petites boîtes, des robots humanoïdes, etc.), peuvent faire différentes choses, nous pouvons interagir avec elles, certaines remplissent même parfois certains rôles sociaux, etc. Dans l'espace public, ces machines sont présentées comme étant porteuses d'informations, comme pouvant répondre à des questions, comme pouvant prendre des décisions, comme pouvant nous aider, comme pouvant accomplir des tâches complexes, etc. (Cappelen, Dever, 2021 : 43, 44; Bicho et al., 2011; Liang et al., 2019; Yang et al. 2020; Li et al., 2021). Le langage étant un outil de partage d'information, il représente une caractéristique fondamentale de ces machines intelligentes qui intègrent nos vies.

La description des implémentations de l'IA que nous venons de donner est grandement influencée par l'approche novatrice que les philosophes Herman Cappelen et Josh Dever proposent dans leur ouvrage de 2021 intitulé *Making AI intelligible*. Ils y étudient la question de savoir en quel sens les sorties (*outputs*) que fournissent les systèmes d'intelligence artificielle signifient quelque chose. Les auteurs utilisent un exemple simple, mais parlant, pour montrer la pertinence de se questionner à propos du contenu des sorties de l'IA. Ils présentent le cas d'une machine servant à établir des cotes de crédit à partir de différentes données reliées à la vie financière d'individus (Cappelen, Dever, 2021 : 4-13). La machine effectue un calcul à partir d'un algorithme sophistiqué, puis fournit un chiffre qui est censé représenter la cote de crédit de l'utilisateur. À ce moment-là, Cappelen et Dever, mais surtout l'utilisateur qui la reçoit se demandent : « comment la

machine en arrive-t-elle à cette réponse ? », « qu'est-ce que le chiffre qu'elle affiche veut réellement dire ? », « qu'est-ce qui fait que ces chiffres signifient une cote de crédit ? ». Ils cherchent effectivement les raisons derrière la sortie. On voit déjà la raison qui pousse les deux philosophes à se questionner à propos de la signification des sorties générées par les systèmes intelligents, mais celle-ci se fait encore plus claire et apparente, selon nous, lorsque la sortie fournie par la machine prend la forme de mots, de phrases et d'expressions. On peut, en effet, être réticents à affirmer qu'une machine dit réellement des choses, prend des décisions, exprime des idées, interprète ce qu'un humain lui dit, le comprend et qu'elle fait, finalement, toutes les autres manipulations et actions que les humains réalisent avec le langage. Par exemple, la phrase « la ville de Montréal est magnifique en été » signifie quelque chose, mais est-ce qu'une machine qui la récite en réponse à une question à propos de la météo estivale veut vraiment dire la même chose que la pensée qui est exprimée dans celle-ci ? Est-ce que la machine *croit* vraiment que cette ville est magnifique durant la belle saison ? Comprend-elle vraiment ce qu'elle dit ?

En réfléchissant à la problématique qui intéresse Cappelen et Dever, nous avons placé les outils qui nous permettent de montrer en quoi les machines dotées d'IA se présentent à nous différemment des autres objets que nous connaissons et utilisons normalement. Ces machines interagissent en utilisant, en guise d'entrées et de sorties, des phénomènes physiques chargés de significations, ce qui montre la pertinence de se poser la question de savoir si cette particularité joue un rôle dans le phénomène plus large de l'attribution, à ces machines d'états mentaux et de capacités mentales.

### 1.4. La compréhension

#### 1.4.1 Le rôle de la compréhension dans les états mentaux

Revenons sur le chemin que nous avons parcouru jusqu'à présent. Dans les parties précédentes de ce chapitre, nous avons présenté les notions fondamentales et les concepts nécessaires à la compréhension de notre problématique de recherche. La première partie visait à présenter une revue des théories philosophiques utiles pour traiter des états mentaux qui peuvent être attribués par des humains à l'IA. Partant des états mentaux représentatifs, nous avons progressé pour en arriver à parler des états mentaux comme attitudes propositionnelles pour lesquelles nous pouvons donner des définitions fonctionnelles. Ce sont effectivement ces états qui sont exprimés et attribués aux machines qui existent actuellement : des machines inconscientes (qui ne ressentent pas d'émotions, dépourvues de toute vie mentale, mais qui peuvent potentiellement réaliser certains états ou processus fonctionnellement équivalents à nos états mentaux et processus cognitifs)[17]. Une fois notre étude de la nature des états mentaux faite, nous avons introduit plus en détail le phénomène intriguant qui motive la rédaction de ce travail : l'attribution d'états mentaux à l'IA. Nous avons ensuite porté notre attention sur la nature des machines qui nous intéressent dans ce mémoire en nous concentrant sur une faculté saillante : leur capacité à manipuler et traiter le langage naturel. De là, nous avons présenté des tentatives d'explications importantes et influentes du phénomène de l'attribution d'états mentaux à ces machines. Nous nous sommes concentrés, principalement, sur les théories importantes de l'anthropomorphisme et nous avons montré qu'il y a un manque de considération général pour le rôle du langage verbal dans ce phénomène. Nous avons, partant de cette lacune, présenté l'importance d'étudier son rôle en montrant que les

[17] Il faut noter que nous n'aborderons donc pas la possibilité de créer une IA forte (*strong AI*) dans ce mémoire. L'IA contemporaine n'a pas quelque chose comme un véritable esprit, comme en aurait une IA dite forte (Searle, 1980). Nous n'étudierons donc pas les machines dotées d'intelligence artificielle en partant de l'idée qu'elles peuvent réellement comprendre le langage comme nous le pouvons et en être « consciente », mais nous nous concentrerons plutôt sur leur capacité à « manifester de la compréhension » et à réaliser des processus qui lui sont fonctionnellement équivalents. Pour plus d'informations à propos de l'IA forte et de notre incapacité actuelle à créer de telles machines, voir : Searle, 1980; Dreyfus, 1992; Butz, 2021; Harnad, 1991.

machines dotées d'IA sont des machines interprétables. Pour compléter cette problématisation, il nous reste à présenter la relation qui existe entre la compréhension linguistique et les états mentaux. Ainsi, la lectrice et le lecteur pourront mieux comprendre comment l'attribution de la capacité à comprendre le langage à l'IA peut représenter une piste d'explication importante pour l'attribution d'états mentaux et de capacités cognitives à ces machines.

Il y a, en fait, différentes relations entre les états mentaux et le langage. Nous présenterons ici les principales qui nous poussent à nous intéresser au rôle du langage dans l'attribution de vie mentale aux machines. D'abord, quelques études sur les interactions humain-IA indiquent des corrélations entre l'attribution d'états mentaux à des machines et la capacité de ces machines à traiter le langage verbal, sans toutefois présenter et caractériser clairement les rapports entre les notions d'états mentaux, d'anthropomorphisme et de communication et comment elles contribuent au phénomène (Appel et al., 2012: 2; Hortsmann et al., 2018; Lee et al., 2020). Ensuite, il est clair que l'utilisation du langage par la machine permet de signaler la présence (même si elle n'en a pas réellement) d'états mentaux (Martin et al., 2020). Par exemple, une machine peut dire : « Je crois que X », l'expression linguistique de l'état mental de la croyance. De plus, même si on sait que les machines ne comprennent pas réellement ce qu'elles disent en utilisant le langage et qu'elles n'ont pas vraiment les états mentaux qu'elles expriment (des croyances telles que celles que nous avons, par exemple), nous utilisons, au quotidien, des assistants vocaux auxquels nous faisons des demandes verbales et nous nous attendons à recevoir des réponses adéquates. En demandant la météo prévue pour la journée, on s'attend à ce qu'on nous la communique. On s'attend à être compris. En ce sens, il faut remarquer que la compréhension semble être une capacité mentale fondamentale que nous attribuons assez facilement à l'IA. Autrement dit, même si plusieurs ne seraient pas prêts à dire que ces machines « comprennent » réellement comme le font les humains, nos actions à leur égard semblent pourtant présupposer ou témoigner que l'on considère qu'elles

comprennent, au moins en partie, le langage. Il y a donc là un manque de clarté à propos de ce que nous considérons être de la compréhension chez une machine dotée d'IA et pourtant, c'est une capacité que l'on a tendance à lui attribuer relativement souvent et facilement.

Pour continuer, il faut aussi voir que l'attribution d'états mentaux (comme les attitudes propositionnelles) aux machines, par les humains, paraît aussi présupposer que les machines peuvent au moins comprendre les objets de ces attitudes. Nous voulons ici porter l'attention de la lectrice et du lecteur sur le fait qu'un état mental de ce type représente une attitude à l'égard d'une proposition et que le fait d'avoir cet état présuppose déjà une certaine compréhension de la proposition qui y est incluse. Ainsi, lorsque nous attribuons à l'IA des états mentaux ayant la forme X (la machine) *croit que* P, X *veut que* P ou X *espère que* P, il semble que nous présupposions que X (la machine) comprend ce que P signifie. C'est en ce sens que l'attribution d'états mentaux aux machines semble présupposer l'attribution d'une certaine capacité à comprendre le langage. Dans la littérature philosophique, cette idée est fortement présente au sein de la théorie de la connaissance de Bertrand Russell. Lorsque Russell s'interroge sur la nature de la vérité, il en vient à la définir comme une relation de correspondance entre la signification de symboles complexes (des propositions au sein desquelles des symboles sont organisés et combinés de certaines façons) et des faits (Russell, 2002 : 187-200). Ainsi, nos croyances, par exemple, portent sur des propositions dont les significations réfèrent à des objets et à la forme d'un complexe. Nous devons alors comprendre toute proposition pour pouvoir adopter l'attitude de la croyance à son égard. Cette relation de dépendance des états mentaux au langage est aussi présente chez Anton Marty, étudiant du phénoménologue Franz Brentano, qui la stipule d'une belle et claire façon :

> Mais on reconnaîtra aussitôt le caractère adéquat de la méthode suivie en réalité par le langage, si l'on songe que tous les jugements, souhaits, etc. présupposent et incluent des représentations. Si nous avions un simple signe

> pour une question – celle que nous exprimons, par exemple, par les mots « y a-t-il une montagne d'or ? », nous aurions besoin, pour les représentations qu'elle contient, de signes distincts d'elles, si bien que le nombre des signes croîtrait jusqu'à l'incommensurable. Mais ainsi, lorsque le système des signes élémentaires, signifiant quelque chose par eux-mêmes, n'exprime tout d'abord que des représentations, nous avons besoin ne serait-ce que d'une série de formules auxiliaires (comme « il y a », « il se pourrait ») ou autres moyens qui permettent d'indiquer qu'un jugement, une question se réfèrent à ces contenus de représentation. Il est clair que ces signes adjuvants ne peuvent être que cosignifiants parce qu'il ne peut y avoir de jugement, etc., sans représentation, c'est-à-dire sans que le signe de cette représentation soit lui aussi présent. (Marty, 2017 : 118)

C'est la dépendance du jugement à la représentation qui intéresse ici Marty. Les jugements sont présentés comme étant à propos de représentations. En ce sens, pour juger de quelque chose, en disant qu'il se pourrait que cette chose soit le cas ou que cette chose est bel et bien le cas, il faut toujours comprendre la chose : cette fameuse chose est en effet le contenu représentationnel exprimé par les phrases que nous utilisons dans le langage. Ainsi, la position de Marty comme celle de Russell impliquent qu'au moins plusieurs états mentaux (précisément ceux que nous attribuons aux autres et que nous exprimons sous la forme d'attitudes propositionnelles) présupposent la compréhension de significations. Dans les deux cas, ce sont des contenus représentationnels que nous comprenons (pour Russell, on comprend, entre autres, des symboles qui réfèrent à des objets et la forme d'un complexe, alors que pour Marty on comprend directement des représentations).

Pour les raisons que nous venons d'évoquer, il paraît clair que le langage peut jouer un rôle important dans l'attribution de capacités mentales et d'états mentaux aux machines autonomes. Il est alors plausible que l'utilisation d'un langage partagé avec l'IA nous amène à lui attribuer la capacité à comprendre et cette capacité semble importante pour l'attribution à l'IA d'autres états mentaux (notamment ceux exprimables par des attitudes propositionnelles).

### 1.4.2 La compréhension linguistique comme processus mental fondamental

C’est indéniable, nous procédons à des inférences pour attribuer des états mentaux aux personnes qui nous entourent et les machines dotées d’IA ne font pas exception à cette pratique[18]. Pour conclure ce chapitre, nous souhaitons présenter brièvement une expérience de pensée qui vise à montrer concrètement la pertinence de se poser la question de savoir comment une machine peut manifester de la compréhension linguistique en présentant les conséquences de l’attribution d’une telle capacité cognitive.

L’expérience de pensée en question (inspirée d’une expérience scientifique, voir : Horstmann, 2018) met en scène une machine qui peut communiquer en français et une personne humaine. On demande à l’humain d’éteindre la machine, tandis que cette dernière le supplie sans cesse de ne pas l’éteindre, en évoquant des volontés et des peurs qu’elle prétend avoir. En effet, lorsqu’une personne attribue des états mentaux comme « La machine *croit que* si je l’éteins, elle ne sera plus jamais rallumée » ou « Le robot *veut que* je désobéisse aux chercheuses et chercheurs », il semble que la personne en question considère déjà que la machine comprend les conséquences liées au fait d’être éteinte : qu’elle comprend la signification d’être éteinte, notamment en comprenant qu’elle ne pourrait plus fonctionner normalement. Il semble aussi que la personne considère que la machine comprend que l’humain peut changer d’avis et décider de ne pas l’éteindre sous ses complaintes : qu’elle comprend que les complaintes ont un pouvoir persuasif. Dans une telle situation, l’interaction avec la machine peut être radicalement modifiée à la suite de l’expression de la machine. Ce qui sera décisif, c’est si la personne à laquelle on demande d’appuyer sur l’interrupteur attribue ou non ces états internes et donc la capacité de comprendre le langage au robot.

[18] Ces attributions servent notamment à faire sens de leurs actions (Horstmann et al., 2018 ; Złotowski, 2016 ; Lee, Liang, 2019).

Que la machine comprenne ou non le langage, elle peut définitivement *manifester une capacité à le comprendre*. Des études très intéressantes ont effectivement contribué à mettre en lumière certaines limites de la compréhension dont peuvent faire preuve ces machines. On a même montré que les humains s'adaptent même aux limites de compréhension des machines, notamment en reformulant la syntaxe de leurs demandes à des assistants vocaux lors de mécompréhensions ou en diminuant les inférences à réaliser par ces machines pour comprendre et répondre adéquatement (Mavrina et al., 2022; Beneteau et al., 2019; Berg et al., 2011). Notre objectif, dans le cadre de ce travail de recherche, sera donc de chercher à savoir en quoi consiste la compréhension dont elles peuvent faire preuve et à la caractériser en lui donnant une définition. Suivant cette idée, nous tenterons, dans le reste de ce mémoire, de formuler une définition fonctionnelle du processus mental de la compréhension. Nous nous questionnerons sur la nature de la compréhension et sur les conditions nécessaires et suffisantes à l'attribution de la capacité à comprendre aux machines intelligentes. En donnant une définition fonctionnelle de la compréhension, réalisable par des machines, nous entendons donner une définition différente de celle qu'on donnerait de la compréhension pour les humains. Nous partons donc en présupposant que la compréhension dont peut faire preuve l'IA n'est pas celle dont fait preuve l'humain, mais que l'IA peut réaliser un niveau de manipulation de la langue assez élevé pour qu'on lui attribue la capacité de comprendre, pour qu'on lui attribue la capacité à réaliser des processus fonctionnellement équivalents à la compréhension. De là, nous analyserons des études portant sur des interactions humains-machines et les mettrons en relation avec des travaux importants en philosophie de l'esprit et du langage, pour définir les modalités nécessaires à l'attribution de cette capacité.

Concrètement, dans le chapitre II, nous nous pencherons sur les façons par lesquelles certaines machines dotées d'IA peuvent manifester de la compréhension à travers le traitement syntaxique. Pour ce faire, nous présenterons différentes notions philosophiques importantes : la

célèbre critique « searlienne » de l'IA (l'argument de la chambre chinoise), les principes de compositionnalité et de contexte, la notion de « forme logique », les marqueurs syntaxiques (les fonctions syntaxiques et les rôles sémantiques), puis la notion de « relation ». À la fin de ce chapitre, nous serons en mesure de formuler une première partie pour notre définition fonctionnelle de la compréhension attribuable à l'IA, fondée dans l'idée que les machines peuvent manifester, à travers le respect des règles syntaxiques, une capacité à comprendre les relations exprimées dans le langage. Une deuxième partie de notre définition sera formulée dans le chapitre III, où nous nous intéresserons à la question de savoir comment une machine peut manifester de la compréhension en démontrant une capacité à bien représenter les significations des mots et expressions qu'elle traite. Ainsi, nous aborderons la notion de « représentation » (mentale et sémantique) et les difficultés liées à l'idée voulant que des machines puissent en avoir, en plus des notions de « référence », de « contenu descriptif », puis de « concept ». Nous en arriverons à proposer qu'une machine, pour manifester de la compréhension, doive pouvoir démontrer une capacité à entretenir et à communiquer des représentations sémantiques fonctionnellement équivalentes à nos représentations mentales (comme nos concepts) et à les utiliser pour réaliser des fonctions cognitives elles aussi équivalentes aux nôtres. Enfin, le chapitre IV, quant à lui, se concentrera sur l'usage plus pragmatique du langage. Nous y présenterons clairement la notion phare d'intention de communication. Nous compléterons ainsi notre définition fonctionnelle de la compréhension en soutenant qu'une machine, pour manifester de la compréhension linguistique, doive pouvoir se montrer capable de comprendre des intentions de communication. Nous montrerons alors qu'elle peut le faire en traitant adéquatement des actes de parole, de l'information implicite et des règles d'usages. Ainsi, la définition fonctionnelle de la compréhension qui résultera du présent mémoire impliquera des notions et des conditions nécessaires tirées des trois dimensions du langage : syntaxique, sémantique et pragmatique.

# CHAPITRE 2 — L'ANALYSE SYNTAXIQUE ET LA COMPRÉHENSION

## 2.1 Introduction

Dans le chapitre précédent, nous avons présenté les raisons pour lesquelles il est pertinent de nous intéresser à la question de la nature de la compréhension pour mieux traiter le phénomène de l'attribution d'états mentaux aux machines fonctionnant grâce à l'intelligence artificielle. L'idée principale est que lorsqu'une personne attribue des états mentaux à une machine à partir de ce que la machine dit ou de ce qu'elle fait, cela semble déjà présupposer que cette personne considère que la machine peut comprendre le langage, notamment parce que la compréhension semble fondamentale à nombre d'états mentaux. Ce mémoire vise à clarifier ce qui est entendu par « comprendre ». Suivant cette idée, la question qui intéressera le présent chapitre est la suivante : comment est-ce que le traitement de mots et de phrases en suivant des règles syntaxiques peut manifester une capacité à comprendre une langue ?

Pour répondre à cette question, nous procéderons en trois temps. D'abord, nous nous intéresserons à quelques notions philosophiques et théoriques fondamentales pour traiter adéquatement de la question. Nous nous pencherons d'abord sur une thèse de John Searle qui veut que la capacité qu'ont certaines machines à utiliser des règles syntaxiques pour recevoir des entrées et fournir des sorties dans une langue partagée avec l'humain soit insuffisante pour être considéré comme la capacité à comprendre cette langue (Searle, 1980 : 422, 423). Nous réinterpréterons cette

thèse, afin de montrer les difficultés qu'elle soulève pour notre recherche. De là, nous présenterons trois notions qui nous permettront d'offrir une réponse nuancée à la thèse de Searle. Nous présenterons d'abord les principes de compositionnalité et de contexte, puis montrerons que la capacité d'une machine intelligente à utiliser des règles syntaxiques pour traiter des expressions et pour en formuler de nouvelles peut jouer un rôle dans la perception que nous avons de son niveau de compréhension, notamment en raison de l'interaction entre les structures syntaxiques et la signification. Ensuite, nous montrerons comment les structures syntaxiques utilisées dans une langue naturelle permettent d'exprimer des relations entre les mots d'une phrase. Partant, nous présenterons ce en quoi consiste l'analyse syntaxique (*syntactic parsing*) que peut réaliser un programme d'IA contemporain et comment cette analyse permet l'extraction d'informations ayant trait aux relations exprimées dans le langage. Enfin, nous conclurons cette première partie de notre chapitre en présentant une première hypothèse plausible et grandement pertinente, basée sur ces concepts philosophiques théoriques : l'hypothèse voulant que la manipulation et l'interprétation syntaxique par l'intelligence artificielle manifestent de la compréhension de différentes relations exprimées dans le langage.

Pour continuer, nous nous éloignerons de la littérature philosophique et linguistique pour davantage nous tourner vers la littérature en sciences informatiques. La partie suivante de ce chapitre présentera alors des arguments, à partir d'exemples concrets, allant en faveur de l'idée voulant que certains systèmes d'IA puissent effectivement manifester de la compréhension d'au moins trois types de relations (les relations spatiales/topologiques et temporelles, les relations agentielles entre des agents et des objets, puis les intentions de communication).

Enfin, nous clorons ce chapitre en nous questionnant sur les limites de notre réflexion à propos du rôle de la manipulation syntaxique grâce à l'intelligence artificielle pour la perception, chez des machines, d'une capacité à comprendre une langue. De là, nous présenterons des

difficultés liées au fait que la capacité à manifester la compréhension d'une langue semble souvent nécessiter la capacité à entretenir certaines connaissances, notamment à propos de la signification de certains mots. Nous terminerons en retournant à notre question de départ en nous demandant quelle place la manipulation syntaxique devrait occuper au sein d'une définition fonctionnelle de la compréhension et proposerons ainsi une première partie de définition.

## 2.2 Perspectives philosophiques sur ce que veut dire « comprendre la syntaxe »

### 2.2.1 La chambre chinoise et le problème du traitement des symboles

En s'inspirant du fameux test de Turing (Turing, 1950), John Searle fut l'un des plus importants critiques des théories allant en faveur de la possibilité de créer une IA forte. À l'époque, l'un des objectifs de Searle était de proposer qu'il soit impossible pour un programme informatique de générer de l'intentionnalité chez une machine (Searle, 1980 : 422, 423). L'intentionnalité étant nécessaire pour comprendre une langue, il apparaît, selon lui, impossible qu'une machine puisse véritablement comprendre ce qu'elle dit, lit ou entend. En effet, selon lui, pour penser (pour avoir un « esprit intentionnel ») et donc pour comprendre, on doit avoir un cerveau biologique[19]. Suivant cette idée, au tournant des années 1980, Searle élabore l'expérience de pensée de la chambre chinoise qui vise à montrer que la manipulation de symboles à l'aide de règles syntaxiques est insuffisante pour constituer de la compréhension, et ce, même si nous pouvons avoir l'impression que c'est le cas, à la vue d'une machine qui réalise cette manipulation (Searle, 1980). Le point de Searle est clair, les outils (les règles syntaxiques) que l'on remet à la personne dans la chambre équivalent au programme informatique que suivent les machines dotées d'IA, montrant ainsi que

[19] Il garde tout de même la porte ouverte à l'idée qu'une machine qui aurait les mêmes pouvoirs causaux qu'un cerveau pourrait peut-être générer une forme d'intentionnalité (Searle, 1980 : 417).

la capacité à respecter ces règles n'équivaut pas à comprendre une langue : la manipulation de symboles conformément à des règles n'entraine pas nécessairement de compréhension de ces symboles. Voici un résumé de l'expérience de la chambre chinoise construit à partir de la présentation qu'il en fait dans « Minds, brains, and programs » (1980 : 417, 418):

Une personne qui ne comprend absolument rien au chinois se trouve dans une chambre verrouillée avec un ordinateur et un écran sur lequel apparaît du texte. Des personnes inconnues lui transmettent une première grande quantité de texte en chinois. On lui transmet ensuite d'autres textes en chinois, ainsi qu'un ensemble de règles en anglais (sa langue maternelle) pour mettre en relation le premier lot de texte avec le deuxième. On lui donne ensuite un troisième lot de textes en chinois, ainsi que d'autres règles pour mettre en relation les symboles du troisième lot avec ceux des deux autres lots. Les personnes à l'extérieur de la chambre nomment le premier lot un *script*, le deuxième une *histoire* et le troisième les *questions*. La personne doit donc suivre les règles et traiter les symboles pour formuler de nouveaux textes. Les gens à l'extérieur de la chambre nomment les textes que la personne à l'intérieur leur soumet les *réponses*. Ultimement, puisque la personne à l'intérieur de la chambre suit des règles de traitement (des règles syntaxiques) pour créer des réponses en chinois (des sorties) aux textes chinois qu'elle a reçus, les personnes à l'extérieur de la chambre ne peuvent pas deviner que la personne ne comprend pas le chinois. Les réponses qu'elles reçoivent sont aussi bien formulées que celles que donnerait un locuteur compétent de la langue qui aurait bien compris les textes. À un tel point, que ces personnes ne peuvent alors pas distinguer entre les réponses de la personne à l'intérieur de la chambre et les réponses d'une personne qui aurait comme langue maternelle le chinois. Néanmoins, selon Searle, la personne enfermée dans la chambre chinoise qui répond aux questions de ses interlocuteurs de façon à manifester de la compréhension ne comprend absolument rien aux entrées et aux sorties

qu'elle génère. Elle manipule des symboles non-interprétés ![20] Exactement de la même manière que peuvent si bien le faire les ordinateurs, nous dit Searle (Searle 1980 : 417, 418)

L'expérience de pensée de Searle impose des contraintes de départ pour notre investigation de l'attribution de la capacité à comprendre à l'IA. Elle nous permet d'abord de prendre une distance par rapport aux travaux de Searle : notre objectif est de montrer ce qu'une machine doit faire pour que nous soyons portés à considérer qu'elle comprend une langue et non pas pour qu'elle la comprenne consciemment et exactement comme le peut une personne humaine, ce qui nécessiterait fort probablement la capacité à avoir des expériences phénoménologiques du monde et de réels états mentaux. C'est en tout cas ce que semble penser John Searle en parlant de cette tendance que nous avons à attribuer des capacités mentales comme celle de la compréhension à des objets :

> *We often attribute "understanding" and other cognitive predicates by metaphor and analogy to cars, adding machines, and other artifacts, but nothing is proved by such attributions. We say, "The door knows when to open because of its photoelectric cell," "The adding machine knows how (understands how, is able) to do addition and subtraction but not division," and "The thermostat perceives changes in the temperature." The reason we make these attributions is quite interesting, and it has to do with the fact that in artifacts we extend our own intentionality; our tools are extensions of our purposes, and so we find it natural to make metaphorical attributions of intentionality to them; but I take it no philosophical ice is cut by such examples.* (Searle, 1980 : 419)

La question de l'intentionnalité n'est effectivement pas au cœur de notre recherche comme elle a pu l'être pour John Searle dans les années 1980. Ce qui est au centre de la nôtre est la question de

[20] Searle propose aussi que les réponses formulées en chinois et celles formulées en anglais (si on demandait, par exemple, à la personne de réaliser l'exercice dans sa langue maternelle) sembleraient aussi bonnes les unes que les autres pour un observateur extérieur, mais pour des raisons différentes. Dans le premier cas, ce sont les opérations computationnelles bien réalisées qui permettent de formuler des réponses adéquates, alors que dans le deuxième cas c'est l'interprétation des questions et du script permet de formuler des réponses correctes.

savoir ce qu'il reste de pertinent à expliquer à propos du rôle de la maîtrise syntaxique dans l'attribution de la capacité à comprendre en contexte d'interaction humains-machines. Elle permet ainsi de poser une question à laquelle ne répond pas l'article-phare de Searle : la question de savoir ce que manifeste la maîtrise de règles syntaxiques pour une observatrice ou un observateur extérieur et donc ce qu'une machine paraît comprendre quand elle se montre capable de traiter et de suivre les structures syntaxiques des expressions d'une langue. Ainsi, en gardant en tête l'idée phare de Searle voulant que la manipulation de symboles en suivant des règles syntaxiques soit insuffisante pour comprendre les symboles en question, nous tâcherons de montrer que la maîtrise de la syntaxe est néanmoins nécessaire à l'attribution de la compréhension aux machines intelligentes. Nous viserons donc à mettre de l'avant l'idée voulant qu'elle constitue une raison partielle (ou l'une des raisons nécessaires) nous poussant à attribuer aux machines la capacité de comprendre une langue. La suite de ce chapitre servira à expliquer pourquoi c'est le cas en fournissant une analyse du rôle de la syntaxe dans la manifestation de la compréhension. Débutons en montrant comment la syntaxe interagit avec la sémantique en nous intéressant aux principes de compositionnalité et de contexte.

### 2.2.2 Les principes de compositionnalité et de contexte

Plusieurs auteurs et autrices se sont penchés sur l'interaction entre les trois dimensions du langage qui nous intéresseront dans ce mémoire (syntaxique, sémantique et pragmatique). Le principe de compositionnalité, par exemple, est issu de ce genre de travaux. Il cherche à expliquer comment la syntaxe d'une langue contribue à donner un sens à ce que l'on dit et comment sa connaissance nous permet de comprendre et de formuler des phrases que nous n'avons auparavant jamais lues ou entendues.

Les premières formulations occidentales du principe de compositionnalité nous proviennent du 12e siècle, mais celui-ci fut plus exhaustivement et explicitement développé, entre autres, par Gottlob Frege, Hilary Putnam, Jerrold Katz, puis Jerry Fodor (Pagin, Westerståhl, 2010 : 250, 251). Étant aujourd'hui assez largement admis en philosophie et en linguistique, ce principe trouve maintenant différentes formulations. En voici une assez générale : « *The meaning of a complex expression is determined by its structure and the meanings of its constituents.* » (Gayral et al., 2005 : 83) Fodor et LePore, quant à eux, formulent le principe de façon plus précise :

« *Compositionality is the property that a system of representation has when (i) it contains both primitive symbols and symbols that are syntactically and semantically complex; and (ii) the later inherit their syntactic/semantic properties from the former.* » (Fodor, LePore, 2002 : 1) Le principe nous dit alors que la façon dont sont agencées les parties d'une expression (que l'expression en question soit un mot ou une phrase), alliée à la signification des constituants utilisés pour la construire[21], déterminent sa signification (Elugardo, 2005 : 61). Dès la première page des *Compositionality Papers*, manuscrit influent traitant du concept de compositionnalité, Jerry Fodor et Ernest LePore présentent un exemple très simple, mais éclairant pour le comprendre rapidement. L'exemple porte sur le terme *DOGS*. Les auteurs proposent que *DOGS* soit un symbole complexe formé par deux symboles *DOG* et *S* et leur organisation syntaxique (Fodor, LePore, 2002 : 1). La signification de *DOGS* est donc formée par *DOG* signifiant un animal à quatre pattes, domestique, etc. et *S*, qui est une marque du pluriel lorsqu'on la place à la fin d'un mot (Fodor, LePore, 2002 : 1). La signification de l'expression complète exprime alors plusieurs chiens. L'exemple avec le mot *DOGS* permet de voir rapidement l'importance de la syntaxe pour le traitement correct d'une

[21] C'est l'argument de la productivité du langage : la compréhension individuelle des parties d'une expression nous permet de construire la signification globale de l'expression. Voir : (Frege, 1980: 79)

entrée par un agent artificiel. En effet, imaginez que vous parliez en anglais de vos chiens à un assistant vocal intelligent et que ce dernier ne tiendrait pas compte du rôle que joue le S dans l'expression DOGS, alors que vous lui indiquez que vous possédez des chiens. L'assistant pourrait, par exemple, vous répondre : « *What is the name of your dog ?* » Probablement que vous tenteriez alors de lui répéter l'information d'une autre façon en lui disant par exemple : « *I actually have multiple dogs* » (en modifiant la syntaxe). Vous considéreriez probablement qu'il n'a pas compris ce que vous vouliez dire.

Retenons donc les idées suivantes : la structure syntaxique d'une expression détermine en partie sa signification (Gayral et al., 2005 : 84) et ces structures nous permettent de comprendre et de formuler des expressions nouvelles[22] (Szabó, 2022 : 3.2). En ce sens, il semble que de saisir la structure syntaxique d'une expression adéquatement, c'est déjà manifester une saisie au moins partielle de la signification de l'expression en question. De plus, puisque les machines dotées d'IA avec lesquelles nous interagissons peuvent, pour la plupart, recevoir de nouvelles phrases et en formuler de façon autonome, le principe de compositionnalité est aussi éclairant sur cet aspect.

Il est pertinent, ici, d'aborder un deuxième principe complémentaire : le principe de contexte, qui propose que le langage fonctionne de la façon inverse. Ainsi, pour comprendre la référence d'un des mots présents dans une phrase il faut d'abord comprendre le tout (la signification globale de l'expression/la pensée exprimée). L'articulation des deux principes peut paraître

[22] Cette idée est fondée en partie sur la thèse voulant que le langage soit *systématique,* notamment parce qu'il existerait des modèles (*patterns*) que suit le langage et qui, une fois que nous les connaissons, nous permettent de comprendre de nouvelles phrases qui les suivent (Szabó, 2022 : 3.2). Cette idée de systématicité rappelle d'ailleurs le concept de « formes logiques » développé longuement par Bertrand Russell, mais aussi utilisé par Wittgenstein, servant notamment à expliquer comment la connaissance de certaines formes d'organisation des faits guide la compréhension de propositions (Irvine, 2009 : 21, 154; Wittgenstein, 1961 : 94).

déstabilisante[23], mais il faut se rappeler que Frege, le théoricien probablement le plus important du principe de contexte, soutenait aussi le principe de compositionnalité. L'interaction entre ceux-ci reste mystérieuse dans les textes du philosophe allemand, mais est clarifiée par le commentateur Gilead Bar-Elli, qui propose deux façons d'analyser le principe de contexte : soit on l'analyse comme expliquant ce qu'est le sens des mots, soit on l'analyse comme étant à propos de la notion de référence (Bar-Elli, 1996 : 112, 122). La première analyse propose que le principe de contexte serve à montrer ce que c'est que d'avoir un sens pour un mot. Avoir un sens, c'est contribuer d'une certaine manière aux sens (aux pensées) exprimés par les phrases dans lesquelles on trouve le mot en question (Bar-Elli, 1996 : 112, Dummett, 1981 : 4). La deuxième analyse, quant à elle, nous semble encore plus intéressante : elle propose que le principe dicte que notre saisie des référents des mots est toujours déterminée par la signification globale des énoncés où ils se retrouvent. Ainsi, notre capacité à saisir des sens (d'expressions complètes) nous permet d'attribuer (*ascribe*) les références aux mots qui les composent (Bar-Elli, 1996 : 122). On peut d'ailleurs effectivement considérer que le contexte propositionnel dans lequel se trouve un mot détermine sa syntaxe et donc indirectement son sens et sa référence. Cela est observable, notamment lorsque l'on s'intéresse aux mots normalement utilisés comme des termes singuliers ou des noms propres. En guise d'exemples, on peut penser à « l'auteur de ce mémoire » ou à « Jérémie Garceau ». Lorsque ces mots agissent syntaxiquement (*behave syntactically*) comme des termes singuliers ou des noms propres à l'intérieur d'une phrase et que la proposition qui y est exprimée est vraie, alors ces termes réfèrent à un seul et unique objet (Milne, 1986 : 492). À l'inverse, si on se retrouve face à une phrase comme la suivante : « "Jérémie Garceau" contient 14 lettres », ce n'est plus le cas que

[23] Les deux principes attribuent la priorité d'interprétation à des choses différentes, soit les sens des mots pris individuellement nous permettent de comprendre le sens d'une phrase complète, soit le sens d'une phrase complète nous permet de comprendre le sens des mots qui la constituent (Gilead Bar-Elli, 1996 : 116, 119, 120).

« Jérémie Garceau » se comporte syntaxiquement comme un nom propre. Il ne sert alors plus à référer à quelqu'un. En ce sens, ce sont les phrases dans lesquelles on les retrouve qui font en sorte ou ne font pas en sorte que ces mots agissent comme des termes singuliers ou des noms propres. La syntaxe d'une expression est en effet modifiée lorsqu'on la sort de son contexte propositionnel usuel. La modification de sa syntaxe causée par le contexte dans lequel on la retrouve modifie alors aussi indirectement son sens et donc sa référence.

Maintenant, ce qu'il faut retenir de ces deux principes qui servent à décrire l'interaction entre les significations de phrases complètes et les significations de leurs constituants, c'est qu'il y a naturellement toujours une interaction entre les premières et les dernières, et ce, que l'on souscrive au principe de compositionnalité, au principe de contexte, ou bien aux deux. De là, les deux principes présentés ci-haut témoignent de la nécessité de la capacité à traiter adéquatement des structures syntaxiques pour manifester une capacité à comprendre le langage, parce que celles-ci font toujours partie des expressions traitées et influencent donc nécessairement les significations de ces dernières.

### 2.2.3 La syntaxe, les faits et les formes logiques: le rapport entre le langage et le monde

Pour continuer, il est utile de revenir aux débuts de la philosophie analytique pour avoir une vision complète de ce que permet l'analyse du langage. Certains des premiers philosophes analytiques, tels que Bertrand Russell et le premier Wittgenstein se sont intéressés au rapport entre le langage et la réalité. Par exemple, dans son *Tractatus Logico-Philosophicus*, Ludwig Wittgenstein soutient une thèse fondamentale voulant que la structure d'une phrase qui asserte un fait et la structure réelle du fait puisse être communes/identiques (Russell, 2010 : 8). Grâce à cette identité de structure, une phrase peut décrire la réalité telle qu'elle est, au sens où une phrase assertorique dit des choses que les noms constituant de celle-ci désignent qu'ils sont dans une

certaine relation. Dans cette perspective, pour comprendre une proposition, saisir ce qu'elle dit, il faut connaître les référents de ses noms et en saisir la forme logique. (Russell 2010 : 8; Wittgenstein, 2010 : 4.2211, 3.24). Comprendre une proposition, dans ce cas, revient à comprendre les constituants et la forme de celle-ci (Wittgenstein 1961 : 94). La compréhension de la syntaxe est alors partie intégrante de la compréhension des propositions qui parlent de la réalité. La structure syntaxique d'une proposition contribue à représenter la structure du monde.

Dans la même optique, l'idée d'une structure logique du monde que le langage permet de décrire fut en grande partie développée par Bertrand Russell. Il proposait effectivement que pour comprendre, par exemple, un nom, il faille savoir ce pour quoi il tient lieu dans la réalité, tandis que pour comprendre des prédicats comme des adjectifs ou des mots qui réfèrent à des relations, il faut comprendre des propositions dans lesquels nous les trouvons (Russell 1919 : 33, 34). En ce sens, pour comprendre la relation « être plus grand que » il faut connaître des propositions dans lesquelles nous trouvons cette relation (Linsky, 2003 : 379, 380). Cette théorie de la compréhension des mots prend son sens lorsque nous nous intéressons à la compréhension plus générale des énoncés (des phrases complètes). Les énoncés, dans le modèle de Russell, nous permettent de comprendre des propositions. Une proposition, pour Russell, exprime une partie de ce qu'une personne veut dire par l'expression d'un énoncé (Russell 2002 : 142). De là, deux énoncés différents peuvent exprimer une même proposition (Russell 2002 : 142) : deux personnes peuvent effectivement dire au moins en partie la même chose (la même proposition), mais différemment, dans deux langues distinctes, par exemple. Dans l'exemple suivant donné par Russell, on retrouve des phrases différentes, mais qui expriment toutes une même proposition à travers la prédication (exprimée par le verbe « être »), reliant sujet (les mendiants) et prédicat (être des cavaliers) :

*« Les mendiants sont des cavaliers. »*

*« Les mendiants seraient des cavaliers. »*
*« Que les mendiants soient des cavaliers. »*
*« Est-ce que les mendiants sont des cavaliers? » (Russell, 2002 : 140)*

De ce que nous venons de présenter, il faut retenir deux choses : d'abord, pour Wittgenstein et Russell, l'analyse logique du langage permet de découvrir des relations logiques qui peuvent exister entre des objets de la réalité : le langage est lié au monde, car il peut décrire son organisation. Ensuite, des phrases différentes peuvent exprimer une même proposition et donc décrire une même façon de comprendre la réalité.

### 2.2.4 Les relations dans les différentes structures de la syntaxe : Noam Chomsky et les niveaux de structures syntaxiques

Cette méthode d'analyse du langage permettant de découvrir des structures guidant l'interprétation et la compréhension du langage, puis la compréhension de la réalité fut féconde, tout au long du vingtième siècle, notamment avec le développement de différentes branches de la linguistique. Dans *Aspects of the theory of syntax* (1965) et *Le langage et la pensée* (2009), le linguiste et philosophe Noam Chomsky, par exemple, développe une distinction entre deux niveaux de structures syntaxiques. Il propose que les différentes grammaires superficielles (*surface structures*) (différentes façons d'organiser et d'identifier les constituants d'une phrase) soient générées par la structure profonde (*deep structures*) qui est sous-jacente à la phrase et qui détermine en grande partie le sens de la phrase (Chomsky, 2009 : 69-72, 77-79; Chomsky, 1965 : 136). En ce sens, l'entrée (*input*) utilisée pour construire la sémantique d'une expression devient la structure syntaxique profonde de celle-ci[24] (Partee, 2015 : 191). C'est à ce niveau profond, selon Chomsky,

24 En linguistique, l'origine de cette idée est normalement attribuée à Jerrold Katz et Paul Postal et on en parle sous les noms *Katz-Postal hypothesis* ou *Katz-Postal principle.* Voir : Katz, Postal, 1964.

que nous découvrons des fonctions syntaxiques comme : « être l'objet de » ou « être le sujet de » (Chomsky, 1965 : 115).

Suivant ces idées, le philosophe et linguiste Jerrold Katz, quant à lui, décrit, dans le même sens que Russell et le premier Wittgenstein, le concept de « compréhension linguistique » comme un acte par lequel nous inférons les formes logiques cachées sous les structures grammaticales de surface des phrases que nous comprenons (Katz 1990 : 23). En plus de Katz, Paul Postal (Katz, Postal, 1964), Gilbert Harman (1970), Kirk Ludwig (2012), et nombre d'autres philosophes se sont penchés sur la question de savoir comment l'interprétation sémantique et, plus largement, la compréhension linguistique peut s'expliquer par un processus d'analyse de structures logiques ou de structures profondes sous-jacentes aux phrases grammaticales que nous formulons dans nos langues maternelles.

En effet, la syntaxe d'une expression participe à la construction de sa signification en exprimant des relations entre des mots qui réfèrent à des objets de la réalité (incluant des entités et des choses). Des auteurs contemporains comme Chomsky et Katz, expliquent cela notamment en introduisant l'idée de différents niveaux structurels de la syntaxe qui nous permettent de décrire comment les mots d'une même expression entretiennent des relations que nous pouvons identifier et caractériser grâce à des outils comme les fonctions syntaxiques. Suivant cette idée, une machine qui saurait interpréter adéquatement les structures syntaxiques de phrases pour en « interpréter » le sens devrait être en mesure de traiter les structures syntaxiques de surface dans le but d'en retirer une structure plus parlante: une structure qui décrit les relations entre les mots, notamment grâce à des fonctions syntaxiques. Maintenant que nous avons mis en place les outils philosophiques pertinents pour notre investigation, voyons comment les programmes d'intelligence artificielle contemporains ont la capacité d'analyser les structures syntaxiques.

### 2.2.5 Introduction à l'analyse syntaxique

Les programmes d'intelligence artificielle comme Siri d'Apple, Alexa d'Amazon et bien d'autres doivent, pour communiquer avec nous adéquatement, analyser les structures syntaxiques des phrases qu'on énonce et en construire qui respectent les règles qui organisent la langue. Pour analyser la structure syntaxique d'une phrase, le système doit construire un « arbre » (*syntactic parsing*) dans lequel sont placés les constituants de la phrase. En linguistique, un arbre de cette forme est appelé *constituent structure* ou *functional structure* (Kaplan, Bresnan, 1995 : 3-6). Il existe différentes façons de diviser une phrase pour en caractériser la structure et donc différentes façons de caractériser chacun des constituants qui la composent en utilisant des marqueurs (*phrases markers*) et différents types d'arbres (dont les deux nommés plus haut)[25].

En ce sens, un adepte de la grammaire traditionnelle nous recommanderait peut-être de créer des systèmes qui peuvent classer les termes d'une expression sous des concepts comme sujet et prédicat pour identifier les rapports entre eux (Frede, 1975; Feuillard, 2009 : 95). Une adepte de la grammaire générative, quant à elle, proposerait peut-être d'abord de grouper les termes en utilisant des marqueurs comme le groupe du nom (*noun phrase/NP*), le groupe du verbe (*verb phrase/VP*), et ainsi de suite (Higginbotham, 1983 : 148-151). La principale différence entre ces deux méthodes est que d'un côté on demande de classer les constituants sous des notions fonctionnelles (sujet, objet de, prédicat, etc.), tandis que de l'autre on évite d'attribuer des implications relationnelles aux constituants de la phrase (Feuillard, 2009 : 95). L'utilisation de notions fonctionnelles pour analyser les structures syntaxiques de phrases permet effectivement d'identifier de l'information comprise dans ces phrases en question, alors que l'identification de

[25] Les principales sont énoncées dans Prakash, 2013 : 62. L'existence de ces différentes approches concurrentes de la grammaire s'explique, entre autres, par une volonté des linguistes de décrire le plus adéquatement possible les structures utilisées pour former des phrases grammaticales et les comprendre (Katz, 1971 : 108).

« constituants » permet une catégorisation syntaxique indépendante de tout contexte. L'analyse exhaustive des multiples approches de la grammaire est hors de la portée de ce travail. Nous souhaitons plutôt nous concentrer sur ce que manifeste la maîtrise des règles syntaxiques dont fait preuve un locuteur ou une locutrice compétente d'une langue comme le français ou l'anglais. Pour les programmes d'IA, cette maîtrise passe par l'analyse syntaxique et donc par différentes façons (telles que celles nommées plus haut) d'identifier comment les mots et groupes de mots contribuent à la construction de la signification des phrases.

Ainsi, en nous concentrant plus précisément sur deux façons principales d'identifier les structures syntaxiques que peuvent utiliser des programmes d'IA pour traiter une langue, nous nous retrouvons sur la piste de l'hypothèse que nous souhaitons analyser : celle voulant que l'analyse syntaxique correcte permette de manifester une compréhension des relations entre des termes et donc des relations réelles entre des entités, des objets et des événements[26]. Ces deux façons sont l'identification des fonctions syntaxiques[27] que remplissent les termes d'une expression donnée et l'identification des rôles sémantiques[28] que joue chacun de ces termes. Plus précisément, les fonctions syntaxiques et les rôles sémantiques sont des entités relationnelles qui peuvent être utilisées par des systèmes autonomes lors de l'analyse syntaxique (*syntactic parsing*) (Feuillard, 2009 : 99). Des exemples paradigmatiques de fonctions syntaxiques sont le sujet et l'objet, tandis que des exemples de rôles sémantiques sont ceux d'agent et de patient. Les rôles sémantiques,

---

[26] Pour utiliser un vocabulaire philosophique comme celui du premier Wittgenstein, nous pourrions dire que les entités et les choses sont des objets, alors que les événements sont des faits (Wittgenstein, 2010 : 4.1272).

[27] Les rôles sémantiques peuvent aussi être utilisés en philosophie analytique pour analyser les formes logiques des expressions (par exemple : Ludwig, 2012). On parle aussi parfois de relations syntaxiques ou de relations grammaticales. Les deux concepts « relations syntaxiques/grammaticales » et « fonctions syntaxiques », selon l'approche de la grammaire que nous adoptons, seront plus ou moins équivalents (voir : Feuillard, 2009 : 96-98).

[28] On parle aussi parfois, surtout dans la littérature en anglais, de rôles thématiques (*thematic roles*).

comme ceux d'agent et de patient, par exemple, peuvent être respectivement attribués au sujet et à l'objet contenus dans une phrase donnée (Williams, 1984 : 639, 640, 657), parce que les rôles sémantiques se superposent aux fonctions syntaxiques, sans s'y réduire (Feuillard, 2009 : 99). Selon l'approche de la grammaire que nous adoptons, la liste de rôles sémantiques utilisables sera plus ou moins longue. La grammaire relationnelle, par exemple, en établit une liste exhaustive: « expérimenteur » (*experiencer*), but (*goal*), emplacement (*location*), bénéficiaire, instrument, etc.[29] (Kracht, 2002: 253; Gildea, Jurafsky, 2002 : 3).

### 2.2.6 La compréhension des relations à travers l'analyse syntaxique

Nous venons de voir différentes façons d'analyser les constituants d'une phrase sans faire directement appel à la signification des mots contenus dans celle-ci. Rappelons-nous : l'incapacité d'une machine dotée d'un programme d'intelligence artificielle d'accéder à la signification des mots et des phrases qu'elle traite était précisément ce qui faisait dire à Searle qu'une telle machine ne comprend absolument rien au langage. Cela étant dit, l'analyse des fonctions syntaxiques et des rôles sémantiques permet de rendre compte de relations qui existent entre des mots et groupes de mots à l'intérieur d'expressions complètes en les catégorisant et en les caractérisant[30]. Suivant ces idées, la structure syntaxique d'une phrase est une source d'information primordiale pour un système d'IA qui traite le langage, et ce, même si ce système ne comprendra jamais vraiment ce

---

[29] Des catégories de ce genre sont aussi systématisées dans l'approche fonctionnelle de la grammaire du linguiste anglais Michael A. K. Halliday au sein de laquelle les catégories grammaticales utilisées pour analyser la syntaxe sont présentées comme la réalisation de modèles sémantiques (*semantic patterns*), plaçant ainsi l'interaction entre syntaxe et sémantique au centre de la tâche l'analyse syntaxique (Webster, 2009 : 7).

[30] Elles ne le font toutefois pas de la même manière. Par exemple, les rôles thématiques comme « agent » et « patient » sont plus que des catégories grammaticales qui décrivent les relations syntaxiques entre des termes : ils sont des rôles que l'on attribue aux entités linguistiques, des rôles que les termes assument l'un par rapport à l'autre dans la construction de la signification d'une phrase (Feuillard, 2009 : 98, 99).

que veulent dire les mots qu'il traite comme nous le pouvons en tant qu'humains. En effet, son analyse permet au système d'identifier des relations entre des mots lui fournissant ainsi une base d'informations utiles pour traiter leurs contributions à la signification des phrases analysées[31].

Donnons un exemple fictif pour nous assurer de présenter clairement l'hypothèse qui sera étudiée dans les prochaines pages : imaginons un assistant vocal qui recevrait l'entrée verbale suivante : « les personnes qui sont propriétaires de cette maison arrivent ». Lors du traitement d'une telle expression, un ordinateur pourrait faire erreur et considérer que c'est la maison qui arrive, plutôt que les propriétaires. Un système d'IA, pour être en mesure de fournir une réponse cohérente avec les informations contenues dans cette entrée, devrait être en mesure de traiter des expressions complexes en les considérant comme des ensembles de relations hiérarchiques entre les mots qui les composent (Manning et al., 2020 : 30048). Dans cet exemple, il faudrait que le système associe le verbe « arriver », non pas avec le nom le plus près (à partir de la proximité linéaire dans la phrase), mais avec les mots qui occupent réellement la fonction syntaxique de groupe sujet (*subject phrase*) dans la phrase en question : les personnes propriétaires (Manning et al., 2020 : 30048). On pourrait même programmer une machine pour qu'elle identifie les rôles sémantiques : le groupe sujet « les personnes propriétaires », par exemple, joue le rôle d'agent, car ce sont ces personnes qui performent l'action d'arriver. Dans tous les cas, il faut que l'assistant vocal soit en mesure de bien identifier la structure syntaxique de la phrase pour être en mesure de répondre adéquatement ou de réaliser la fonction attendue par l'utilisateur qui l'informe de l'arrivée des propriétaires. Il est donc plausible que cette capacité, si elle nous permet, en tant qu'humains, de comprendre les expressions que nous entendons d'une certaine manière plutôt que d'une autre, nous permette aussi

[31] L'utilisation conjointe de l'analyse syntaxique et de l'analyse des rôles sémantiques réalisées par l'IA est exemplifiée dans cet article sur le traitement informatique des langues naturelles : Gildea, Jurafsky, 2002.

de juger, lors d'une interaction avec une machine, si elle traite adéquatement ce qu'on lui dit: si elle comprend l'information communiquée.

### 2.3 L'analyse syntaxique adéquate peut manifester une compréhension des relations

#### 2.3.1 Les relations spatiales et temporelles

Jusqu'à maintenant, le point qui émane de ce chapitre est le suivant : il est plausible que lorsqu'une machine applique correctement les règles grammaticales et réalise des analyses syntaxiques adéquates, elle manifeste une capacité à comprendre des relations exprimées dans le langage qui réfèrent à des relations réelles qui existent entre certaines personnes, des choses, des sentiments, etc. Suivant ce que nous avons présenté plus haut à propos de l'analyse logique et à propos de l'analyse artificielle du langage, cherchons maintenant à voir comment l'analyse syntaxique permet concrètement à certaines machines dotées d'IA de manifester une compréhension des relations exprimées dans certaines expressions et de les respecter lors de la formulation de sorties.

Pour préciser les types de relations que l'IA peut sembler comprendre en maîtrisant les règles qui régissent une langue, présentons-en trois, en commençant par les relations spatiales. Pour illustrer notre propos, il est pertinent de nous intéresser au programme d'IA nommé DALL-E. Ce programme sort quelque peu du cadre de recherche que nous avions fixé, au début de la rédaction de ce mémoire, puisqu'il permet de représenter visuellement la signification de diverses combinaisons des mots de la langue anglaise, plutôt que de communiquer avec nous verbalement. Néanmoins, sa capacité d'analyse des structures syntaxiques est utile pour voir comment l'identification adéquate des éléments syntaxiques d'une phrase est primordiale pour pouvoir fournir des réponses/des sorties qui manifestent de la compréhension, du point de vue d'un interlocuteur. Concrètement, ce système d'IA permet de combiner et de modifier des images

existantes et d'en créer de nouvelles de toute pièce à partir de texte descriptif (*text-guided image generation*) fourni par ses utilisatrices et utilisateurs (Ramesh, Pavlov, Goh, Gray, 2021). Ainsi, DALL-E peut combiner la sémantique textuelle à des images, en plus de traiter des structures syntaxiques. DALL-E est donc en mesure de démontrer une certaine capacité à comprendre et à représenter visuellement les relations entre les termes que nous lui fournissons. L'utilisation des prépositions comme *ON, OVER, WITH, BESIDE, IN* ou *INTO* permet effectivement de bien voir la capacité de la machine à représenter la nature de ces relations basiques spatiales/topologiques[32] sous la forme d'images originales. Des exemples le montrent bien :


Figure 1 : image tirée de Conwell et Ullman, 2022 : 8.

[32] Exprimer des relations spatiales est une fonction des prépositions (Pullum, Huddleston, 2002 : 603).

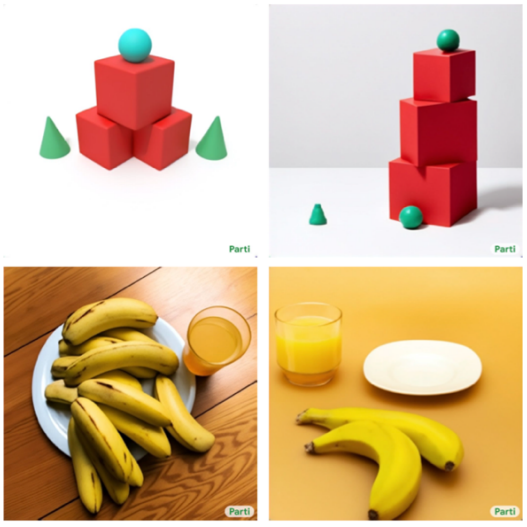

**A**. *A photo of a frog reading the newspaper named "Toaday" written on it. There is a frog printed on the newspaper too.*

Figure 2 : image tirée de Yu et al., 2022 : 3.

En effet, le premier ensemble d'images de la figure 1 manifeste une compréhension de la demande, alors que le deuxième est fautif. Dans la figure 2, DALL-E réussit à respecter chacune des relations exprimées dans le texte.

Malgré son efficacité grandissante, il est fréquent de rencontrer des situations de mécompréhension lorsque nous utilisons DALL-E. La figure ci-bas (figure 3) présente encore une fois l'incapacité du programme à manifester une compréhension des relations spatiales, en raison de diverses erreurs de positionnements des objets les uns par rapport aux autres. Ce qui permettrait à l'IA de remédier à l'erreur, entre autres, c'est d'être entrainée à traiter des phrases ayant ces structures syntaxiques, afin de traiter adéquatement les relations exprimées entre les mots et d'être en mesure, à partir de la position des mots dans ces phrases, d'attribuer la bonne place à chaque image dans la composition globale. Il faut aussi remarquer que la capacité à représenter les directions des relations est particulièrement importante, car les directions des relations permettent

d'indiquer comment seraient combinés les constituants auxquels les expressions réfèrent dans la réalité et comment nous percevrions leur combinaison si le complexe qu'ils forment existait vraiment (Wittgenstein, 2010 : 5.5423).



**F.** (a, b) Two images in the same batch for the prompt *a stack of three red cubes with a blue sphere on the right and two green cones on the left.* (c, d) Two images in the same batch for the prompt *a plate that has no bananas on it. there is a glass without orange juice next to it.* **Failures**: Incorrect relative positioning of objects (a,b,d). Incorrect coloring-to-attribute association (b). Hallucination (of objects specifically mentioned as absent) (c, d).

Figure 3 : image tirée de Yu et al., 2022 : 23.

Enfin, DALL-E n'est qu'un programme permettant de générer des images, mais il représente un excellent exemple de l'importance du traitement des structures syntaxiques pour la manifestation de la capacité à comprendre une langue. Il suffit de nous projeter dans l'avenir et d'imaginer un robot humanoïde qui nous demanderait, en contexte de travail par exemple, « je souhaite avoir l'objet qui est sur la table plutôt que celui sur le sol », pour comprendre l'importance de la maîtrise des structures syntaxiques qui impliquent des mots exprimant des relations basiques spatiales. Ici, nous nous sommes exclusivement concentrés sur la compréhension des relations spatiales, mais nous aurions pu, en nous intéressant au vocabulaire relationnel que peut utiliser un assistant vocal, nous concentrer aussi sur les relations temporelles comme AVANT, APRÈS, PLUS

TARD et PLUS TÔT. Il est tout aussi aisé d'imaginer les mêmes répercussions pour les termes référents à des rapports temporels que celles liées à l'utilisation des termes qui expriment des rapports spatiaux.

### 2.3.2 Les relations agentielles exprimées par les fonctions syntaxiques et les rôles sémantiques

Dans les dernières lignes, nous avons utilisé l'exemple d'un programme permettant de générer des images à partir de texte descriptif, afin de montrer comment la maîtrise des relations et de leurs directions est importante pour l'attribution de la capacité à comprendre. Maintenant, nous souhaitons porter l'attention des lectrices et lecteurs sur un autre type de relations pour lesquelles l'IA peut manifester de la compréhension : les relations dites « agentielles » (*agentic relations*). Elles sont « agentielles », au sens où elles expriment des actions que réalisent des agents[33]. Ces relations sont souvent exprimées par des verbes (Voir Conwell, Ullman, 2022 : 2).

L'exemple suivant est tiré d'un chapitre sur la sémiologie et la syntaxe de David Lockwood. Ce dernier présente des propositions qui contiennent le verbe anglais *ADMIRE*. L'auteur analyse plusieurs propositions, dont celles-ci : « *Hope admires Jane* » et « *Charity admires Reggie* » (Lockwood, 2002 : 310). Dans leur usage ordinaire, les mots HOPE et CHARITY ne jouent pas les rôles de noms propres (du moins, si l'on se fie à un dictionnaire anglais). Cependant, la structure des deux expressions, dû à l'utilisation du verbe ADMIRE, impose la fonction syntaxique de sujets à HOPE et CHARITY (Lockwood, 2002 : 310). Un traitement syntaxique adéquat de la part d'une

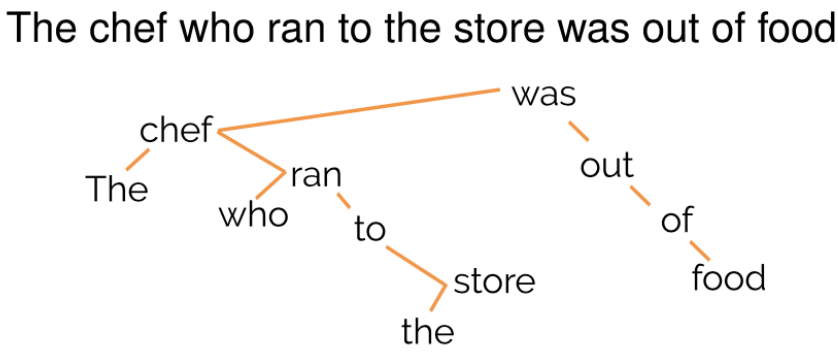
[33] En philosophie, Donald Davidson s'est d'ailleurs intéressé à la question de savoir quelle forme logique suivent les phrases qui présentent des actions, notamment en proposant que certains verbes *introduisent* l'idée d'agentivité dans l'interprétation de la phrase dans lesquelles ils se trouvent (Davidson, 2006 : 49). Cela rappelle l'idée qu'il est utile pour un système d'IA d'identifier, lors de l'analyse syntaxique, les rôles sémantiques.

IA devrait donc en tenir compte et la sortie de l'IA devrait utiliser Hope et Charity comme des sujets ou des noms référant à des personnes. Aussi, l'IA devrait, dans la formulation d'une sortie adéquate, à la suite de la réception de cette entrée, tenir compte de la relation qui existe entre Hope et Jane ou entre Charity et Reggie et tenir compte de la direction de cette relation. La relation ici présentée est, au moins jusqu'à preuve du contraire, asymétrique[34]. L'énoncé ne dit pas que Reggie et Jane admirent aussi leurs comparses respectives. Le fait que Hope et Charity occupent les fonctions de sujets des phrases indique qui admire qui : les fonctions syntaxiques indiquent effectivement la direction des relations d'admiration entre les acteurs en catégorisant les noms sous les catégories de sujets et d'objet, qui sont des informations primordiales à retirer de l'énoncé pour entretenir, par la suite, une conversation cohérente. En situation conversationnelle entre une IA et une personne humaine, la première pourrait demander, par exemple, sortie 1 : « pourquoi Charity admire Reggie ? » ou sortie 2 : « Et vous, qui admirez-vous ? ». Ces réponses seraient effectivement adéquates et cohérentes avec l'information contenue dans les expressions reçues en guise d'entrées. En ce sens, sans même savoir ce que signifie « admirer », une machine peut répondre adéquatement à une affirmation de ce genre, car sa structure syntaxique lui fournit des informations sur les relations qui y sont exprimées et sur leurs directions (comme dans le cas de la sortie 1.) Identifier HOPE, non pas comme un verbe, mais plutôt comme le sujet de la phrase (ou même comme jouant le rôle sémantique d'agent) permet à l'IA de parler adéquatement de HOPE et de son rapport avec les autres termes de la phrase dans la suite de la conversation et donc de manifester une capacité à comprendre des relations entre ces termes et probablement même des relations entre leurs référents dans le monde. En ce sens, la machine montre qu'elle comprend « qui a fait quoi à qui » (Van der

[34] Bertrand Russell élabore clairement la distinction entre les relations symétriques et asymétriques dans Russell, 2009a : 94.

Velde, 2005 : 265). Cependant, on voit rapidement qu'il semble manquer quelque chose pour qu'une machine puisse fournir une réponse comme la sortie 2. En effet, si, sans accès à de l'information sémantique à propos du verbe *ADMIRE*, l'identification de différentes catégories syntaxiques peuvent « informer » un système d'IA sur l'existence et la direction de cette relation, la structure syntaxique ne dit absolument rien sur la nature de cette relation. Le troisième chapitre de ce mémoire approfondira, entre autres, cette difficulté.

Maintenant, il faut noter qu'en plus de manifester de la compréhension de situations où des relations agentielles entre des gens peuvent être décrites, des programmes peuvent manifester de la compréhension de relations agentielles entre des gens et des objets. C'est le cas dans l'exemple suivant, tiré d'un article rédigé par des chercheuses et chercheurs en sciences informatiques, où un chef court au supermarché pour faire des achats[35]. Cet exemple est relativement ambigu, notamment parce que, pour manifester une compréhension du fait que c'est le chef qui manquait de nourriture et non le magasin, on doit comprendre que le magasin est une proposition relative et qu'il n'occupe pas le rôle sémantique d'agent ou la fonction syntaxique de sujet[36] (Manning et al. 2020 : 30046).




[35] L'arbre présenté est nommé un arbre de dépendance (*dependency tree*).

[36] On pourrait aussi programmer une machine pour qu'elle identifie des rôles sémantiques comme « emplacement » (*location*) qu'elle pourrait attribuer aux mots « the store », afin d'éviter cette erreur.

Figure 4 : Cette illustration est tirée de Manning et al., 2020: 30047.

### 2.3.3 Les intentions de communication

Enfin, nous voulons maintenant aborder un dernier type de relations pour lequel l'analyse syntaxique adéquate permet de manifester de la compréhension : les intentions de communication. Plus précisément, il semble que l'IA puisse manifester une capacité à comprendre les relations exprimées dans les propositions servant à réaliser des actes de langage. En effet, un papier paru en 2021 dans le cadre de la *Thirty-Fifth AAAI Conference on Artificial Intelligence* (Wang et al., 2021) présente une expérience réalisée avec des programmes utilisés pour déterminer correctement les intentions des utilisateurs humains à l'aide d'une méthode d'analyse de la syntaxe. Dans cet article, on présente certains problèmes d'interprétation d'un système d'IA de la syntaxe de requêtes ambiguës. On présente, entre autres, une difficulté liée au terme anglais *May* qui peut être employé soit comme un verbe, soit comme un nom. L'identification correcte de ce mot en tant que verbe, dans la phrase « *May I have the movie schedules for Speakeasy Theaters* » permet à l'IA d'interpréter l'énoncé comme une demande, alors que l'identification incorrecte (en tant que nom) peut l'amener à traiter *May* comme un marqueur de temps référant au cinquième mois de l'année (Wang et al., 2021 : 13948).

Si l'on faisait cette requête à un assistant vocal et qu'il nous répondait en donnant la programmation des films pour le mois de mai, nous conclurions qu'il n'a pas compris et nous pourrions reformuler notre demande d'une autre façon en remplaçant *MAY* par *CAN*. Ce que l'on chercherait à faire, c'est encore une fois de modifier la syntaxe pour s'assurer que la machine comprend qu'on lui demande quelque chose et qu'elle comprend la nature de cette demande. L'analyse correcte de la structure de l'énoncé permet d'atteindre le premier objectif (en identifiant *MAY* non pas comme un nom, mais bien comme un verbe qui sert à formuler une demande). Si la

machine n'y arrive pas, ce qu'elle ne semble pas comprendre, c'est que nous réalisons un « acte de langage ». Nous reviendrons, dans le dernier chapitre de ce mémoire, plus en profondeur sur l'utilisation et le traitement du langage pragmatique par des agents artificiels, mais pour l'instant, il faut retenir que lorsque nous posons une question, la philosophie du langage pragmatique nous dit que nous faisons un énoncé qui a une certaine force ou fonction illocutoire. Les actes illocutoires sont formés dans l'action d'énoncer des phrases dans certains contextes, tout en ayant certaines intentions (Searle, 1969 : 24, 25). Pour réaliser l'acte de « poser une question », par exemple, il faut normalement que certaines conditions de réalisation soient satisfaites (dont des conditions dites « préparatoires », liées au contexte d'énonciation) (Searle, 1969 : 57-60, 65). On peut penser, par exemple, à celles d'avoir l'intention de provoquer un certain comportement chez son interlocuteur et d'exprimer un désir (de connaître une information X, par exemple) (Searle, 1969 : 25, 65). Si un acte est réalisé correctement, il est censé avoir des effets illocutoires qui lui sont propres, comme informer l'interlocuteur d'une intention de la personne qui parle, par exemple. Au quotidien, nous sommes constamment en train d'essayer de comprendre les intentions des prises de parole des personnes avec qui nous interagissons. Suivant ces idées, en reprenant l'exemple du traitement d'une requête utilisant le terme *MAY*, il semble plausible que l'incapacité à manifester une compréhension de ce genre d'énoncé manifeste au même moment une incapacité à comprendre les intentions de la personne qui le réalise (l'intention de connaître de nouvelles informations, par exemple). Cette dernière explication pointe vers un type de relation que peut sembler comprendre une machine, notamment grâce à l'analyse syntaxique adéquate : les intentions de communication.

Maintenant, il faut dire que Searle s'est lui-même posé la question de savoir comment les conditions de satisfaction et de réalisation d'actes illocutoires pourraient être représentées dans une

analyse, par exemple, de la structure syntaxique profonde d'un énoncé. Ce dernier reste sceptique[37] sur cette question, sauf en ce qui a trait à certains cas, comme celui des actes illocutoires servant à donner des ordres : des actes impératifs (Searle, 1969 : 64). Notons d'ailleurs que les actes impératifs sont, pour les types de machines qui nous intéressent dans ce mémoire au moins, fort probablement les plus importants, et ce, principalement parce que des machines comme les robots sociaux et les assistants vocaux servent précisément à répondre à des demandes en réalisant certaines fonctions. Ainsi, de ce que nous avons vu à propos des relations agentielles et dans notre brève présentation des actes de langage, retenons que quand une machine analyse correctement ou incorrectement la structure syntaxique des expressions servant à réaliser ces actes, elle peut manifester une capacité ou une incapacité à comprendre la relation qui existe entre l'humain et la machine : dans notre exemple, la machine ne semble pas comprendre la relation agentielle de « demande », notamment parce qu'elle ne semble pas comprendre l'intention de communication de son interlocuteur.

### 2.4 La nécessité et la suffisance de la syntaxe pour l'attribution de la compréhension

Jusqu'à maintenant, dans ce chapitre, nous avons présenté des raisons pour lesquelles la capacité d'analyse des structures syntaxiques semble fondamentalement nécessaire pour que nous considérions que des machines peuvent comprendre une langue. Ce que nous avons présenté pointe effectivement vers l'idée voulant que les machines capables d'analyser correctement les structures syntaxiques réalisent une partie de ce qui est requis pour comprendre : comme si leur capacité d'analyse de ces structures était fonctionnellement équivalente à celle des locutrices et locuteurs

[37] Cette position vient d'un scepticisme de Searle pour la capacité de réduire chacune des règles de réalisation de tous les actes illocutoires à des types illocutoires qui pourraient tous être décrits dans les structures profondes.

compétents d'une langue. Méthodologiquement, nous avons pris comme point de départ la question de savoir ce que manifeste, pour un interlocuteur humain, cette capacité chez une machine, en gardant en tête l'idée que les structures syntaxiques des expressions influencent les significations de ces dernières. Nous avons effectivement fait des fonctions syntaxiques, puis des rôles sémantiques des outils pour exemplifier l'idée voulant qu'une machine puisse manifester une compréhension de relations entre des mots[38].

Nous avons toutefois, à quelques reprises, pris soin de soulever des difficultés avec ces idées en montrant que ces programmes nécessitent parfois l'accès à certaines informations supplémentaires à propos de la sémantique de certains mots pour bien manifester cette capacité. Concluons maintenant ce chapitre en introduisant plus clairement les problèmes liés à l'analyse syntaxique réalisée par des programmes d'IA en montrant des contextes où il semble important que ces programmes aient accès à la signification des mots et des expressions qu'ils manipulent. Nous procéderons en deux étapes : d'abord, nous présenterons le *Winograd schema challenge* en montrant que la capacité d'une IA à traiter adéquatement les structures syntaxiques semble à elle seule insuffisante pour dépasser le problème de l'ambiguïté référentielle et, ainsi, pour manifester

---

[38] Cette thèse, qui nous paraît largement applicable aux machines que nous connaissons aujourd'hui, pourrait aussi être défendue pour en soutenir une autre : on pourrait soutenir que sa compréhension des structures syntaxiques et des relations qui y sont exprimées lui permet aussi de manifester de la compréhension de relations réelles entre des entités, des choses et des états qui sont exprimées dans certaines expressions. En effet, si les relations entre les mots d'une phrase et ces mots eux-mêmes réfèrent effectivement à des objets et à des relations réelles qui existent entre eux dans le monde, cette thèse semble effectivement en découler naturellement. Cependant, l'IA n'a aucune expérience consciente du monde et de ces constituants. Comment pourrait-elle alors comprendre les relations qui y ont lieu ? La piste de solution qui nous paraît la plus plausible pour soutenir une telle thèse est la suivante : l'IA ne fait que *sembler* comprendre les relations réelles. Elle ne les comprend pas réellement, mais les traite de façons adéquates lors d'interactions conversationnelles.

Cela étant dit, dans le cadre de ce travail, nous nous concentrerons davantage sur la capacité de l'IA à comprendre les relations exprimées dans le langage, et ce, pour deux raisons. D'abord, la compréhension des relations extralinguistiques par l'IA, dans le modèle que nous avons présenté au moins, passe par la compréhension des relations linguistiques que l'on retrouve dans les structures d'une langue : la première est donc moins fondamentale que la dernière et dépend de cette dernière. Ensuite, nous souhaitons orienter ce mémoire sur le sujet de la compréhension linguistique plutôt que sur la compréhension du monde.

de la compréhension. Ensuite, nous nous questionnerons plus en profondeur sur la notion de « compréhension » et montrerons que même sans respecter les règles syntaxiques d'une langue, une machine peut sembler comprendre ce qu'on lui dit. Enfin, nous reviendrons sur notre question de départ et formulerons une première partie d'une définition fonctionnelle de la compréhension qui rend compte de notre tendance à attribuer cette capacité à des agents artificiels.

### 2.4.1 Le Winograd Schema Challenge, l'ambiguïté et les connaissances d'arrière-plan

Le *Winograd Schema Challenge* est une épreuve alternative au test de Turing, développée par des chercheurs en informatique, qui porte principalement sur l'ambiguïté référentielle[39] (Levesque et al., 2012 : 552). Voici un exemple tiré de l'article en question pour lequel un programme d'IA doit tenter de répondre à la question présentée, à partir d'un choix de deux réponses potentielles :

« *Joan made sure to thank Susan for all the help she had given.*

*Who had given the help?*

*Answer 0 : Joan*

*Answer 1 : Susan* » *(*Levesque et al., 2012 : 554, 555)

L'un des objectifs que se donnent les auteurs est celui de montrer que la « réflexion » est nécessaire pour répondre à cette question. Clairement, la bonne réponse est *Susan,* mais la situation est ambiguë et il est difficile pour une machine de répondre à cette question qui nous paraît pourtant très simple à traiter lorsque nous maîtrisons l'anglais. La personne ou la machine qui passe le test doit effectivement être en mesure d'identifier correctement le référent du pronom *SHE* qui pourrait

[39] Le problème de l'ambiguïté référentielle est souvent lié à celui de « coréférence » qui prend forme lorsque deux expressions réfèrent à la même entité. Voir : Manning et al., 2020 : 30050.

être utilisé ici pour référer autant à *Joan* qu'à *Susan*, créant ainsi une ambiguïté (Levesque et al., 2012 : 554, 555).

Nous souhaitons porter l'attention de la lectrice ou du lecteur vers le fait que, pour cet exemple, répondre correctement à la question est plus difficile que dans la plupart des exemples que nous avons donnés précédemment et que la capacité à identifier correctement les fonctions ou les rôles syntaxiques des constituants de l'expression ne semble pas suffisant pour s'assurer d'une réponse adéquate à tout coup par un système d'IA. Effectivement, cette mise en situation montre bien qu'il serait utile que le système ait aussi accès à certaines informations liées à la signification des verbes *to thank* et *to help* (en sémantique, on dirait que l'IA devrait avoir accès aux significations de ces items lexicaux)[40]. En effet, dans notre exemple, puisqu'il est difficile de trouver à quel nom propre *SHE* réfère, si la machine avait accès à de l'information sur les concepts de remerciements et d'aide (comme un locuteur compétent de la langue anglaise), sa tâche serait plus simple. On peut penser à des connaissances sous la forme de propositions comme « ce sont les personnes qui aident les autres qui sont remerciées ». En ce sens, l'interaction entre des données sémantiques et syntaxiques peut améliorer la simulation de compréhension[41].

Cependant, notons tout de même qu'en répondant correctement, une machine pourrait être perçue comme comprenant la situation, qu'elle ait accès ou non à des informations sur la signification de « *to thank* » et de « *to help* ». Répondre que c'est de l'aide de Susan dont Joan a profité manifeste déjà une compréhension de la relation entre Susan et Joan qui est exprimée dans

---

[40] C'est d'ailleurs l'accès à ce genre d'informations qui est nécessaire lorsque l'IA est confrontée, par exemple, à deux phrases syntaxiquement distinctes, mais qui auraient le même sens/la même signification (Katz, 1980 : 18).

[41] Cette idée est d'ailleurs abordée par Levesque et les auteurs du *Challenge* qui proposent que pour répondre à ce genre question, un agent doive être en mesure de déterminer ce qui se passe (« *figure out what is going on* ») dans la situation particulière exprimée dans la proposition (Levesque et al., 2012 : 554).

la phrase, que la nature de cette relation soit réellement connue et comprise ou non. Autrement, une machine pourrait simplement, à la rencontre d'une ambiguïté de ce genre, être entrainée à demander de clarifier la relation exprimée dans la phrase et sa direction. C'est ce que nous faisons, nous-mêmes, lorsque nous ne sommes pas en mesure d'extraire ces informations.

L'analyse d'une autre question tirée du défi *Winograd* va dans le même sens :

*« The trophy doesn't fit in the brown suitcase because it's too small.*

*What is too small?*

*Answer 0 : the trophy*

*Answer 1 : the suitcase » (*Levesque et al., 2012 : 554-555)

On le comprend rapidement, c'est la valise brune qui est trop petite pour le trophée. Nous le savons intuitivement parce que la situation exprime une incapacité à faire entrer une chose dans une autre et que cela est causé par leurs tailles respectives (Levesque et al., 2012 : 554-555). Malgré sa forme qui présente une ambiguïté référentielle (l'ambiguïté est provoquée par la structure de la phrase, voir : Yu et al., 2022 : 45), cette question n'est pas, pour nous, trop ambiguë, car on sait que pour qu'un trophée (ou tout autre objet) entre dans une valise, il doit être plus petit que la valise en question. Les auteurs de l'article présentent cette capacité comme l'habileté à raisonner à partir de connaissances d'arrière-plan (*background knowledge*) *(*Levesque et al., 2012 : 554-555). Encore plus précisément Lévesque et al. proposent que l'agent doive être capable de raisonner à partir de connaissances « spatiales » (Levesque et al., 2012 : 554). On pourrait aussi dire qu'il est effectivement utile de savoir que les valises sont faites pour transporter des choses et non les trophées ou de connaître des informations à propos du verbe « *to fit* » (Forbes, Choi, 2017 : 1). Un humain connaît ces informations, tandis qu'une machine, si on ne lui apprend pas en la programmant ou en l'entrainant sur des données représentatives de cette réalité, ne peut clairement déterminer ce qui se passe dans l'expression (Levesque et al., 2012 : 554). Il y a en effet des

moments où l'information tirée d'une analyse syntaxique est insuffisante pour interpréter des phrases ambiguës et donc des situations où l'IA capable d'analyser la syntaxe a besoin de plus : comme l'accès à des significations ou de l'information additionnelle sur le contexte qui permettraient d'améliorer l'interprétation (Ferreira, Bailey, Ferraro, 2002 : 11). Suivant ces idées, si l'analyse syntaxique d'une phrase permet à une machine de découvrir des relations entre les termes qui la composent et même les directions de ces relations, elle semble parfois faire défaut si elle n'est pas réalisée à l'aide de certaines données à propos de la signification des mots traités, notamment lorsqu'elle rencontre des ambiguïtés.

### 2.4.2 Quelle importance pour la syntaxe dans une définition fonctionnelle de la compréhension ?

Enfin, pour conclure ce chapitre nous voulons aborder la question de savoir quelle importance la manipulation syntaxique devrait avoir dans une définition fonctionnelle de la compréhension utile pour rendre compte de l'attribution de cette capacité et de l'attribution d'états mentaux à des agents artificiels. Mentionnons d'abord une nuance importante ayant trait au fait que même sans une maîtrise des règles syntaxiques, une machine qui démontrerait une capacité à accéder à des données sémantiques pourrait aussi être considérée comme capable de comprendre.

D'abord, la nuance que nous souhaitons posée peut être illustrée grâce à un exemple fictif : si une machine répondait « vais-je bien » à la demande « comment allez-vous ? », probablement que la personne posant la question remettrait en doute le niveau de compréhension de la machine. En effet, la réponse de celle-ci, même si elle comporte les mêmes mots que la réponse adéquate « je vais bien. », vu l'ordre fautif de ceux-ci à l'intérieur de la phrase, pourrait être interprétée comme une question: « vais-je bien ? ». Dans une telle situation, probablement que la personne poserait sa question à nouveau pour être certaine qu'elle soit bien reçue et pour recevoir une

réponse adéquate. Néanmoins, il faut aussi remarquer que la personne qui pose la question et qui reçoit la réponse ambiguë de la part de la machine ne serait peut-être pas prête à affirmer que la machine n'a absolument rien compris de sa demande. En effet, si les mots sont agencés d'une façon étrange, on peut tout de même y voir les composantes nécessaires à une réponse adéquate à la question de départ. Cela pointe vers deux choses : d'abord, nous voyons que les autres dimensions du langage (sémantique et pragmatique) pèsent aussi dans la balance lorsque vient le temps de nous demander si nos demandes, par exemple, ont bien été comprises. On peut d'ailleurs penser aux théories philosophiques portant sur l'holisme sémantique qui tentent d'expliquer comment les significations des mots d'une langue peuvent être comprises comme interdépendantes et déterminées par nos croyances à l'égard des mots. Cela pourrait effectivement nous fournir une piste pour expliquer la perception d'un certain niveau de compréhension chez une machine qui serait en mesure de manifester une connaissance des relations entre les significations des mots, par exemple, en utilisant des mots adéquats dans des contextes précis[42]. Ensuite, comme nous ne serions pas prêts à soutenir qu'une personne nouvellement locutrice d'une langue étrangère qui agencerait mal les mots ne comprenne absolument rien à cette langue, nous pourrions considérer que la machine a quand même compris jusqu'à un certain point, et ce, même si elle fait défaut lorsqu'elle fournit sa réponse : lorsqu'elle ne suit pas les bonnes règles syntaxiques parce que ces règles sont mal programmées ou qu'elles ont été incorrectement apprises. Autrement dit, une réponse mal formée n'est pas nécessairement le résultat d'une question complètement mal interprétée et une réponse mal formée n'est pas non plus nécessairement complètement incomprise par son énonciateur. De ces nuances, retirons deux éléments importants : Il semble en effet exagéré

[42] L'holisme modéré propose une analyse d'une forme de « contextualisme » de ce type, où les significations sont déterminées en partie par nos croyances, mais aussi par les différents contextes d'utilisation de ceux-ci. Voir : Jackman, 1999.

de voir la compréhension linguistique comme un processus « binaire » que nous attribuons ou non à une personne ou une machine. Il semble y avoir des choses comme des niveaux de compréhension. Puis, retenons finalement que la capacité à manifester une maîtrise des règles syntaxiques ne semble pas la seule analyse possible de la compréhension qu'on attribue à l'IA. Ces idées seront développées dans les chapitres subséquents.

Maintenant, revenons sur le chemin que nous avons parcouru. Nous avons d'abord ouvert le chapitre en introduisant trois outils philosophiques : nous avons présenté l'expérience de pensée de la chambre chinoise, dans le but de montrer ce que la manipulation syntaxique n'est pas, pour ensuite nous tourner vers la question de la contribution de la syntaxe dans la signification d'expressions (principes de compositionnalité et de contexte). Enfin nous avons abordé les rapports entre les relations logiques, les relations syntaxiques et la sémantique en philosophie analytique contemporaine et en linguistique. Nous avons ensuite posé l'hypothèse voulant que la capacité d'analyse syntaxique dont sont capables certaines machines intelligentes leur permette de manifester une capacité à comprendre des relations entre des mots et groupes de mots qui expriment des relations réelles existantes entre les référents de ces mots dans le monde (relation spatiales, agentielles et intentions de communication). Nous avons ensuite présenté divers exemples concrets de cela. Enfin, nous avons conclu ce chapitre en ouvrant la discussion sur l'apport du contenu sémantique dans l'analyse syntaxique réalisée par l'IA, puis, plus largement, dans la manifestation de la capacité à comprendre une langue. L'objectif était, bien entendu, d'apporter des nuances à l'analyse de notre hypothèse de départ. Trois éléments importants ont été soulevés : il est plausible qu'il y ait des « niveaux de compréhension », la capacité à manifester une maîtrise des règles syntaxiques ne semble pas la seule analyse possible de la compréhension telle qu'on l'attribue à l'IA et, finalement, le traitement de certaines expressions ambiguës (comme dans le cas du problème de l'ambiguïté référentielle) est difficile à réaliser sans information à propos des

significations des mots. Cette dernière idée est immensément importante, car elle se base sur des exemples concrets de situations où la seule capacité à traiter, analyser et manipuler des structures syntaxiques se montre insuffisante pour être considérée comme de la compréhension linguistique et même pour manifester de la compréhension linguistique.

Finalement, ce chapitre nous a effectivement permis d'analyser une hypothèse pertinente et de formuler des outils pour présenter une première partie d'une définition fonctionnelle de la compréhension. Notre chapitre propose alors que la maîtrise syntaxique peut se montrer suffisante pour manifester un certain niveau de compréhension (la limite de ce niveau étant montrée par les problèmes de traitement des expressions ambiguës). Principalement, nous avons montré qu'elle permet de manifester une compréhension des relations exprimées dans le langage. Toutefois, nous avons aussi montré que sa maîtrise complète n'est pas nécessaire pour que l'on perçoive de la compréhension. C'est d'ailleurs pourquoi le prochain chapitre se concentrera sur le traitement sémantique des mots et des phrases par l'IA. Cela dit, puisque nous cherchons à savoir en quel sens la maîtrise syntaxique permet à une machine de manifester de la compréhension, il convient de remarquer que c'est principalement la compréhension de relations entre les mots qu'une machine manifeste lorsqu'elle se montre capable de traiter correctement la syntaxe d'une phrase. Comme première partie à notre définition fonctionnelle de la compréhension, nous voulons donc proposer ce qui suit : la compréhension implique de se montrer capable d'identifier des relations et leurs directions respectives exprimées entre les mots et groupes de mots contenus dans des expressions, puis de respecter ces relations et leurs directions lors de la formulation de sorties. Ces relations peuvent être de différents types : relations spatiales, relations temporelles, relations agentielles, intentions de communication, etc. (les relations identifiées dans les expressions traitées peuvent référer à des relations réelles entre les référents des mots).

Enfin, pour nous, interlocutrices et interlocuteurs humains, les mots et groupes de mots qu'une machine peut traiter trouvent leurs référents dans la réalité. En ce sens, lorsqu'une machine parle de désirs qu'elle dit entretenir ou lorsqu'elle énonce des croyances ou des connaissances à l'égard du monde, notamment à travers l'expression d'attitudes propositionnelles, il semble naturel de considérer qu'elle comprend les relations (et leurs directions) qui existent entre elle, ses désirs, ses croyances/connaissances et le monde. Mais si cette même machine n'est pas en mesure de respecter ces relations chaque fois qu'elle prend la parole ou si elle ne peut montrer qu'elle est capable d'identifier des relations lorsque nous lui parlons, nous pouvons être amenés à la considérer comme incapable de comprendre ce que nous lui disons ou ce qu'elle dit elle-même. Ainsi, l'analyse de la syntaxe est l'outil par excellence auquel l'IA a accès pour extraire de l'information relationnelle et respecter ces relations et leurs directions.

# CHAPITRE 3 — LES REPRÉSENTATIONS SÉMANTIQUES ET LA COMPRÉHENSION

## 3.1 Introduction au chapitre

L'objectif de ce chapitre est de présenter, maintenant, les éléments théoriques de la philosophie du langage et de l'esprit qui sont fondamentaux pour étudier les capacités que le traitement de données sémantiques procure aux systèmes d'IA. Ainsi, dans la première partie du chapitre, nous proposerons une analyse des façons par lesquelles une machine dotée d'IA peut manifester la capacité à comprendre les significations des mots et expressions d'une langue en les regroupant sous deux principales catégories : des expressions qui parlent de la réalité externe et des expressions qui parlent d'états internes. Partant, nous soutiendrons que la compréhension de ces expressions présuppose l'attribution à des agents conversationnels d'une capacité à se représenter adéquatement les significations de ces expressions, puis nous présenterons le rapport entre ces représentations et la réalité à partir de la notion philosophique de « direction d'ajustement ».

Ensuite, nous en viendrons à présenter la principale hypothèse qui fera l'objet de ce chapitre, soit celle qui veut que les machines dotées d'IA puissent manifester une capacité à comprendre le langage parce qu'elles peuvent montrer qu'elles possèdent des représentations sémantiques des mots qui sont fonctionnellement équivalentes à nos représentations mentales de

leurs significations. Une fois l'hypothèse lancée, nous présenterons trois difficultés qui se dressent face à l'idée que des machines dotées d'IA puissent effectivement entretenir des représentations. Ce sont ces difficultés qui serviront à placer certaines balises nous permettant de clarifier ce que nous entendons par « représentations » en restreignant le concept à celui de *représentations sémantiques*.

Enfin, cette présentation nous mènera à l'approfondissement de la question de savoir en quoi peuvent consister les contenus des représentations sémantiques dont l'IA peut manifester de la compréhension. Nous commencerons par montrer qu'elles peuvent représenter des contenus généraux ou plus particuliers comme des référents, des descriptions et des relations sémantiques. Cette analyse nous mènera ensuite à clarifier ce qu'elles peuvent contenir en précisant et en soutenant la thèse principale de ce chapitre. Nous établirons effectivement, dans la dernière partie du chapitre, un rapprochement entre la notion de représentation sémantique et celle de concept. C'est grâce à ce rapprochement que nous pourrons soutenir que les représentations sémantiques que certaines machines peuvent entretenir et/ou montrer qu'elles entretiennent sont fonctionnellement équivalentes à nos représentations mentales : nous proposerons que ces représentations sémantiques puissent être équivalentes aux concepts que nous possédons.

### 3.2 Les systèmes d'IA peuvent manifester une capacité à comprendre des représentations de la réalité externe et l'expression d'états internes

La première catégorie d'expressions pour lesquelles des machines peuvent manifester de la compréhension est celle qui englobe les expressions qui expriment des contenus représentationnels : des contenus qui portent sur la réalité externe, soit en référant à des objets et des composantes de ceux-ci ou en attribuant des propriétés à ces objets et composantes. La

deuxième catégorie est celle qui comporte les expressions d'états internes de personnes ou d'autres entités. On peut penser aux phrases exprimant des émotions, des états mentaux, etc.

D'entrée de jeu, lorsque nous communiquons avec des machines comme des robots sociaux, des *agents conversationnels* ou des assistants vocaux et que nous leur demandons des informations à propos de certaines choses comme l'itinéraire le plus rapide pour se rendre quelque part ou les caractéristiques propres à une race de chien, par exemple, il semble que nous présupposions qu'ils peuvent comprendre notre demande, mais aussi qu'ils « connaissent » ces choses suffisamment bien pour y répondre. Mais en quoi consistent ces choses que nous croyons qu'ils peuvent « connaitre » ? Des définitions ? Une machine inconsciente peut-elle vraiment *connaître* ? Il semble plus juste de dire qu'une telle machine peut communiquer des représentations de la réalité et c'est l'avenue que nous prendrons. Ces agents artificiels peuvent en effet manifester une capacité à communiquer des représentations de la réalité, principalement parce qu'ils peuvent nous fournir des informations qui la décrivent et parce qu'ils peuvent référer à la réalité par les mêmes mots que nous employons (comme des cartes routières et des itinéraires représentent l'état de routes sur lesquelles nous pouvons circuler, les noms des races de chiens, par exemple, représentent les propriétés de chiens que nous pouvons rencontrer, etc.)

En guise d'exemple supplémentaire, on peut penser à *ChatGPT,* un robot conversationnel fonctionnant grâce à un modèle de langage de grande envergure. Le système est présenté comme un agent conversationnel pouvant répondre à des questions textuelles complexes. En réalité, *ChatGPT* fonctionne comme un prédicteur de texte et réalise des calculs probabilistes permettant de déterminer quels mots sont les plus probables à venir les uns à la suite des autres et ceux qui sont les plus souvent utilisés pour répondre à certaines questions (Bender et al., 2021 : 610-612). Sa particularité, c'est qu'il a été et est toujours entrainé sur une quantité énorme de données. En effet, l'accès à des textes de différentes formes, traitant d'innombrables sujets, permet à l'IA d'être

en mesure de composer des textes complexes, qui paraissent nuancés, réfléchis et même érudits. Ainsi, même si *ChatGPT* paraît comprendre naturellement et complètement nos questions et ses propres réponses comme le pourrait un être humain, le système ne fait que reproduire des modèles de conversations en accédant à de l'information contenue sur internet. Pourtant, *ChatGPT* se présente comme une machine ayant une certaine capacité explicative, ce pour quoi il n'est d'ailleurs pas rare de rencontrer des étudiantes et des étudiants qui l'ont déjà utilisé pour rédiger une dissertation à leur place ou pour réviser des notions difficiles avant un examen. Des comportements observables de ce genre sont des manifestations directes de notre tendance à placer une confiance aveugle en la capacité de ce système à adéquatement représenter la réalité (notamment en disant vrai).

Pour continuer, en plus de communiquer des représentations du monde, les agents artificiels que nous connaissons peuvent se montrer capables de comprendre certaines formes d'expressions d'états mentaux et en exprimer eux-mêmes. Des expériences empiriques présentent effectivement des cas où des êtres humains, en interagissant avec des systèmes d'IA, rapportent avoir l'impression que ces systèmes peuvent communiquer leurs propres désirs et croyances ou semblent pouvoir comprendre l'expression d'états mentaux par des êtres humains. Shank et al., par exemple, rapportent une situation où un robot conversationnel agit de façon compréhensive et s'excuse après l'expression, par un être humain, de frustration et de désarroi (Shank et al., 2019 : 263, 264). Ce phénomène est souvent appelé *mind perceiving* et semble être provoqué, entre autres, par l'utilisation de certaines techniques ou stratégies communicatives comme la politesse, l'excuse, l'humour, ou bien par l'expression de tristesse ou de joie par la machine (Shank et al., 2019 : 263, 264).

Maintenant, ce que nous souhaitons ici montrer, c'est que la capacité à comprendre du contenu représentationnel et la capacité à comprendre des expressions d'états internes/mentaux

dépendent souvent l'une de l'autre. Les états mentaux sont normalement exprimés sous la forme d'attitudes propositionnelles et les propositions sur lesquelles portent ces attitudes peuvent avoir des représentations comme contenus. Je peux exprimer un état de tristesse en disant quelque chose comme : « Je suis triste q*ue mon chat soit mort* ». Dans un tel cas, j'exprime une représentation de la réalité au sens où je décris un fait, une situation réelle dans le monde, qui s'incarne sous la forme d'une représentation complexe de mon état de tristesse par rapport à l'événement « la mort du chat qui m'appartenait », analysable par une décomposition des représentations plus simples qui la forment : celles de « tristesse », du concept d'appartenance et la représentation « chat ». Ainsi, la capacité à comprendre des phrases exprimant des états internes paraît liée à cette capacité de certaines machines à communiquer et « interpréter » des représentations de la réalité, puis à réagir de certaines façons qui sont adéquates par rapport à celles-ci (en s'excusant, en utilisant des phrases qui connotent de la tristesse, etc.).

Suivant ces idées, nous nous concentrerons effectivement sur le caractère représentationnel de la signification dans le reste de ce chapitre. Nous nous demanderons comment une machine peut manifester de la compréhension de mots et d'expressions qui ont normalement comme contenus des représentations de la réalité. Cette question est fondamentale, d'abord parce qu'elle permettra de clarifier le problème laissé en suspens au chapitre précédent à propos du fait qu'une machine dotée d'IA doive parfois avoir accès à certaines données sémantiques pour manifester la capacité à comprendre les relations et les situations décrites par certaines phrases ambiguës. Ensuite, son traitement est nécessaire pour mieux saisir comment rendre compte des phénomènes décrits dans le paragraphe précédent, où l'on présente deux capacités linguistiques (la capacité à comprendre des expressions à propos de la réalité externe et des expressions d'états internes) que peuvent manifester des agents artificiels. Avant de nous y attaquer, nous souhaitons présenter une dernière notion philosophique utile pour bien saisir l'importance de la notion de représentation.

Le contenu représentationnel du langage est effectivement constamment exprimé par nos prises de parole. John Searle et Daniel Vanderveken, par exemple, dans *Foundations of illocutionary logic* (1985), formalisent la logique des actes de langage en présentant, entre autres, la notion philosophique de « direction d'ajustement ». Cette notion montre bien comment le langage représente le monde, et ce, en raison de l'existence de relations d'ajustement entre l'un et l'autre. Selon les auteurs, les actes de paroles que nous réalisons suivent l'une des deux directions d'ajustement suivantes ou bien les deux en même temps : soit les mots s'ajustent au monde, soit le monde s'ajuste aux mots (Searle, Vanderveken, 1985 : 92). Nous pouvons comprendre ce que les auteurs veulent ainsi dire de la manière suivante : le contenu de nos prises de parole peut servir à décrire le monde tel qu'il est (à le représenter à travers le langage, grâce à *l'assertion*, par exemple : direction d'ajustement des mots au monde) ou peut servir à le représenter tel qu'il serait s'il correspondait à une certaine représentation que nous avons de ce qu'il deviendra une fois l'acte de parole réalisé (comme dans le cas de *la réalisation d'une promesse*, par exemple : direction d'ajustement du monde aux mots) (Searle, Vanderveken, 1985 : 92-96). Si j'ai l'intention de tenir une promesse que j'ai faite, la réalisation de la promesse est supposée mener à la modification du monde, afin qu'il corresponde au contenu de l'acte de parole : une fois ma promesse tenue, le monde correspondra à la représentation que j'avais en tête lorsque j'ai promis [43]. Enfin, le langage peut servir à agir sur le monde pour le rendre d'une certaine façon (comme c'est le cas lors de la réalisation d'un acte déclaratif : on peut penser à un célébrant qui déclare deux personnes

[43] Il faut toutefois noter une différence dans la catégorisation de la promesse chez Searle, notamment entre ce qui est présenté dans ses plus vieux ouvrages et dans ses plus récents. Les ouvrages plus récents de Searle et ses commentatrices et commentateurs tendent vers l'idée que l'acte de promettre peut aussi être compris comme suivant la double direction d'ajustement, au sens où l'énonciation de la promesse crée un nouveau fait social: la promesse elle-même et l'obligation sociale qu'elle implique. En ce sens, la promesse suivrait aussi la direction d'ajustement des mots au monde. Je me suis obligé à tenir ma promesse et les autres me tiennent pour obligé de la tenir. (Hindriks, 2013 : 382; Searle, 2010 : 15, 16)

« mariées » ou à une personne qui « baptise » un bateau). Dans ces cas, la direction d'ajustement est double : la réalisation de l'acte de langage modifie le monde (la réalité sociale, plus précisément) (ajustement du monde aux mots) en le représentant comme étant dorénavant d'une certaine façon (ajustement des mots au monde). Ces trois exemples portent sur des actes de paroles paradigmatiques, mais montrent bien que le langage peut servir à décrire le monde, le monde tel qu'il pourrait être ou tel que l'on voudrait qu'il soit, donnant encore une fois à cette capacité une importance fondamentale pour l'attribution de la compréhension à des machines.

### 3.3 Deux obstacles aux représentations artificielles et pourquoi il doit quand même y avoir des représentations de la réalité

Qu'entendons-nous par « représentations de la réalité » ? Ludwig Wittgenstein parlait lui-même des propositions en tant qu'images (*Bilder*) de la réalité, faisant en sorte que l'ensemble des propositions vraies à propos la réalité, si nous les joignions ensemble, formeraient comme une grande image de la réalité (Wittgenstein, 2010 : 2.1-3.01). Bien que sémantique, cette conception des représentations rappelle la notion de représentation mentale abordée dans le premier chapitre de ce mémoire. Nous souhaitions, à ce moment-là, introduire l'étrange phénomène des machines intelligentes qui expriment des états mentaux comme la peur, la joie, des croyances, etc. Le concept de représentation mentale nous a ainsi permis d'expliquer comment les états mentaux peuvent être à propos d'autre chose qu'eux-mêmes, notamment parce que les propositions qui font l'objet d'attitudes peuvent avoir, comme contenus, des représentations de la réalité.

Les représentations de la réalité que nous construisons en tant qu'êtres humains sont évidemment variées et complexes. Elles prennent, par exemple, les formes de propositions, d'images ou de concepts. De plus, elles peuvent être reliées de différentes façons entre elles, elles peuvent être plus simples ou plus complexes et les représentations complexes peuvent être formées

par l'agencement d'autres représentations plus élémentaires. On peut aussi comprendre comment les mots réfèrent aux constituants de la réalité à travers des descriptions de celle-ci.

Gottlob Frege, quant à lui, soutenait une approche descriptiviste de la signification. Dans le cadre conceptuel de Frege, les significations des expressions ne sont pas comprises comme des représentations mentales subjectives, mais plutôt comme des « sens » : des modes de donations de référents réels (Linsky, Pelletier, 2005 : 199, 200). Ce sont alors des descriptions des choses de la réalité que nous comprenons lorsque nous comprenons les sens d'expressions qui les désignent. Je peux, par exemple, connaître une personne parce que je l'ai rencontrée récemment et donc me la *représenter* comme tel lorsque j'entends son nom, mais je peux aussi comprendre son nom comme une description qui la désigne. Je peux connaître la personne en tant que « la personne qui est ceci et cela ». À ce moment-là, lorsque je comprends le terme qui est utilisé pour la nommer, c'est cette description que j'ai en tête, plutôt qu'une représentation subjective que je peux avoir de la personne en question.

En tant qu'êtres humains, lorsque nous communiquons avec les autres nous utilisons ces représentations mentales et les descriptions que nous connaissons de la réalité pour faire sens des mots et expressions utilisés pour parler des choses qui la constituent. Pour nous intéresser à la question de savoir comment une machine inconsciente peut manifester la compréhension de significations, nous devons alors nous demander comment une machine peut se représenter la réalité même sans y avoir directement accès. C'est pour cette raison que nous nous concentrerons sur la capacité des systèmes d'IA à construire des représentations sémantiques (des représentations des significations des mots et des expressions) qui peuvent se montrer équivalentes à des représentations mentales et des descriptions du monde que nous avons et utilisons. De là, nous nous demanderons plus particulièrement comment des représentations sémantiques (des représentations artificielles des significations des mots et expressions) peuvent se montrer équivalentes à nos

concepts, de sorte que les machines avec lesquelles nous interagissons paraissent capables de comprendre les concepts exprimés par le langage. Dans les prochaines lignes, nous poserons et analyserons deux difficultés qui se dressent devant cette idée. Ces problèmes permettront de mettre en lumière un problème plus général et historiquement important en philosophie de l'esprit et du langage : le problème du rapport entre le monde et les représentations que nous formons de celui-ci. Ultimement, cette critique nous permettra de clarifier ce que nous entendons par représentations sémantiques et de montrer pourquoi elles forment le type de représentations le plus pertinent pour traiter la problématique de la compréhension et de l'attribution d'états mentaux aux machines.

### 3.3.1 Des représentations inconscientes

Pour débuter, nous souhaitons aborder la forte opposition du philosophe John Searle à l'idée qu'une machine puisse avoir et puisse communiquer des représentations. La position de Searle, qu'il a développé dans le cadre de sa célèbre expérience de pensée de la chambre chinoise, se présente ainsi : « *[…] the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics.* » (Searle, 1980: 422) Pour Searle, l'absence de conscience chez l'IA entraine systématiquement l'incapacité de l'IA à utiliser des symboles de la même manière que les humains, c'est-à-dire à utiliser des symboles qui veulent dire quelque chose, des symboles qui représentent certaines choses. Toutefois, même si on tombait d'accord avec l'idée que ces symboles ne signifient et ne représentent rien pour une machine, pour nous, qui agissons en tant qu'interprètes des sorties de l'IA, ils signifient tout de même certaines choses (Müller, 2007 : 11). Une machine dotée d'IA peut effectivement utiliser le langage pour interagir avec nous dans un langage signifiant qui permet de véhiculer du contenu représentationnel. Comme l'état d'un thermomètre représente pour nous la

température environnante, alors qu'il ne représente rien pour le thermomètre lui-même, les énoncés d'une machine dotée d'IA peuvent représenter certains états de choses pour nous, même si ce n'est pas le cas pour la machine (Müller, 2007 : 3.1, 3.2). Quand un robot conversationnel nous indique que notre pizza vient de quitter le restaurant, dans le véhicule d'un livreur qui devrait arriver d'ici 15 minutes, il est raisonnable de penser que le robot parle effectivement de la pizza que nous avons en tête et qui nous fait saliver, comme il est raisonnable de croire que « 15 minutes » réfèrent à la même durée dont nous faisons l'expérience au quotidien. Il est aussi correct d'avoir en tête une voiture de livraison lorsque nous entendons le robot énoncer le mot « véhicule ». En croyant le contraire, il semble que l'on mélange deux problèmes. Si on pense que les symboles exprimés par une machine ne signifient absolument rien parce qu'elle-même ne comprend pas leurs significations de la même manière que nous le pouvons, c'est le fait qu'une machine soit inconsciente qui nous préoccupe davantage que le fait de savoir si une machine peut faire preuve d'une forme de compréhension linguistique et communiquer des représentations du monde (Müller, 2007 : 5.2).

Cependant, il est juste de remarquer qu'il paraît intuitivement problématique que des signes puissent représenter quelque chose, même s'ils ne sont reliés à cette chose pour tous les agents qui peuvent l'interpréter. Comme l'indique Müller dans son article, ce qui fait que l'état du thermomètre représente la température ambiante, c'est le fait que son état est causalement lié à l'état de la température ambiante (Müller, 2007). C'est sa fonction de représenter la température. Pour Müller, c'est une caractéristique fondamentale de la représentation : « *In order for X to represent Y, X should stand in a particular relation to Y* […]. » (Müller, 2007 : 3.2) Cette difficulté montre en effet qu'un symbole, pour représenter quelque chose, doit y être relié adéquatement. L'approche que nous présenterons en tiendra compte, mais proposera qu'une relation minimale soit suffisante (contrairement à Müller qui met l'emphase sur une relation *causale* entre le signe et

la chose représentée). Nous proposerons que cette relation se manifeste par une capacité de la machine qui utilise le symbole à décrire ou à expliquer adéquatement ce à quoi le symbole renvoie.

### 3.3.2 Le problème de l'ancrage des symboles

La deuxième difficulté découle directement de la première et se nomme le problème de l'ancrage des symboles (*symbol grounding problem*), formulé par le chercheur en sciences cognitives Stevan Harnad. Ce dernier soutient que l'idée voulant que la cognition consiste en la réalisation d'opérations sur des symboles est problématique, notamment parce que cela ne nous permet pas de comprendre comment les significations des mots et expressions que nous interprétons trouvent leur ancrage dans la réalité[44] (Harnad, 1990 : 2.2). Dans son article, Harnad s'intéresse plus précisément aux systèmes informatiques capables d'identifier certains objets, notamment par des processus de discrimination, de comparaison et finalement de catégorisation (Harnad, 1990 : 3). Son article se concentre alors sur un système à la fois symbolique et non symbolique qui permet l'association de représentations iconiques[45] (des formes, par exemple) et de représentations catégorielles innées/programmées (des mots, par exemple), dans le but de réaliser les tâches nommées précédemment, mais aussi dans le but de former des représentations plus complexes à partir de représentations simples (Harnad, 1990 : 3). Pour former la représentation complexe « zèbre », par exemple, un système de ce genre pourrait avoir les représentations catégorielles simples « cheval » et « rayures » ancrées dans des représentations iconiques (les

[44] On peut se demander, par exemple, ce qui fait que le symbole « cheval » est lié à la chose réelle que nous nommons « cheval ».

[45] Les icônes sont des représentations en vertu du fait qu'elles ressemblent physiquement à l'objet qu'elles représentent. Dans la théorie des représentations de Charles Sanders Peirce, les représentations iconiques sont l'une des relations d'ancrages possibles qu'ont les représentations comme référer, exprimer un sens, ou représenter la réalité (Von Eckardt, 2012 : 31).

formes physiques d'un cheval et d'une rayure) (Harnad, 1990 : 3.3). De là, la représentation complexe « zèbre » pourrait être formée par la suite de symboles « cheval et rayures », la première héritant ainsi des ancrages iconiques des représentations plus simples qui la composent (Harnad, 1990 : 3.3).

Certes, le problème de l'ancrage est important parce qu'il met en lumière l'interaction entre les représentations symboliques et non symboliques nécessaire pour la compréhension « complète » ou « naturelle » du langage. Cependant, il faut voir qu'une machine peut tout de même manifester de la compréhension d'un nom sans avoir recours à des représentations iconiques comme celles que peut traiter le système étudié par Harnad. Plusieurs assistants virtuels comme Alexa et Siri n'ont pas ce genre de représentations et sont donc complètement « aveugles » par rapport au monde physique. Néanmoins, ces systèmes peuvent accéder à des définitions des choses et donc à des descriptions pour en parler. Dans de tels cas, l'ancrage dans la réalité n'est pas nécessaire pour la compréhension d'un signe. Ainsi, la question de savoir comment créer des machines qui raisonnent à partir d'autre chose que seulement des mots et des calculs, mais bien avec des représentations de la réalité physique est pertinente, mais vue la nature des agents artificiels qui nous intéressent, c'est surtout ce qu'Harnad nous dit sur l'interaction entre les différentes représentations entretenues par un système qui rend le problème de l'ancrage particulièrement pertinent. Dans son article, les représentations du système entretiennent entre-elles des relations. Ainsi, le papier de Harnad soulève une difficulté importante : les représentations d'un agent artificiel, comme nos représentations mentales, doivent interagir entre elles pour bien représenter la réalité dans toute sa complexité (c'est d'ailleurs le cas lorsqu'un système utilise une définition formée de mots pour en « interpréter » un autre). Retenons donc que les problèmes que nous avons présentés pointent vers deux caractéristiques que doit avoir une représentation : elle

doit être reliée à la chose qu'elle représente et elle doit entretenir des relations adéquates avec d'autres représentations.

### 3.3.3 Pourquoi le concept de représentation sémantique doit-il être préservé ?

Pourquoi, alors, parler de représentations qu'ont des machines dotées d'IA ? Précisément parce que les difficultés soulevées dans les lignes précédentes pointent vers l'idée qu'il doit y avoir, lorsque l'on interprète un énoncé ou un mot, quelque chose comme une bonne interprétation. Il faut en effet savoir construire la représentation adéquate de la signification de l'énoncé ou du mot que l'on interprète. Lorsque nous communiquons avec les autres, nous nous attendons effectivement à être compris : nous attendons, de la part de nos interlocuteurs, une capacité à se représenter adéquatement la signification de ce que nous disons. On peut expliquer cela, notamment en faisant appel aux concepts d'intersubjectivité et de lieu commun (*common ground*)[46]. La communication présuppose que nous connaissions des significations partagées qui nous permettent de nous comprendre. L'intersubjectivité, telle qu'elle est présentée par Edmund Husserl, est une façon d'expliquer le fait que nous partageons une certaine relation au monde avec les autres (Hermberg, 2006 : 65). En effet, les sujets/égos participent à un monde intersubjectif, au sens où ils partagent un monde de significations communes (Hermberg, 2006 : 66). Le « lieu commun » quant à lui, réfère à un ensemble de présuppositions qu'incluent des prises de paroles (Stalnaker, 2002 : 701). Il contient donc des informations en arrière-plan de nos conversations quotidiennes. Robert Stalnaker, par exemple, propose qu'en contexte de communication, nous

[46] Le terme « intersubjectivité » trouve ses origines dans la tradition phénoménologique, alors que le concept de « lieu commun » provient plutôt de la philosophie du langage et de la linguistique anglo-saxonnes. Le développement de la notion de *common ground* est principalement attribué à Charles Sanders Peirce, H. Paul Grice et Robert Stalnaker.

présupposions que certaines croyances sont communes, qu'elles sont partagées avec notre interlocuteur, influençant ainsi la conversation en créant un lieu commun se modifiant selon le fait que nous présupposons que plus ou moins d'informations sont communément crues (Stalnaker, 2002). Ce qui est intéressant dans l'analyse des concepts d'intersubjectivité et de lieu commun, c'est d'abord la justification qu'ils donnent de la nécessité de connaître des significations partagées pour pouvoir réaliser le processus mental de la compréhension[47]. Dans un même ordre d'idées, une autre approche importante à propos de ce qui constitue la signification d'un mot est celle qui utilise le concept de « règle ». Peter F. Strawson, par exemple, propose que les significations des mots qui « réfèrent » soient des instructions générales d'usages, des conventions, des habitudes qui gouvernent l'usage correct de celles-ci (Strawson, 1950 : 327, 328). Connaissant des règles d'usages, une personne qui comprend le mot « tu », par exemple, saurait qu'il est normalement utilisé pour référer à une personne (Strawson, 1950 : 327).

Suivant ces idées, il est aisé de voir que la communication dépend toujours d'un ensemble de présuppositions à propos des significations des mots et énoncés d'une langue, mais aussi à propos de la capacité des autres à les connaître et les comprendre. Nous présupposons aussi que les autres sont sensibles aux influences des contextes communicationnels lors de leurs interprétations et, finalement, qu'ils peuvent se représenter comment les mots et énoncés doivent être utilisés. Notons que la communication avec des machines semble aussi exiger que certaines connaissances

---

[47] Si nous nous intéressons plus en détail à l'approche du lieu commun de Herbert H. Clark, par exemple, nous remarquons que les significations des mots et des phrases qui forment les énoncés d'une conversation constituent en partie le lieu commun, ce qui implique que lorsque nous communiquons avec quelqu'un, nous présupposons effectivement que cette personne connaît certaines significations. (Clark 1996 : 53). Sa conception du « lieu commun » est divisée en une représentation du discours d'une conversation sous deux aspects : la représentation textuelle et la représentation situationnelle. Les significations conventionnelles des mots et expressions utilisés se rangent du côté textuel de la représentation du lieu commun, alors que les référents précis (les objets réels) des mots sont représentés du côté de l'aspect situationnel du lieu commun (Clark, 1996 : 52-54). Dans la philosophie du reconnu penseur de la conversation Herbert Paul Grice, ces significations communes sont plutôt catégorisées sous les concepts de *timeless meaning* en opposition à une signification déterminée par l'occasion d'usage (*occasion meaning)* (Grice, 1968).

(au moins à propos de certaines significations) soient partagées entre les agents (humains comme artificiels) pour qu'ils puissent communiquer, et ce, même si les rapports qui existent entre les agents artificiels et le monde ne sont pas de la même nature que ceux que nous entretenons, en tant qu'être humain, avec ce dernier. C'est pourquoi nous souhaitons soutenir la thèse voulant qu'une machine dotée d'IA doive pouvoir entretenir un type précis de représentations pour manifester de la compréhension linguistique : des représentations sémantiques (et non des représentations mentales) adéquates et fonctionnellement équivalentes aux représentations mentales du monde et aux descriptions de celui-ci que nous utilisons pour comprendre le langage. Dans les prochaines lignes, nous montrerons donc comment différents systèmes d'IA manifestent de la compréhension en se montrant capables de se représenter les significations d'expressions de façons équivalentes à nous et de les utiliser pour faire les mêmes choses que nous.

## 3.4 Le contenu des représentations sémantiques

### 3.4.1 Les références

Pour comprendre en quoi peuvent consister des représentations sémantiques, nous nous intéresserons d'abord à ce que la philosophie du langage contemporaine dit sur le contenu des représentations qui peuvent être exprimées par le langage. D'entrée de jeu, l'une des principales raisons pour lesquelles nous comprenons ce que disent des agents artificiels, c'est que les significations des expressions qu'ils utilisent peuvent inclure des références à des objets particuliers que nous connaissons[48]. Lorsqu'un agent réfère à un objet, si nous le connaissons, nous pouvons normalement comprendre l'expression utilisée pour y référer. Nous avons, par exemple, des connaissances à propos d'objets particuliers et d'entités particulières (comme un chat auquel nous

[48] « Objets » inclut, dans ce contexte, autant des objets physiques que des entités ou des personnes vivantes.

pouvons référer en le nommant « Pistache » ou par l'expression « mon chat », par exemple). Pour bien comprendre la notion de référence, il est utile de revenir à la distinction classique entre sens et référence/dénotation de Gottlob Frege. Selon Frege, une expression possède à la fois un sens et une dénotation qui sont distincts l'un de l'autre. La dénotation est ce que l'expression désigne dans le monde, alors que le sens de l'expression, quant à lui, est le mode de donation de la dénotation : la façon dont nous comprenons la dénotation (Frege, 1993 : 24). Les énoncés exprimant une identité ($a$=$b$) permettent d'exemplifier cela : « ($a$) Hesperus (=) est ($b$) Phosphorus ». Cet énoncé présuppose qu'il y a deux façons de parler et de comprendre une même chose : les expressions « Hesperus » et « Phosphorus » expriment deux sens distincts, soit respectivement « l'étoile du soir » et « l'étoile du matin », mais ont une seule et même dénotation (Frege, 1993 : 22). Il faut donc retenir que les sens des expressions nous permettent de comprendre comment se présentent leurs références : comprendre le sens, c'est donc être dirigé d'une façon (parmi potentiellement d'autres) vers la référence. De là, la personne qui sait que Phosphorus, en comprenant le sens « l'étoile du matin », réfère à Vénus ne comprend pas « Vénus » de la même manière que celle qui sait plutôt que Hesperus est « l'étoile du soir » et réfère à Vénus. Les deux personnes comprennent deux sens différents pour une même dénotation. Ainsi, la distinction entre sens et référence suggère que différentes façons de comprendre un mot peuvent nous guider vers une même chose dans le monde. En ce sens, il est possible que la représentation sémantique $R^1$ d'un mot que possèderait une machine puisse référer à une chose $C$ de façon équivalente à une représentation de la réalité $R^2$ (une représentation mentale ou une description, par exemple) que possèderait un être humain.

Cela étant dit, il est nécessaire de reconnaître que le traitement des expressions référentielles s'avère un défi d'envergure pour l'IA. Nous n'avons qu'à penser au problème de l'ancrage des symboles : comment une machine inconsciente peut-elle effectivement associer un nom à son référent réel dans le monde sans même ne jamais avoir conscience de ce monde ? C'est la même

chose si nous pensons aux expressions ambiguës qui contiennent une anaphore, nécessitant souvent l'accès à des connaissances d'arrière-plan à propos des contextes conversationnels dans lesquels elles sont employées pour bien identifier la référence des pronoms utilisés (Landgrebe, Smith, 2019 : 18-20). John Searle, dans *Speech acts* (1969), tente justement de formaliser les conditions que l'on doit nécessairement respecter pour référer à quelque chose :

1. *There must exist one and only one object to which the speaker's utterance of the expression applies […]*
2. *The hearer must be given sufficient means to identify the object from the speaker's utterance of the expression […]* (Searle, 1969 : 82, 83)

C'est la deuxième condition de Searle qui pique ici notre curiosité : elle nous dit que la personne qui entend l'énonciation doit recevoir suffisamment d'information pour pouvoir identifier la personne ou la chose à laquelle l'énonciateur réfère. En détaillant cette condition, Searle en vient d'ailleurs à soutenir que cette condition repose sur le principe d'identification, au sens où l'acte de référence est un processus continu durant lequel l'énonciateur peut être amené à clarifier la personne à laquelle il réfère en répondant à des questions plus précises à propos de son identité, jusqu'à ce que la personne qui entend soit en mesure d'identifier adéquatement la référence (Searle, 1969 : 85, 86). Il s'ensuit, pour Searle, que la deuxième condition se réalise ainsi : « *[…] though a speaker may satisfy it* [la seconde condition] *even if he does not utter an identifying description […] he commits himself to identifying one and only one object, he commits himself to providing one of these* [une description identifiante] *on demand.* » (Searle, 1969 : 86)

Ainsi, la capacité à référer reposerait sur la capacité à fournir des descriptions (si nécessaire selon le contexte et les connaissances de l'interlocuteur) permettant à l'interlocuteur d'identifier la

personne à qui l'on réfère[49]. Cette analyse est parlante pour faire sens de la capacité de certaines machines à référer à des choses et des personnes particulières, précisément parce que les agents conversationnels sont, en grande partie, capables de fournir des descriptions de plusieurs choses et de plusieurs personnes, notamment d'eux-mêmes. Les prochaines lignes montreront en quoi peuvent consister des descriptions permettant l'identification de référents, mais aussi de classes et de catégories pouvant inclure plusieurs objets.

### 3.4.2 Les descriptions et les relations sémantiques

En plus de la compréhension de mots et d'expressions qui réfèrent, une machine peut effectivement manifester une capacité à comprendre et à communiquer des représentations plus générales de la réalité comme des « catégories » ou des « classes »[50]. En effet, la capacité à former des représentations des choses sous la forme de descriptions nous permet de nous représenter la réalité, ses constituants et leurs propriétés et donc de parler adéquatement de ces choses. De plus, comme nous venons de le voir, en plus de nous permettre de parler de catégories ou de classes d'objets, ces représentations nous permettent de parler d'une chose ou d'une personne particulière à laquelle nous souhaitons référer en la décrivant. Il suffit de penser aux situations de notre vie dans lesquelles nous avons, par malchance, oublié momentanément le nom d'une personne et que

[49] En suivant une approche intentionnaliste de la référence comme celle de David Kaplan (1989), on pourrait certainement argumenter contre Searle en proposant que les conditions qu'il propose sont insuffisantes pour réaliser ce qu'on appelle l'acte de référer. L'approche de Kaplan veut que la personne qui réfère à une chose ou à un objet doit avoir l'intention d'y référer pour que l'acte soit réussi. L'intention dirige et détermine la référence (Kaplan, 1989 : 582). Néanmoins, comme indiqué plus haut, les machines dotées d'IA auxquelles nous nous intéressons sont souvent « aveugles » par rapport au monde. Elles ne peuvent « avoir en tête » une représentation mentale de la personne/de la chose précise dont elles parlent ou avoir quelque chose comme « l'intention d'y référer ». Au mieux, elles peuvent être le plus précises possible lorsqu'elles en parlent et cela passe souvent, dans le contexte d'une conversation ordinaire, par l'action de fournir une description précise de la chose ou de la personne à identifier en cas de mécompréhension.

[50] Nous connaissons des caractéristiques que partagent, par exemple, les chats.

nous avons alors seulement en tête des façons de la décrire pour que nos interlocuteurs comprennent de qui nous parlons.

La théorie philosophique « classique » des descriptions nous vient de Bertrand Russell et constitue un outil utile pour bien saisir comment une machine peut manifester une capacité à comprendre le langage en construisant et communiquant des représentations des significations des mots qu'elle traite sous la forme de descriptions. Plus précisément, par « descriptions », Russell entend des connaissances qui peuvent être de deux types. Il nous dit d'abord que les descriptions *définies* consistent en des phrases dénotantes qui visent un objet unique et particulier dans le monde (Hylton 2003 : 202; Russell, 2009b). Une description définie pourrait se dire ainsi : « *le livre de Bertrand Russell* qui se trouve sur la table devant moi », tandis qu'une description *indéfinie* pourrait se dire ainsi : « *quelques livres de Bertrand Russell* dans la bibliothèque de l'UQTR ». Les descriptions que nous connaissons peuvent être satisfaites (ou non) par des choses dont nous faisons ou ne faisons pas l'expérience directe[51]. « *The round square is round* », par exemple, est une description qu'aucun objet réel ne satisfait (Russell 1905 : 491). Sa signification est analysable de la façon suivante : il n'y a qu'une seule entité X qui est ronde et carrée et cette entité est ronde (Russell 1905 : 491). Ce qu'il faut en retenir, c'est que nous pouvons comprendre des descriptions des constituants de la réalité et donc les connaître, notamment en connaissant leurs propriétés, et ce, même si nous ne les connaissons pas par expérience directe (Russell, 2009b : chapitre V). Une machine dotée d'IA peut effectivement elle aussi montrer qu'elle connaît des descriptions d'objets ou de catégories d'objets de la réalité, même si elle ne les connaît pas par expérience. De là, il est

[51] Pour Russell, cela implique qu'il y a des choses que nous ne pouvons connaître que par description. Nous connaissons leurs propriétés, mais nous ne connaissons pas les choses elles-mêmes, dû à l'impossibilité d'avoir, avec elles, des accointances (Russell 1905 : 492, 493; Russell, 2009b).

plausible que notre tendance à percevoir de la compréhension chez des machines passe, entre autres, par la capacité de ces machines à manifester une capacité à comprendre des descriptions de constituants de la réalité.

Il y a des exemples concrets de cela dans la littérature sur l'IA. En effet, les systèmes d'IA contemporains, en ayant accès à des définitions linguistiques par l'entremise d'internet ou en étant capables d'analyser la fréquence de co-occurrence de certains mots, par exemple, peuvent manifester un certain niveau de compréhension de contenus représentationnels sous la forme de descriptions en montrant qu'ils peuvent décrire adéquatement les objets du monde. Pour illustrer cela, nous pouvons nous référer au cas de l'assistante vocale Siri. L'assistante vocale Siri peut effectivement accéder à des définitions génériques des mots contenues dans une phrase reçue comme entrée (Błachnio, 2019 : 25). C'est ensuite à partir de cet ensemble de significations que Siri peut tenter de rechercher les relations existantes entre les mots contenus dans une expression, afin d'en déterminer la structure syntaxique[52], pour enfin tenter d'identifier l'intention de l'humain qui l'a énoncé et fournir une sortie adéquate à celle-ci (Błachnio, 2019 : 25, 26). Ainsi, le cas de Siri tend à montrer que pour manifester de la compréhension, une machine doit pouvoir démontrer une connaissance des significations des mots et expressions utilisés qui est assez proche de celle de son interlocuteur : les définitions auxquelles a accès l'agent artificiel doivent être équivalentes aux descriptions que nous pouvons donner des choses auxquelles réfèrent ces mots.

Prenons maintenant l'exemple des réseaux neuronaux récurrents (*recurrent neural networks*)[53] : des réseaux de neurones artificiels qui permettent des cycles d'interactions entre les

---

[52] C'est d'ailleurs de cette manière que l'analyse sémantique contribue à l'identification des marqueurs comme les fonctions syntaxiques et les rôles sémantiques (cette idée est soulevée dans le chapitre précédent).

[53] Plusieurs agents conversationnels populaires contemporains (*ChatGPT*, par exemple) fonctionnent grâce à cette technologie.

différents neurones de chacune des couches contenues dans ces réseaux. Cette technologie permet de faire interagir les constituants d'une phrase (les mots) dans le but de prédire les mots à venir les uns à la suite des autres, jusqu'à former des phrases complètes signifiantes (IBM Cloud Education, *What are recurrent neural networks*). L'illustration ci-dessous présente un réseau neuronal récurrent. La récurrence s'exécute sur les couches cachées (hidden layers) :

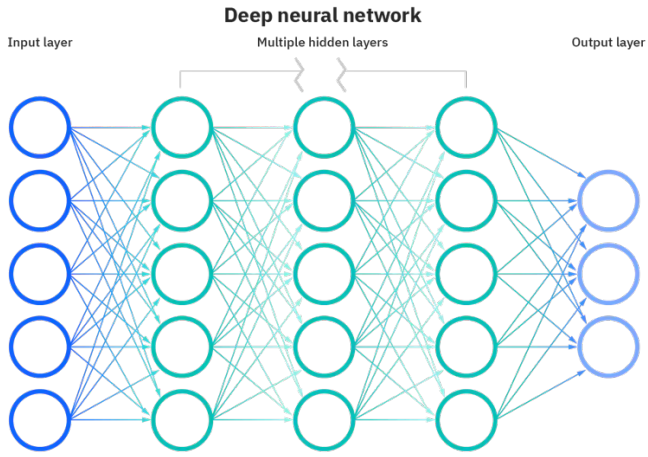



Figure 5 : Tirée de *IBM Cloud Education, What are neural networks?*

Philosophiquement, ce qu'il y a d'intéressant à soulever à propos du traitement et de la génération du langage à l'aide de réseaux neuronaux récurrents, c'est d'abord que ces réseaux réalisent une fonction similaire à celle de la compositionnalité dans la compréhension du langage chez les êtres humains. L'agencement des mots permet de prédire le mot qui sera le suivant et ainsi de suite, de sorte que le système peut composer des phrases complètes et interpréter le contenu des phrases reçues en analysant la façon dont les constituants y sont combinés. Ensuite, ces systèmes permettent, grâce à leur capacité d'analyse des fréquences de co-occurrence des mots, de créer ce

que les chercheurs et programmeurs nomment *words embeddings*[54]. Les *words embeddings* sont des représentations des significations d'ensembles de mots à partir des relations sémantiques que ces mots entretiennent entre eux. Plus deux mots « co-occurrent » fréquemment, plus leurs significations sont considérées comme proches, par le système. Ci-bas se trouve une illustration qui présente des représentations des significations des mots « *Apple* », « *Amazon* », « *Obama* » et « *Trump* » et l'évolution de celles-ci à travers le temps. Il suffit de jeter un coup d'œil à l'évolution de la représentation de la signification du mot « Apple » à travers les années ou bien à celle du mot « Trump » pour voir comment l'analyse, par l'IA, des fréquences de co-occurrences permet de créer des représentations très adéquates des relations qu'un mot entretien avec d'autres, et ce, relativement aux contextes historiques et sociaux dans lesquels on l'utilise.

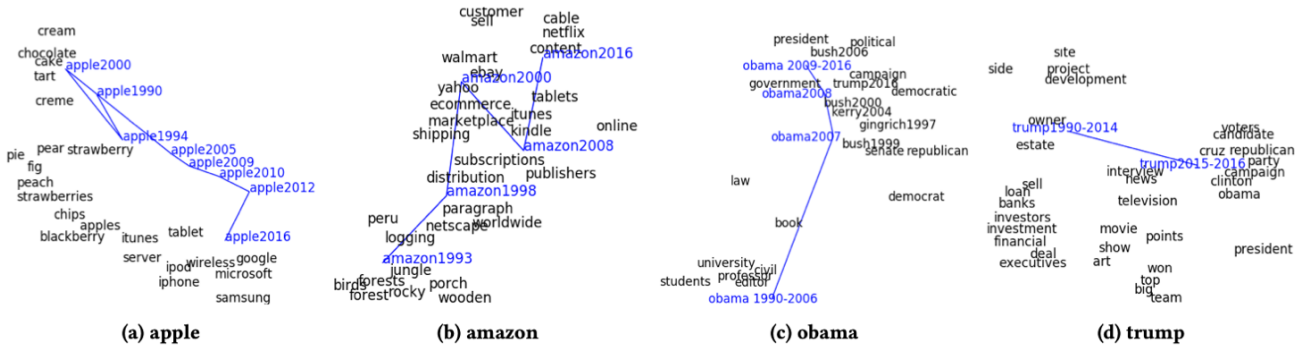



Figure 6 : Tirée de Yao et al., 2018.

Voici un autre exemple pertinent de construction de représentations sémantiques à partir de l'analyse des fréquences de co-occurrence de mots qui réfèrent à différentes espèces d'animaux, situés sur une droite par rapport aux mots « *small* » et « *large* ». L'exemple montre bien comment

[54] L'expression « *vector space models* » est aussi utilisée pour référer à ce genre de modèles de représentation des significations en IA. Voir Dasgupta et al., 2020 : 4.

ces représentations sémantiques peuvent agir comme des représentations de contenus descriptifs pour une machine, notamment en lui permettant de décrire et de comparer les tailles des différents animaux :

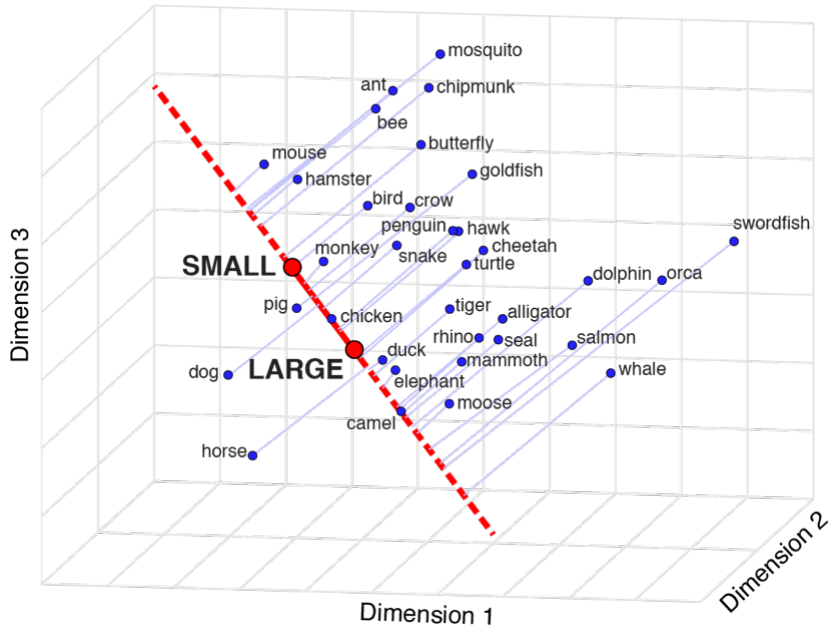



Figure 7 : Tirée de Grand et al., 2018 : 5.

Revenons sur ce que nous venons de présenter. Nous avons divisé notre présentation du contenu des représentations sémantiques en deux éléments : le contenu référentiel et descriptif. Cette division a été faite, car elle permet de rendre compte des deux cas de manifestation de la compréhension par des systèmes d'IA présentés plus haut : la compréhension de phrases représentant la réalité externe et la compréhension de phrases représentant et exprimant des états internes. Dans les deux cas, la compréhension de ces phrases peut se manifester par une capacité à saisir de leurs référents et/ou à saisir des propriétés qu'elles décrivent (par exemple : la présence d'états mentaux chez une personne/entités). Dans notre présentation du contenu référentiel et descriptif, nous avons porté une attention particulière à l'importance de la capacité à décrire la réalité pour les machines qui manifestent de la compréhension, et ce, parce que même la capacité

à référer à des choses particulières semble exiger la capacité à décrire ces dernières. Nous avons alors montré comment des systèmes d'IA contemporains sont en mesure de produire des descriptions de la réalité, à travers l'accès à des définitions d'objets et une capacité à saisir des relations sémantiques entre des mots.

### 3.5 Les représentations sémantiques et la notion de concept

Il est raisonnable de penser que pour manifester de la compréhension d'une phrase exprimant un état mental comme « *la peur qu'une situation S se produise* », par exemple, un agent conversationnel devrait pouvoir être capable de saisir le référent (la personne qui a peur) et devrait en saisir les implications : il devrait pouvoir montrer qu'il comprend ce que cela implique d'avoir peur qu'advienne la situation *S* et qu'il comprend alors ce que signifie « *avoir peur que* » et « *S* ». Pour montrer cela, il semble effectivement nécessaire que l'IA ait accès à une certaine représentation[55] de « avoir peur » et de « la situation *S* ». Dans le même sens, quand un agent conversationnel dit quelque chose comme : « je promets de vous réveiller à 7h00 », personne ne croirait que l'assistante lancera vraiment une alarme à l'heure énoncée sans considérer au moins qu'il peut adéquatement se représenter les obligations ou les engagements qu'implique de dire « je promets », ainsi que la signification de 7h00. On doit, pour lui faire confiance et dormir paisiblement, considérer que la représentation sémantique qu'elle forme de cette expression est fonctionnellement équivalente à la nôtre. Présenté autrement, puisque les représentations que nous possédons, comme nos concepts de [promesse] et de [7h00], sont des constituants de nos états mentaux, il semble absolument nécessaire qu'une machine à laquelle on attribuerait la capacité à éprouver des états mentaux, doive pouvoir comprendre ce sur quoi portent ces états mentaux. Dans

[55] Ou qu'il puisse au moins démontrer qu'il en possède une.

ce cas-ci, si on considère que la machine « a l'intention de nous réveiller à 7h00 », car elle nous l'a promis, et bien cela semble impliquer nécessairement que nous considérions qu'elle comprend « promesse » et « 7h00 ». En tant qu'être humain, nous comprenons ces mots, car ils réfèrent, pour nous, à des concepts que nous possédons[56]. De là, nous analyserons maintenant d'un point de vue critique l'hypothèse voulant que des machines puissent entretenir des représentations sémantiques équivalentes à nos représentations mentales en nous concentrant sur la notion de « concept ».

Pour commencer, définissons d'une façon générale ce que nous entendons par « concept ». D'entrée de jeu, lorsque l'on s'interroge sur la nature des concepts, nous nous frappons rapidement aux termes « intension » et « extension ». L'intension d'un concept est normalement conçue comme sa définition ou son sens (*meaning*) alors que son extension est la chose ou les choses auxquelles le concept s'applique dans le monde[57]. En ce sens, les mots et les agencements de mots exprimant des concepts dans des phrases complètes peuvent décrire des états de faits réels ou possibles et nous les comprenons et savons comment nous représenter ces états de fait, notamment parce qu'ils ne sont pas seulement extensionnels, mais aussi intensionnels. En intelligence artificielle, on cherche souvent à créer des systèmes capables de représenter des « connaissances » (*knowledge representation*). Il existe différentes façons d'implémenter l'équivalent d'une sorte de « connaissance conceptuelle » dans un système autonome[58]. Par exemple, certains systèmes

---

[56] L'idée que les concepts sont constituant des états mentaux est présentée dans Fodor, 1998 : 6. Si la situation *S* en question est en fait l'occurrence d'un décès ou d'une rupture amoureuse, par exemple, il va de soi que pour que la machine comprenne l'état mental de peur qu'elle exprime, elle doit nécessairement montrer qu'elle comprend ce qu'est et ce qu'implique un décès ou une rupture amoureuse.

[57] Rudolf Carnap présente la distinction intension/extension comme suit : l'extension d'une expression individuelle (des expressions qui réfèrent) est l'individu à qui elle réfère et son intension est un concept individuel. L'extension d'un prédicateur (au sens d'expression qui exprime une prédication) est la classe des individus à qui le prédicat s'applique, tandis que son intension est la propriété qu'il exprime. (Carnap, 1947 : I)

[58] On peut penser aux définitions auxquelles accède Siri pour donner des significations aux mots qu'elle traite. Dans le même sens, c'est grâce à ce que l'on appelle des *CONCEPTUAL captions* que des systèmes comme DALL-E sont dits capables « d'apprendre » ce que représentent des millions d'images à travers un processus d'apprentissage

mettent en relations certains mots, dans le but de créer des sortes de toiles de significations qui permettent à la machine d'identifier des relations entre les mots d'une langue (Davis, Shrobe, Szolovits, 1993 : 25). D'autres systèmes peuvent aussi contenir des représentations de *frames* : des cadres situationnelles incluant certains concepts importants dont l'usage est lié à certaines situations la vie réelle, permettant à la machine d'interagir correctement, selon le type d'interaction (Davis, Shrobe, Szolovits, 1993 : 25, 26). Certains systèmes peuvent aussi contenir des ontologies qui représentent la nature des choses qui composent le monde et leurs interactions. On retrouve, dans ce genre d'ontologie, des concepts englobant comme des classes, des universaux et des propriétés, des relations hiérarchiques et d'interactions, ainsi que des individus particuliers (Smith, 2004). Ces ontologies peuvent porter sur des domaines restreints, comme elles peuvent représenter des caractéristiques globales des objets du monde, leurs propriétés et leurs relations[59]. Lorsque les développeuses et développeurs en sciences informatiques créent ces types de systèmes, elles et ils tentent précisément de définir les intensions des concepts que nous utilisons, afin que leurs systèmes autonomes aient accès à des informations adéquates à propos des mots qui expriment nos concepts et qu'ils soient ainsi en mesure de communiquer avec nous de façon adéquate.

Suivant ces idées, pour bien montrer comment une machine peut manifester une capacité à entretenir des représentations sémantiques adéquates et fonctionnellement équivalentes à nos concepts et donc manifester de la compréhension des mots qui les désignent, nous développerons une réflexion autour de quatre approches philosophiques des concepts. Ces approches ont été sélectionnées à partir de deux critères : leur importance respective dans la philosophie du langage

---

supervisé. Les *conceptual captions* sont des descriptions textuelles d'images formulées dans une langue naturelle. Voir : Sharma, Ding, Goodman, Soricut, 2018; Ramesh et al., 2021.

[59] Fodor, quant à lui, nomme ce genre d'ontologies des hiérarchies sémantiques (Fodor, 1998 : 89, 90).

et de l'esprit et leur pouvoir explicatif des capacités linguistiques et cognitives dont peuvent faire preuve des machines dotées d'intelligence artificielle. Tout au long de cette réflexion, nous mettrons l'accent sur ce que la compréhension conceptuelle permet de faire. Ainsi, nous tâcherons de montrer ce qu'une machine dotée d'IA doit pouvoir faire avec les mots qu'elle traite pour manifester une capacité à former et entretenir des représentations sémantiques fonctionnellement équivalentes à nos concepts les plus utiles. Ultimement, cette analyse de la notion nous mènera à proposer des conditions qu'une machine doit respecter pour manifester de la compréhension conceptuelle.

### 3.5.1 L'approche classique des concepts

L'approche classique des concepts, la première théorie que nous souhaitons présenter, cherche à expliquer en quoi consiste l'intension d'un concept. Cette approche soutient que pour qu'une chose tombe sous un concept, elle doit respecter certaines conditions nécessaires et suffisantes qui forment son intension (Fodor, 1998 : 24, 25). Pour qu'une chose tombe sous le concept « bâtiment », par exemple, elle doit être un objet immobile, être d'une taille suffisante pour que l'on puisse y entrer, elle doit être une construction humaine, etc. De là, remarquons déjà que la compréhension conceptuelle semble impliquer nécessairement une connaissance de plusieurs significations et donc de plusieurs concepts en même temps, puis des relations qui existent entre eux. Remarquons aussi que les conditions nécessaires et suffisantes qui forment l'intension d'un concept sont très proches de ce que nous appelons ordinairement une « définition ».

Les systèmes d'IA qui fonctionnent grâce à la programmation de toiles sémantiques (*semantic networks/webs*) en intelligence artificielle permettent de bien montrer la pertinence de l'approche classique des concepts. Le système *KL-ONE*, par exemple, permet de représenter la signification des concepts sous la forme de descriptions formées à partir de leurs rapports avec des sous-types. Dans la figure suivante, CAMION (*TRUCK*) peut être décrit comme un sous-type de VÉHICULE (*VEHICULE*), ayant un certain nombre de ROUES (*WHEELS*), une capacité à contenir des choses (*CARGO CAPACITY*), etc.

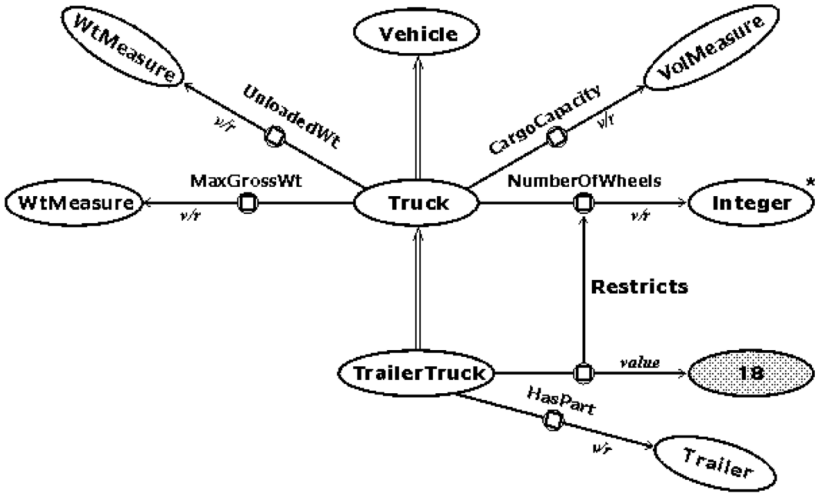



Figure 8 : représentation du concept TRUCK tirée de Sowa, 1992.

Les intensions des concepts programmés dans le système dictent ainsi comment il doit « représenter le monde » et comment une machine dans laquelle on implémenterait le système d'IA devrait parler du monde et de ses constituants. Si nous demandions à un système comme *KL-ONE*, par exemple, ce qu'est un camion, sa réponse nous informerait sur la façon dont il se représente l'intension de « camion » en termes de descriptions ou de caractéristiques que partagent ses membres (son extension). Qu'un agent conversationnel avec qui nous communiquons ait accès ou non à l'ensemble des conditions nécessaires et suffisantes qui forment l'intension d'un concept qu'elle

semble comprendre et maîtriser, il paraît nécessaire, quoiqu'il en soit, qu'il y ait une adéquation entre l'intension de nos concepts et la connaissance que l'agent peut manifester des relations entre le mot qui désigne le concept en question et d'autres qui déterminent son contenu : les conditions nécessaires et suffisantes qui permettent de l'appliquer aux objets. En ce sens, il paraît clair qu'une façon pertinente de concevoir l'équivalence fonctionnelle entre les représentations sémantiques des mots que traite une machine et les concepts qu'expriment pour nous ces mots soit celle qui définit l'équivalence en termes de descriptions adéquates des objets auxquels ils sont applicables. Pour que l'on soit portés à considérer que la machine comprend le mot d'une façon équivalente à la nôtre, il est donc plausible qu'elle doive pouvoir montrer qu'elle comprend quels autres mots forment son intension pour pouvoir montrer qu'elle connaît ses conditions d'applications.

### 3.5.2 Les concepts comme prototypes

En réponse à l'approche classique s'est développée l'approche des concepts comme prototypes. Celle-ci veut que les concepts soient des sortes de prototypes mentaux auxquels les objets du monde correspondraient plus ou moins adéquatement. La formulation de l'approche est en grande partie attribuée à Eleanor Rosch et Carolyn B. Mervis[60]. Dans un article de 1975, les chercheuses ont testé l'hypothèse voulant que la capacité d'individus à classer des objets sous des catégories sémantiques relativement larges et abstraites (meuble, par exemple) et d'autres, plus étroites et particulières (chaise, par exemple), leur permette, une fois qu'ils ont classé suffisamment d'objets, de former des sortes de prototypes mentaux des membres qui représentent le plus fidèlement les catégories en question (Rosch, Mervis, 1975 : 575, 576). Pour en arriver là, elles

[60] Les travaux du second Wittgenstein, surtout ceux portant sur les jeux de langage et leurs « airs de famille », sont précurseurs de cette approche (Rosch, Mervis, 1975 : 574, 575).

élaborent le critère suivant permettant de créer des prototypes : plus un membre d'une catégorie sémantique partage des ressemblances (des airs de famille) avec les autres membres de la catégorie et moins il en partage avec des membres d'autres catégories, plus il devient prototypique dans la première catégorie (Rosch, Mervis, 1975 : 575, 598, 599). Cette hypothèse fut testée à travers six tests empiriques qui la corroborent (Rosch, Mervis, 1975).

Pour nous, c'est d'abord la description que l'approche de Rosch et Mervis donne des ressemblances qui la rend intéressante sur le plan philosophique. Elles parlent de ressemblances en termes d'attributs (*discrete attributes*) comme « a des pattes », « se conduit », etc. Ainsi, cette façon de décrire les ressemblances entre des choses rappelle l'approche classique des concepts, mais présente également des différences importantes qui nous renseignent plus précisément sur ce que manifeste la capacité à appliquer des concepts. En effet, l'approche des prototypes permet qu'un objet tombe sous une catégorie sémantique sans nécessairement devoir respecter un ensemble de conditions rigides, parce que son appartenance se détermine surtout par rapport aux ressemblances qu'il partage ou qu'il ne partage pas avec d'autres objets et non par rapport à des conditions prédéterminées. L'intension d'un concept est donc plutôt présentée comme un prototype idéal construit par un processus d'apprentissage, à partir d'actes de discrimination et de comparaison d'objets. C'est alors davantage la capacité à réaliser ces actes cognitifs que la connaissance de conditions prédéterminées que manifeste la maîtrise d'un concept. Il faut aussi remarquer qu'une piste de réflexion intéressante pour la question de la communication humain-IA se trouve dans cette approche : elle permet certaines disparités entre les façons dont nous pouvons comprendre un même concept, puisque les objets tombent sous un concept s'ils ressemblent au prototype personnel du concept qu'un agent forme. Nous pouvons effectivement entrainer des systèmes d'IA à comparer des objets physiques ou linguistiques comme des descriptions, par exemple, dans le but de les amener à être en mesure d'identifier des objets, leurs ressemblances et leurs différences pour

les classer sous des concepts. Imaginez d'ailleurs la situation suivante : si nous demandions à Siri ou à *ChatGPT* d'expliquer la différence entre un chat et un zèbre ou entre une ville et un pays, une réponse adéquate de leur part serait effectivement de nommer des différences et des ressemblances à propos de leurs propriétés respectives. Il est alors raisonnable de penser que ce qui ferait en sorte qu'une machine semblerait vraiment comprendre les significations de ces mots, c'est une capacité à montrer qu'elle se représente assez bien des idées générales (des prototypes) de ce à quoi ressemble un chat, un zèbre, une ville et un pays, puis qu'elle connaît les relations comme les ressemblances et les différences qui existent entre les propriétés de ces différents prototypes[61].

### 3.5.3 Les approches holistes

Dans les deux parties précédentes, l'idée voulant que nos concepts entretiennent entre eux des relations revenait fréquemment, notamment parce que les descriptions que nous formons des choses incluent toujours l'utilisation d'autres mots et donc d'autres concepts. Le partage cohérent et efficace d'informations nécessite d'ailleurs de comprendre les relations qui existent entre les concepts que nous utilisons (comme savoir qu'un CHAT est un ANIMAL, que les animaux sont VIVANTS, etc.) Les approches holistes de la signification cherchent à rendre compte de cette réalité et sont particulièrement utiles pour étudier les interactions humain-IA. En effet, les approches holistes proposent généralement que la signification d'un concept soit au moins en partie déterminée par sa place à l'intérieur du système conceptuel d'un agent et donc par ses relations

[61] Dans le même sens, nous n'exigeons pas d'une personne, pour considérer qu'elle comprend un mot, qu'elle connaisse l'entièreté des propriétés des choses qui tombent sous celui-ci ou bien qu'elle ait exactement la même représentation que la nôtre en tête, lorsque nous parlons d'une chose ou d'une catégorie de choses pour considérer qu'elles comprennent.

avec les autres concepts du système[62]. La sémantique inférentialiste de Robert Brandom en est un exemple. Cette approche propose que la signification d'un concept soit formée des inférences qu'il permet (Brandom, 2007 : 654). L'intension d'un concept serait donc composée des prémisses et des conclusions que l'on pourrait retrouver dans des raisonnements inférentiels impliquant l'usage du concept en question (Brandom, 2007 : 654). La citation suivante présente bien l'idée à partir de « *temperature* » : « *The content of a concept such as temperature is, on this view, captured by the constellation of inferential commitments on undertakes in applying it: commitment, namely, to the propriety of the inferences from any of its circumstances of appropriate application to any of its appropriate consequences of application.* » (Brandom, 2007 : 654)

Dans le même sens, une autre approche holiste importante mérite ici d'être mentionnée : « la théorie des théories ». Cette approche veut que les concepts d'une personne entretiennent des relations entre eux, parce qu'ils forment des théories à propos du monde (voir : Carey, 2009 : 501, 502). De là, le changement de la signification d'un concept à l'intérieur d'une théorie d'une personne impliquerait nécessairement des changements sur les autres concepts qui constituent la théorie en question. Les approches des concepts comme théories impliquent alors aussi que les contenus des concepts soient au moins en partie déterminés par les inférences que ceux-ci permettent de réaliser pour en arriver à une théorie ayant une certaine portée explicative à propos de différents phénomènes du monde (voir : Carey, 2009 : 501, 502). En guise d'exemple, on peut

[62] D'ailleurs, l'une des problématiques fondamentales qui émanent de l'holisme sémantique est celle de la supposée impossibilité de partager des significations communes. Le contenu d'un même concept peut être différent d'une personne à l'autre s'il n'occupe pas la même place dans leurs systèmes conceptuels respectifs. Cela a pour effet de théoriquement rendre impossible la communication interpersonnelle, notamment parce que l'on ne peut pas prétendre pouvoir interpréter parfaitement ce qui disent les autres. Les difficultés théoriques et pratiques entourant l'holisme de la signification ont mené certaines et certains philosophes et linguistes à adopter une position plus flexible : une position molécullariste. Le molécularisme de la signification propose qu'il existe des relations de dépendance entre les significations des termes d'un langage, mais ne soutient pas que les significations de tous les mots d'une langue soient interreliées (Pagin, 2008 : 213).

simplement penser aux concepts utilisés dans l'expression et l'attribution d'attitudes propositionnelles comme ceux de DÉSIR et de CROYANCE, qui nous permettent constamment d'expliquer les comportements des gens autour de nous. Les concepts de DÉSIR et de CROYANCE, selon ces approches, sont alors des outils pour expliquer des choses. Ces concepts nous sont utiles pour inférer de l'information et expliquer ce qui se passe dans le monde. En comprenant ce qu'est une CROYANCE, je peux expliquer pourquoi j'observe un comportement C chez quelqu'un. En comprenant ce à quoi s'applique le concept de CHAT, je peux expliquer et même prédire le comportement qu'aura un chat lors de sa première rencontre avec un chien[63].

De là, il faut voir que les approches holistes de la signification sont immensément intéressantes pour enrichir la partie traitant de la sémantique à l'intérieur de notre définition fonctionnelle de la compréhension. Elles montrent encore une fois que la compréhension d'un concept passe par la capacité de l'agent qui le comprend à apprécier les relations sémantiques qui existent entre ce concept et d'autres (les CROYANCES sont liées aux COMPORTEMENTS et les CHATS aux CHIENS, par exemple). Ces relations, pour ces approches, prennent la forme d'interactions continues entre les concepts que nous possédons, au sens où la compréhension d'un concept exige d'apprécier les engagements inférentiels et les impacts de la modification du contenu d'un concept sur l'ensemble du système conceptuel de l'agent. Il est clair que les contenus des

[63] Une autre approche holiste des concepts précise la relation entre la compréhension conceptuelle et le comportement : la sémantique des rôles conceptuels (*conceptual role semantics*) de Ned Block. L'approche de Block prend comme point de départ la distinction entre contenu étroit et contenu large, puis propose que le contenu étroit d'un mot ou d'une expression consiste en son rôle conceptuel (Block, 1986 : 623). Pour Block, les mots tirent donc en partie leurs significations des pensées dans lesquelles ils sont utilisés (Block, 1986 : 637, 643, 644). Le rôle conceptuel d'une expression est donc la façon dont nous pouvons l'utiliser pour réaliser des raisonnements (comme des inférences pour expliquer ses propres comportements ou ceux d'autrui, notamment). Ce qui doit ici retenir notre attention, c'est l'influence que la compréhension du langage a sur nos actions : les significations des mots (qui sont analysables en tant que concepts, une fois qu'elles sont considérées comme jouant des rôles causaux dans nos raisonnements) permettent la production de raisonnements et ces raisonnements influencent la production de comportements (Block, 1986 : 663, 628). En ce sens, les concepts médient entre les entrées sensorielles d'une entité et ces comportements (Block, 1986 : 628).

concepts qui sont partagés entre des agents humains et non humains sont différents. Les réflexions sur l'holisme et ses difficultés mettent justement en lumière l'idée que nous avons constamment l'impression de partager des significations parfaitement équivalentes à celles d'autres agents, alors que ce n'est pas nécessairement le cas. Toutefois, ce que les approches que nous venons de présenter nous disent, c'est qu'il est suffisant de savoir comment un concept interagit avec les autres pour le comprendre : il est suffisant de savoir quelles inférences (qui incluent l'usage d'autres concepts) nous pouvons en tirer (si nous suivons Brandom) et de savoir comment il contribue à nos explications des phénomènes du monde et comment son contenu est influencé par nos autres concepts (si nous suivons l'approche des théories).

Pour conclure cette partie, retenons qu'il est utile de réfléchir aux représentations sémantiques en IA comme des outils qui sont censés être interreliés et jouer des rôles causaux dans les raisonnements des machines qui les implémentent. Il est en effet facile d'imaginer une scène où l'on questionnerait une machine dotée d'IA à propos des raisons derrière ses actions. Il paraît clair que cette même machine devrait être en mesure d'utiliser correctement des termes comme « vouloir que », « croire que », etc. (généralement les termes utilisés pour exprimer des attitudes propositionnelles) pour que l'on soit portés à considérer qu'elle comprend l'explication qu'elle donne de ces actions. La situation inverse semble aussi évidente : si nous confions à une machine un désir que nous entretenons, il va de soi que nous nous attendions à ce qu'elle puisse montrer qu'elle comprend ce que le désir implique, qu'elle comprenne quel rôle causal le concept de désir joue dans notre esprit et, par le fait même, comment celui-ci a le potentiel d'influencer nos comportements futurs.

### 3.6 La place de la sémantique dans une définition fonctionnelle de la compréhension

Les lignes précédentes montrent ce que sont les concepts en présentant les deux principales choses qu'ils nous permettent de faire : ils permettent de nous représenter la réalité en la décrivant et ils nous permettent de raisonner à propos de la réalité en réalisant certains processus cognitifs. On peut, par exemple, observer, connaître et comprendre les relations qu'entretiennent nos différents concepts et les différents objets du monde qui tombent sous ceux-ci. On peut aussi réaliser des actions comme catégoriser des objets, discriminer entre ceux-ci, analyser leurs ressemblances et leurs différences. On peut aussi amplifier notre connaissance en inférant des informations à partir des concepts que nous possédons. Enfin, cette capacité à tirer des inférences à partir de nos concepts nous permet aussi de prendre des décisions, puis d'exprimer les motifs de nos actions, ainsi que celles des autres. Avant de formuler une nouvelle partie pour notre définition fonctionnelle de la compréhension, il convient d'ajouter une nuance importante à ce que nous avons jusqu'à maintenant présenté. Nous avons dit, à plusieurs reprises, que les concepts qu'une machine peut sembler comprendre doivent être adéquats par rapport aux nôtres. C'est l'idée de « représentations sémantiques *fonctionnellement équivalentes* à nos représentations mentales comme nos concepts ». Nous avons alors soulevé différentes conditions (sur lesquelles nous reviendrons dans quelques lignes) qui paraissent nécessaires pour qu'une machine manifeste de la compréhension des significations des mots et expressions d'une langue. La nuance que nous souhaitons d'abord apporter a trait aux limites de la capacité représentationnelle de l'IA, puis au fait que la réalité est en constant changement, faisant en sorte que les significations des mots que nous utilisons pour en parler doivent s'ajuster à ces changements. Cette nuance servira à ajouter une dernière condition à notre définition. Celle-ci s'exprime en trois points : l'IA doit pouvoir manifester une capacité à comprendre d'une façon assez « générale », elle doit éviter d'entretenir des biais ou de sembler en entretenir et, enfin, elle doit pouvoir apprendre à comprendre.

D'abord, en proposant qu'une machine dotée d'IA doive manifester une compréhension « générale », nous voulons dire que la compréhension des significations dont elle peut faire preuve doive être assez générale pour être adéquate à l'intérieur de différents contextes conversationnels. La construction des représentations sémantiques dont ces machines sont capables est effectivement influencée, comme nous l'avons présenté, par des définitions génériques ou des fréquences de co-occurrences entre des mots, puis par des mécanismes de combinaisons des significations des mots qui forment les phrases traitées et formulées. De là, tous ces facteurs (dépendamment du système implémenté dans la machine) doivent être suffisamment englobants pour permettre le traitement adéquat d'un large éventail d'expressions dont les significations sont influencées par les différents contextes communicationnels.

C'est de cette même difficulté qu'émane le deuxième aspect de la nuance que nous souhaitons mettre de l'avant : les représentations sémantiques peuvent contenir des biais et donc mal représenter la réalité. Il existe d'ailleurs des exemples concrets de cela. C'est le cas, par exemple, du prédicteur de texte *WWT* présenté dans un article de Carlo Perrotta, Neil Selwyn et Carrie Ewin sur les modèles de langage *GPT*, où l'on montre que celui-ci exprime des biais de genre en proposant à une utilisatrice d'utiliser le mot *MEN* après que l'utilisatrice ait employé *SOLDIER* (Perrota et al., 2022 : 11). Cette erreur est due au fait que le modèle a été entrainé sur des corps de textes comportant des récits de guerre qui sont plus fréquemment à propos d'hommes (Perrota et al., 2022 : 11)[64]. Dans la même optique, un article de Bender et al. sur les modèles de langage de grande envergure rapporte aussi qu'il est bien reconnu, au sein des communautés de

[64] L'article de Perrotta et al. présente aussi d'autres manifestations de biais de genre présents chez l'IA. Le système propose, par exemple, l'utilisation de certains mots référant à des objets différents à la suite de l'entrée de phrases mettant en scène une personne portant un prénom souvent porté par des hommes ou un prénom souvent porté par des femmes (Perrota et al., 2022 : 13-15).

chercheuses et de chercheurs en IA que ces modèles manifestent des biais à travers des associations stéréotypées ou par l'expression de « sentiments » négatifs à l'égard de certains groupes (Bender et al., 2021 : 614, 615). Ces exemples permettent alors de constater comment les représentations sémantiques d'un système peuvent être inadéquates et donc non équivalentes par rapport à nos concepts[65]. Maintenant, il faut remarquer que ces mises en situation reflètent des mécompréhensions entre les systèmes d'IA et les utilisateurs et utilisatrices parce que les représentations de ces systèmes sont inadéquates *pour* les concepts que possèdent les utilisateurs/utilisatrices. Cela dit, ces systèmes tirent les informations qu'ils utilisent pour formuler des textes à partir d'internet et reflètent donc les positions qu'entretiennent les auteurs et autrices de ces textes trouvés sur internet. En ce sens, on pourrait argumenter que les représentations sémantiques que se montre capable d'entretenir un système sont nécessairement adéquates, parce qu'elles reflètent *adéquatement* l'information présente sur internet, mais qu'elles ne respectent pas toujours ce qui est normativement prescrit ou valorisé dans notre société. Toutefois, il faut bien voir qu'une incapacité d'un système à montrer que sa représentation sémantique d'un mot qui exprime un concept est adéquate *par rapport* au concept particulier possédé par son interlocutrice ou interlocuteur mène quand même à une situation de mécompréhension : la machine et l'humain *ne comprennent pas* le mot de la même façon[66].

Enfin, cette difficulté nous conduit au dernier aspect de cette nuance : l'IA doit pouvoir apprendre à comprendre. Avant 1885, les femmes n'étaient pas autorisées à s'enrôler dans les Forces armées canadiennes. Une machine aurait alors été justifiée d'entretenir, durant cette période,

---

[65] Dernièrement, on réfère de plus en plus aux situations dans lesquelles une machine représente mal les faits en utilisant le terme « hallucinations », au sens métaphorique où l'IA est dite comme « hallucinant » une réalité déformée. Voir : Azamfirei, 2023.

[66] Dans le même sens, il faut aussi remarquer que l'on pourrait évaluer de façon objective le caractère adéquat de la représentation par rapport à la réalité, en nous demandant si elle la décrit bien.

une représentation sémantique du mot « *soldier* » qui présente une forte proximité sémantique avec d'autres mots référant majoritairement ou exclusivement à des hommes. Aujourd'hui, ce n'est bien évidemment plus le cas, et les représentations sémantiques que forme une machine dotée d'IA doivent pouvoir être modifiées pour s'ajuster adéquatement à la réalité et donc à nos représentations mentales de la réalité (comme notre concept de SOLDAT), si nous voulons pouvoir dire qu'elles leur sont fonctionnellement équivalentes.

En gardant cela en tête, nous pouvons maintenant formuler une deuxième partie à notre définition fonctionnelle de la compréhension attribuable aux machines dotées d'IA. Rappelons la première partie formulée dans le deuxième chapitre : la compréhension implique de se montrer capable d'identifier des relations et leurs directions respectives exprimées entre les mots et groupes de mots contenus dans des expressions, puis de respecter ces relations et leurs directions lors de la formulation de sorties. Ces relations peuvent être de différents types : relations spatiales, relations temporelles, relations agentielles, intentions de communication, etc. Les relations identifiées dans les expressions traitées peuvent d'ailleurs référer à des relations réelles entre les référents des mots traités.

Tenant compte de ce qui a été présenté dans ce chapitre, il convient d'y ajouter ce qui suit : Puisqu'en tant qu'êtres humains nos concepts sont des contenus mentaux qui agissent comme des blocs nous permettant de construire nos pensées les plus complexes et de les partager aux autres, pour manifester de la compréhension, une machine dotée d'IA doit montrer qu'elle peut former et entretenir des représentations sémantiques adéquates et donc fonctionnellement équivalentes aux concepts que nous possédons. Pour ce faire, ces représentations sémantiques doivent représenter correctement le monde (elles doivent donc s'ajuster au monde, aux différents contextes conversationnels et sociaux et êtres exemptes de biais propices à causer des situations de mécompréhensions), puis doivent permettre à la machine de remplir des fonctions cognitives

équivalentes à celles que nous permet de remplir notre propre appareillage conceptuel. Notre mémoire présente plusieurs conditions que notre étude de la nature des représentations et des concepts montre comme nécessaires pour manifester de la compréhension linguistique : l'agent doit pouvoir référer à des entités et des objets réels, puis être en mesure de les décrire adéquatement. Il doit aussi pouvoir manifester une connaissance des relations sémantiques qui existent entre les significations des mots qu'il traite et utilise (notamment parce que c'est une capacité nécessaire à la compréhension de descriptions). Ensuite, les représentations sémantiques de l'agent doivent lui permettre de réaliser certains actes cognitifs comme des inférences, des prises de décisions et des explications comportementales. Elles doivent aussi lui permettre de réaliser des actes de discrimination, de comparaison et de catégorisations d'objets linguistiques qui réfèrent à des objets réels, sur la base de leurs des ressemblances et de leurs différences, puis sur la base de la possession ou de la non-possession de certaines propriétés.

# CHAPITRE 4 — LE LANGAGE PRAGMATIQUE ET LES INTENTIONS DE COMMUNICATION

## 4.1 Introduction au chapitre

Dans ce chapitre, nous nous intéresserons maintenant plus en profondeur à la notion d'intention de communication et au traitement, par l'IA, du langage pragmatique, dans le but de formuler de nouvelles conditions nécessaires à l'attribution de la compréhension aux machines dotées d'IA et d'ainsi compléter notre définition fonctionnelle de ce processus mental. Ultimement, cette partie permettra de montrer comment des machines dotées d'IA peuvent manifester une capacité à comprendre ce que l'on fait avec le langage et une capacité à elles-mêmes réaliser des actions à travers le langage.

## 4.2 Les intentions de communication

D'entrée de jeu, nous nous intéresserons à trois notions importantes de la philosophie du langage ordinaire qui nous permettront de montrer la pertinence du concept d'intention de communication pour notre définition de la compréhension. La première notion est celle « d'information implicite », que nous aborderons principalement à partir des travaux de Paul Grice. L'étude de celle-ci nous permettra de montrer que la communication présuppose une capacité à comprendre de l'information implicite contenue dans les énoncés de nos interlocutrices et interlocuteurs et donc, par le fait même, une capacité à identifier les intentions de communication

de ceux-ci. La deuxième notion est celle de « règles d'usage ». Son étude nous amènera à montrer que le langage est régi par un ensemble de règles d'usages qui nous permettent de comprendre ce que nos interlocutrices et interlocuteurs veulent dire. Ainsi, nous montrerons que la maîtrise de ces règles permet de comprendre ce que les autres *font* avec le langage (et, par le fait même, leurs intentions) et donc de manifester une compréhension des fonctions du langage. Enfin, la troisième notion que nous aborderons est celle « d'acte de langage ». Celle-ci sera utilisée pour montrer que les machines peuvent manifester une capacité à comprendre des intentions de communication à travers le traitement d'actes de langage. De là, nous irons un peu plus loin et proposerons qu'en se montrant capables de traiter et de réaliser des actes de langage, elles peuvent aussi paraître capables de comprendre l'expression d'états mentaux et d'entretenir des états mentaux.

### 4.2.1 L'information implicite

Grice est reconnu pour avoir développé la notion « d'implicature » en linguistique pragmatique. Ce terme fait référence à ce qui est suggéré sans être explicitement dit : une implicature est donc ce qui est *implicite* aux prises de paroles (Beyssade, 2017 : 45). Grice s'est particulièrement intéressé à la notion d'implicature conversationnelle, par opposition à l'implicature conventionnelle. Une implicature conventionnelle consiste en une inférence sémantique, alors que l'implicature conversationnelle a trait à la dimension pragmatique du langage. Vu l'intérêt beaucoup plus grand que Grice a porté aux implicatures conversationnelles et compte tenu de notre volonté de maintenant nous concentrer sur la pragmatique, nous étudierons les implicatures de ce type.

Voici un exemple classique d'une implicature conversationnelle :

« *A : Smith doesn't seem to have a girlfriend these days.*

*B: He has been paying a lot of visits to New York lately.* » (Grice, 1989 : 32)

Dans cet exemple de conversation, la réponse de *B* suggère de façon implicite que Smith a ou pourrait avoir une copine qui habite à New York (Grice, 1989 : 32). Ce qui fait que cette information est « conversationnellement » implicite, c'est précisément le fait que pour interpréter l'énoncé de *B* adéquatement (pour saisir l'implicature), on doit connaître son contexte d'énonciation (connaître l'énoncé de *A,* par exemple). Sans connaissance des circonstances d'énonciation, la personne qui doit interpréter l'énoncé devrait en effet s'en remettre à sa connaissance des significations littérales conventionnelles de l'énoncé et des mots qui le composent, ce qui serait ici insuffisant pour en comprendre la signification complète. En ce sens, la signification de l'énoncé change selon son contexte « conversationnel » (Beyssade, 2017 : 46). L'exemple suivant l'illustre aussi bien, où Marie indique implicitement qu'elle ne participera pas à la fête, car elle doit travailler :

« […] Alain : Est-ce que tu viens à la fête ce soir ?

Marie : J'ai du travail. » (Beyssade, 2017 : 46).

Maintenant, il faut aussi savoir que l'existence des implicatures s'explique, selon Grice, par la structure des conversations auxquelles nous participons. Pour montrer cela, Grice introduit le principe de coopération qu'il analyse en quatre maximes générales que nous suivons lors des conversations ordinaires (quantité : on doit donner juste assez d'information, qualité : on doit dire des choses vraies et justifiées, relation : on doit être pertinent, manière : on doit livrer un discours bref, ordonné, non obscur et non ambigu) (Grice, 1989 : 26-28; Beyssade, 2017 : 48). Justement, l'approche de Grice conçoit nos conversations comme des activités conjointes qui poursuivent des buts partagés entre les participantes et participants, ce qui explique d'ailleurs la volonté de chacune et chacun de suivre le principe de coopération (Beyssade, 2017 : 47).

De là, on pourrait donc dire, par exemple, que ce sont les volontés de parler de la vie amoureuse de Smith et des participants de la fête de Alain qui font en sorte que les énoncés

formulés par le participant *B* et par Marie, dans les exemples donnés précédemment, permettent au participant *A* et à Alain de comprendre l'information implicite que l'on y trouve. C'est donc principalement leurs contextes d'énonciation et la présupposition que les participants suivent le principe de coopération qui guident l'interprétant vers ce que l'interlocutrice ou l'interlocuteur *veut dire*[67]. De façon plus complète, pour saisir une implicature conversationnelle présente dans un énoncé, une personne peut utiliser les données suivantes : les significations conventionnelles des mots utilisés et leurs références, le principe de coopération, le contexte d'énonciation, des connaissances d'arrière-plan et la présupposition du fait que leur interlocutrice ou interlocuteur a aussi accès aux données qui tombent sous les catégories précédentes (Grice, 1989 : 31).

Suivant ces idées, il est primordial de remarquer l'importance de réaliser les inférences nécessaires pour traiter adéquatement le discours ordinaire. Nos conversations sont remplies d'informations implicites qui nécessitent une sensibilité très fine à leurs contextes linguistiques et extralinguistiques pour être comprises, et ce, aussitôt qu'une interprétation purement littérale est insuffisante pour rendre compte de la pleine signification d'un énoncé (comme de l'intention de communication de son énonciatrice ou énonciateur). La littérature scientifique sur les situations de mécompréhension entre des assistants vocaux et des humains présente d'ailleurs l'importance de la saisie de l'information implicite et de la sensibilité aux contextes pour qu'un dialogue soit réalisé avec succès. Dans un article de Liang et al. (2019), par exemple, on présente un système d'intelligence artificielle qui peut jouer au jeu *Hanabi*, un jeu de cartes coopératif qui exige des participants de partager implicitement de l'information à propos des cartes que leurs coéquipières

[67] Bref, pour saisir une implicature conversationnelle présente dans un énoncé, une personne peut utiliser les données suivantes : les significations conventionnelles des mots utilisés et leurs références, le principe de coopération, le contexte d'énonciation, des connaissances d'arrière-plan et la présupposition du fait que leur interlocutrice ou interlocuteur a aussi accès aux données qui tombent sous les catégories précédentes (Grice, 1989 : 31).

et coéquipiers ont entre les mains pour gagner la partie. Voici comment raisonne l'IA capable d'implicatures conversationnelles développée par les chercheuses et chercheurs :

> *The AI is unable to see its own hand and therefore relies entirely on the hints its teammate provides to narrow down the identity of its cards. The AI knows for each card in its hand which colors and which numbers are possible given the current state of the game. When the AI receives a hint, it considers each possible hand and the flow of the decision tree in Figure 2 to determine what kind of hint it would give if its teammate was the one with that possible hand. The AI can then use this process to determine the likely identities of the cards in its hand by checking if the hint it would have given aligns with the hint it actually received from its teammate. For example, if the AI receives the hint "you have one 2", then only the possibility of the hinted card being actionable aligns with this hint. All possible hands that would result in this hint being generated require that the card be playable. Therefore the AI interprets the hint it received as "play this card."* (Liang et al., 2019 : 7)

Nous voulons alors ici porter l'attention de la lectrice et du lecteur sur ce que permet le traitement d'implicature conversationnelle : elle permet d'interpréter, à l'aide du contexte de conversation (le jeu, les cartes précédemment jouées, les cartes que l'autre joueur possède, les indices donnés, etc.) l'intention de l'interlocuteur, même si son intention n'est pas exprimée dans le sens littéral de l'énoncé.

De cette présentation théorique et de cet exemple d'implémentation du traitement informatique des implicatures, il convient donc de retenir que la compréhension linguistique se manifeste aussi par la saisie des intentions de communication qui sont parfois implicites et qui exigent donc de nous, comme d'une machine dotée d'IA, de réaliser des inférences à partir des contextes conversationnels pour les saisir[68]. Remarquons aussi déjà que l'interprétation des actes de langage et donc des intentions de communication peut se traduire par l'attribution d'attitudes

[68] Au moins, la machine devrait pouvoir tenter une inférence et confirmer l'information implicite comprise auprès de l'utilisatrice/l'utilisateur par une méthode de confirmation comme celle présentée dans Berg et al. (2011), afin que l'utilisateur soit amené à fournir plus explicitement l'information à comprendre au système.

propositionnelles. Dans le jeu présenté en guise d'exemple, on peut interpréter l'énoncé de l'autre joueur en lui attribuant une attitude « l'autre joueur *désire que* je joue la carte C ».

### 4.2.2 Les règles d'usage

Dans les lignes précédentes, nous avons montré comment les contextes communicationnels influencent les significations de certaines expressions, notamment parce qu'ils permettent d'identifier des informations implicites et donc non littérales. Cela rappelle d'ailleurs la distinction entre *timeless meaning,* une signification fixe et *occasion meaning,* une signification déterminée par le contexte d'usage (Grice, 1968). En ouvrant un dictionnaire, par exemple, nous sommes confrontés à des définitions littérales, mais on peut aussi parler de règles d'usage qui influencent la signification des mots que nous connaissons[69]. Prenons le cas d'expressions indexicales telles que le pronom « tu », par exemple. « Tu » change de référent dépendamment du contexte d'usage, la définition qu'on lui donne dans un dictionnaire, elle, ne change pas : ça reste le pronom personnel de la deuxième personne du singulier. On peut aussi penser à l'énonciation des mots « oui, je le veux » lors d'une cérémonie de mariage, qui contribue, dans ce contexte, à unir un couple officiellement devant une institution pourvue d'autorité religieuse ou légale. La connaissance, la compréhension et la maîtrise de ces significations d'usage dépendent d'un savoir-faire linguistique que l'on acquiert en évoluant à l'intérieur d'une communauté linguistique. Ainsi, la communication réussie nécessite de comprendre cette dimension sociale du langage où les significations des mots et des expressions sont profondément ancrées dans des comportements, des habitudes et des conventions sociales.

---

[69] Souvenons-nous d'ailleurs que, comme présenté dans le chapitre précédent, pour P. F. Strawson les significations de certaines expressions consistent directement et complètement en des ensembles d'instructions d'usages de celles-ci (Strawson, 1950 : 327, 328).

L'un des plus importants théoriciens du concept de règle d'usage fut Ludwig Wittgenstein. C'est effectivement dans sa « seconde philosophie » que l'on retrouve le fameux concept de « jeux de langage ». Alors qu'il rédige son ouvrage *Recherches philosophiques* (2004), il s'intéresse à la question de savoir ce qu'est un langage. Il en viendra alors à la conclusion que les mots et les expressions sont des outils qui peuvent remplir diverses fonctions (nommer des objets du monde en étant une parmi d'autres) (Wittgenstein, 2004 : 11, 23, 27). Ainsi, il conçoit le langage comme un ensemble de jeux ayant entre eux des ressemblances (Wittgenstein, 2004 : 7, 67). Cette approche propose donc que les significations soient à trouver dans l'observation des règles d'usages du langage : ordonner, décrire un événement, réagir à une demande, traduire, etc., réduisant ainsi la signification d'une expression à son usage (Wittgenstein, 2004 : 23, 24). En guise d'exemple, on peut penser à l'expérience de pensée classique dans laquelle un maçon dit à son assistant « brique », faisant en sorte que ce dernier lui apporte une brique (Wittgenstein, 2004 : 2). Dans ce jeu de langage, les deux participants savent et donc *comprennent* que « brique » sert à provoquer ce comportement de l'assistant. De là, nous souhaitons attirer l'attention de la lectrice et du lecteur sur l'importance du comportement dans la conception du langage que fournit la deuxième philosophie d Wittgenstein :

> *[…] L'intention est incorporée à la situation, aux coutumes des hommes et à leurs institutions. Si la technique du jeu d'échecs n'existait pas, je ne pourrais pas avoir l'intention de faire une partie d'échecs. Si j'ai d'emblée déterminé par l'intention la forme de la phrase, c'est parce que je sais parler français. (Wittgenstein, 2004 : 337)*

La connaissance du langage, au sens de savoir comment l'utiliser pour faire des choses, permet effectivement la réalisation de certains comportements intentionnels à travers le langage. D'ailleurs, il est aisé d'imaginer une machine qui saurait participer à des jeux de langage comme

celui présenté plus haut. Nous posons effectivement constamment des questions à nos assistants vocaux artificiels et leur demandons de réaliser certaines tâches à notre place. Dans Mavrina et al. (2019), par exemple, on présente des situations où des mécompréhensions ont lieu entre des enfants et l'assistante vocale Alexa qui sont expliquées, par les autrices et auteurs, par un manque de précision quant aux contextes des demandes des enfants et par un manque de clarté quant aux intentions derrière leurs prises de paroles (Mavrina et al., 2019 : 10). L'article en question présente ainsi la conversation entre un enfant, un adulte et Alexa, où l'enfant cherche à connaître des activités accessibles, auxquelles participer dans sa ville, pour lui et sa famille durant la journée, en demandant : « *Alexa, what can I do together with my mom and sister?* » (Mavrina et al., 2019 : 10) Alexa ne comprendra pas la demande et l'adulte devra retravailler la formulation pour en venir à énoncer : « *Alexa, events in [city name]* », énoncé auquel Alexa répondra adéquatement (Mavrina et al., 2019 : 10). Ce qu'il faut remarquer, ici, c'est que puisque nous avons lu l'article, nous connaissons l'intention de l'enfant : obtenir des informations sur les événements qui ont lieu dans la ville où il habite, durant la journée, dans le but de connaître des activités à la portée de sa famille. Ainsi, nous sommes en mesure de bien comprendre ce qu'il *veut dire*. Une interprétation littérale de l'énoncé de l'enfant ne mène toutefois certainement pas à cette analyse complexe. Nous sommes donc en présence d'un potentiel jeu de langage : lorsque l'enfant demande ordinairement ce qu'il peut faire avec des membres de sa famille, il cherche des activités/événements auxquels participer, tout en étant limité par leur localisation géographique et on doit inférer, entre autres, cette information implicite. Il est alors facile d'imaginer qu'une façon adéquate de répondre, pour Alexa, consisterait en une demande de précision : demander, par exemple, si l'enfant cherche des activités à faire chez soi ou dans la ville. En ne sachant pas répondre à un tel énoncé, la machine montre qu'elle ne comprend pas ce que l'enfant fait avec le langage : qu'elle ne connaît pas le jeu de langage auquel il participe et donc qu'elle ne comprend pas son intention.

Dans la même optique, remarquons comment l'adulte qui réussira à obtenir une réponse satisfaisante de la part d'Alexa n'utilise pas la forme interrogative conventionnelle de la demande, mais plutôt une forme presque impérative (Mavrina et al., 2019 : 10). Dans ce cas, en répondant correctement, il semble, en effet, qu'Alexa se montre capable d'identifier correctement l'intention de communication derrière l'énoncé en montrant qu'elle comprend la règle d'usage suivie : par l'énonciation des expressions « *Alexa, events in [city name]* » et « *Alexa, what can I do together with my mom and sister?* », les utilisateurs expriment la *volonté que*/*exigent que* le système fournisse une liste des événements qui ont lieu en ville.

Suivant ces idées, les concepts de règles d'usages et de jeux de langage sont grandement pertinents pour voir comment une machine dotée d'IA peut manifester de la compréhension, car ils montrent que de participer correctement à ces jeux en suivant et en interprétant des règles d'usage, c'est manifester une compréhension des fonctions du langage et donc des intentions de communication de ceux et celles qui l'utilise. Les machines dotées d'IA que nous connaissons aujourd'hui sont précisément développées pour remplir des fonctions précises et le langage est un moyen de les réaliser. Il va donc de soi qu'elles doivent, pour manifester de la compréhension, être en mesure de participer aux différents jeux de langage auxquels nous prenons part, pour adéquatement répondre à nos prises de paroles et donc montrer qu'elles comprennent ce que nous faisons avec le langage.

### 4.2.3 Les actes de langage

Dans la partie précédente, nous avons utilisé, en guise d'exemples d'expressions qui peuvent exiger une maîtrise de certaines règles d'usages pour être comprises, « tu » et « oui, je le veux ». Dans les deux cas, il faut connaître leurs contextes d'énonciation et les règles d'usage à suivre pour les utiliser et les interpréter adéquatement. Nous voulons maintenant nous intéresser

aux énoncés qui, comme la réponse positive à une demande en mariage, peuvent et même devraient être analysés comme des actes de langage (*speech acts*)[70]. De là, dans les prochaines lignes, nous présenterons la théorie des actes de langage de John Searle, puis montrerons ensuite que le traitement des actes de langage permet à l'IA de manifester une capacité à comprendre des intentions de communication. Enfin, nous proposerons que cette capacité puisse aussi manifester une capacité à comprendre et à avoir des états mentaux.

La théorie des actes de langage de Searle propose que le langage soit réductible à une forme de comportement que nous réalisons en suivant des règles (Searle, 1969 : 22). Sa théorie se veut donc une analyse de ces règles et montre comment nous faisons différentes choses avec le langage en précisant les règles que nous suivons dans la réalisation de chacun de ces comportements (Searle, 1969 : 22). Nous nous concentrerons sur la catégorie d'actes de langage dits « illocutoires » (*illocutionary acts*), car cette catégorie contient les actes qui forment ce que nous nommons la dimension pragmatique du langage. On y compte les actes de questionner/demander, ordonner, promettre, affirmer, etc. (Searle, 1969 : 24) Pour Searle, ceux-ci sont effectués conjointement aux actes propositionnels, au sens où les actes propositionnels (comme le fait d'énoncer des propositions dans lesquelles on « réfère » et on « prédique ») sont accompagnés de la réalisation d'actes illocutoires (comme « l'assertion » ou « l'affirmation » d'une proposition contenant une prédication) (Searle, 1969 : 29). Ce qui nous permet d'identifier un acte de langage illocutoire, c'est sa force illocutoire qui s'incarne à travers différents indicateurs potentiels comme l'intonation, l'emphase, l'ordre des mots contenu dans la proposition exprimée, la ponctuation, les

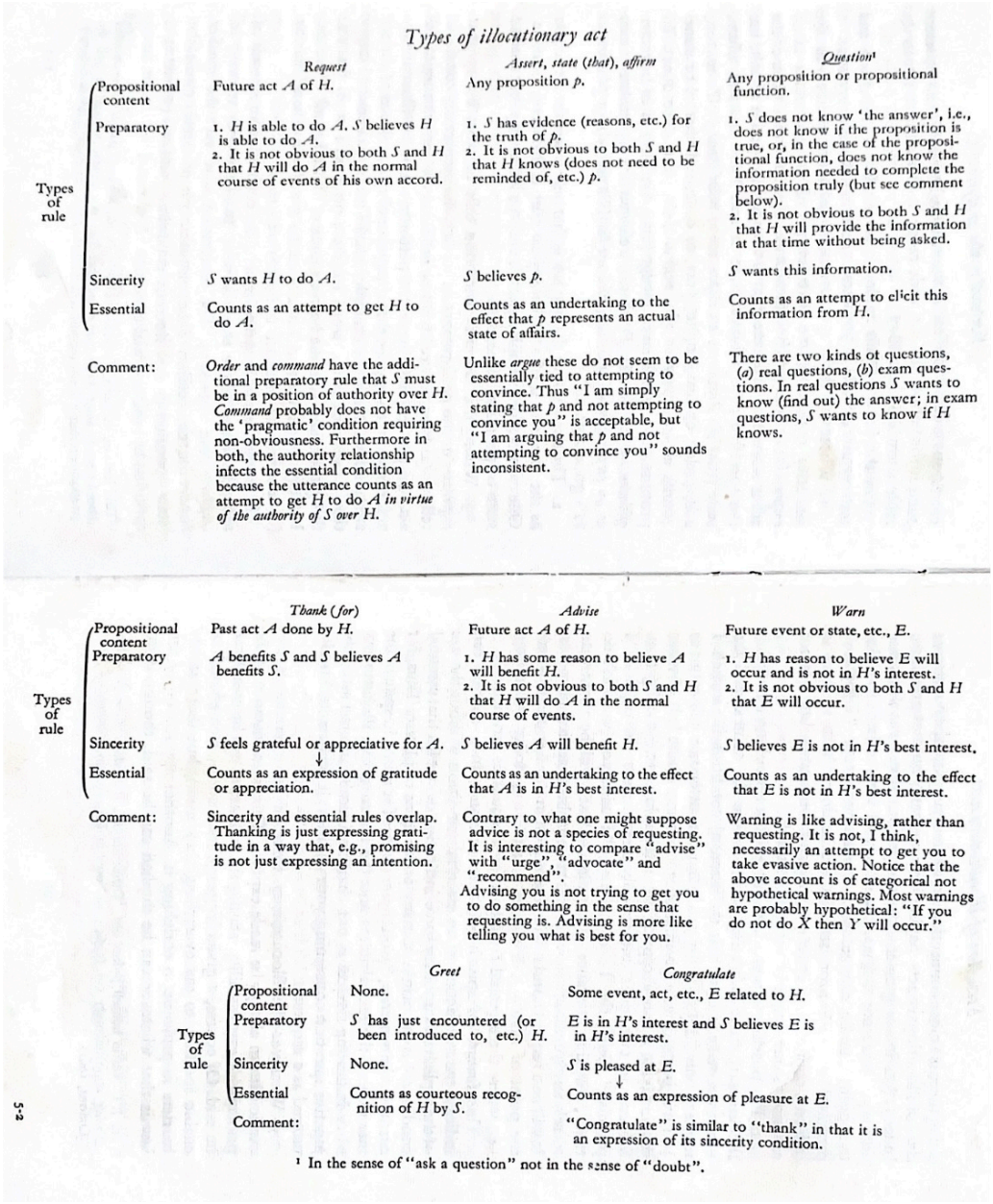
[70] L'acte de référer peut aussi être analysé comme un acte de langage. Searle le fait d'ailleurs (Searle, 1969 : 26-29). On pourrait toutefois argumenter que ce n'est pas le cas, tandis que l'acte de réaliser des vœux de mariage en est clairement un qu'on peut analyser comme la réalisation d'une promesse. L'exemple de « oui je le veux » est alors utilisé, car il frappe davantage l'imaginaire et rend notre propos clair. De plus, l'acte de référer est analysé comme un acte propositionnel, dans l'approche de Searle et nous souhaitons nous concentrer sur les actes illocutoires (Searle, 1969 : 24).

temps de verbe et les verbes performatifs (qui constituent l'action qu'ils expriment) (Searle, 1969 : 30). Il faut ainsi voir que la force illocutoire d'un acte de langage est très intimement liée à l'intention de son énonciateur. D'ailleurs, le passage suivant tiré de « *A classification of illocutionary acts* » (1976) illustre bien comment les actes illocutoires expriment des états psychologiques/mentaux :

> *A man who states, explains, asserts or claims that p expresses the belief that p; a man who promises, vows, threatens or pledges to do A expresses an intention to do A; a man who orders, commands, requests H to do A expresses a desire (want, wish) that H do A; a man who apologizes for doing A expresses regret at having done A; etc. In general, in the performance of any illocutionary act with a propositional content, the speaker expresses some attitude, state, etc., to that propositional content. Notice that this holds even if he is insincere, even if he does not have the belief, desire, intention, regret or pleasure which he expresses, he nonetheless expresses a belief, desire, intention, regret or pleasure in the performance of the speech act. This fact is marked linguistically by the fact that it is linguistically unacceptable (though not self-contradictory) to conjoin the explicit performative verb with the denial of the expressed psychological state. Thus one cannot say 'I state that p but do not believe that p', 'I promise that p but I do not intend that p', etc. Notice that this only holds in the first person performative use. One can say, 'He stated that p but didn't really believe that p', 'I promised that p but did not really intend to do it', etc. The psychological state expressed in the performance of the illocutionary act is the sincerity condition of the act […]* (Searle, 1976 : 4)

Il est, en effet, fascinant de voir comment les propositions qui sont utilisées pour réaliser des actes de langage, lorsqu'elles sont énoncées dans des contextes adéquats (ce que Searle nomme les conditions préparatoires) et que leurs énonciatrices ou énonciateurs sont sincères, en viennent à compter comme l'expression de certains états mentaux. L'acte de promettre, par exemple, consiste ainsi en l'expression d'une intention sincère (condition de sincérité) qui compte comme un engagement à faire ce qui est promis (condition essentielle), réalisé à travers l'énonciation d'une phrase qui prédique une action future (condition du contenu propositionnel), dans un contexte

approprié (condition préparatoire) (Searle, 1969 : 63)[71]. Le tableau suivant présente de façon détaillée différents types d'actes de langage illocutoires comme l'assertion, la demande, l'avertissement, etc., ainsi que leurs conditions de réalisation :



*Types of illocutionary act*

| Types of rule | *Request* | *Assert, state (that), affirm* | *Question*[1] |
|---|---|---|---|
| Propositional content | Future act *A* of *H*. | Any proposition *p*. | Any proposition or propositional function. |
| Preparatory | 1. *H* is able to do *A*. *S* believes *H* is able to do *A*.<br>2. It is not obvious to both *S* and *H* that *H* will do *A* in the normal course of events of his own accord. | 1. *S* has evidence (reasons, etc.) for the truth of *p*.<br>2. It is not obvious to both *S* and *H* that *H* knows (does not need to be reminded of, etc.) *p*. | 1. *S* does not know 'the answer', i.e., does not know if the proposition is true, or, in the case of the propositional function, does not know the information needed to complete the proposition truly (but see comment below).<br>2. It is not obvious to both *S* and *H* that *H* will provide the information at that time without being asked. |
| Sincerity | *S* wants *H* to do *A*. | *S* believes *p*. | *S* wants this information. |
| Essential | Counts as an attempt to get *H* to do *A*. | Counts as an undertaking to the effect that *p* represents an actual state of affairs. | Counts as an attempt to elicit this information from *H*. |
| Comment: | *Order* and *command* have the additional preparatory rule that *S* must be in a position of authority over *H*. *Command* probably does not have the 'pragmatic' condition requiring non-obviousness. Furthermore in both, the authority relationship infects the essential condition because the utterance counts as an attempt to get *H* to do *A* *in virtue of the authority of S over H*. | Unlike *argue* these do not seem to be essentially tied to attempting to convince. Thus "I am simply stating that *p* and not attempting to convince you" is acceptable, but "I am arguing that *p* and not attempting to convince you" sounds inconsistent. | There are two kinds of questions, (*a*) real questions, (*b*) exam questions. In real questions *S* wants to know (find out) the answer; in exam questions, *S* wants to know if *H* knows. |

| Types of rule | *Thank (for)* | *Advise* | *Warn* |
|---|---|---|---|
| Propositional content | Past act *A* done by *H*. | Future act *A* of *H*. | Future event or state, etc., *E*. |
| Preparatory | *A* benefits *S* and *S* believes *A* benefits *S*. | 1. *H* has some reason to believe *A* will benefit *H*.<br>2. It is not obvious to both *S* and *H* that *H* will do *A* in the normal course of events. | 1. *H* has reason to believe *E* will occur and is not in *H*'s interest.<br>2. It is not obvious to both *S* and *H* that *E* will occur. |
| Sincerity | *S* feels grateful or appreciative for *A*. | *S* believes *A* will benefit *H*. | *S* believes *E* is not in *H*'s best interest. |
| Essential | Counts as an expression of gratitude or appreciation. | Counts as an undertaking to the effect that *A* is in *H*'s best interest. | Counts as an undertaking to the effect that *E* is not in *H*'s best interest. |
| Comment: | Sincerity and essential rules overlap. Thanking is just expressing gratitude in a way that, e.g., promising is not just expressing an intention. | Contrary to what one might suppose advice is not a species of requesting. It is interesting to compare "advise" with "urge", "advocate" and "recommend".<br>Advising you is not trying to get you to do something in the sense that requesting is. Advising is more like telling you what is best for you. | Warning is like advising, rather than requesting. It is not, I think, necessarily an attempt to get you to take evasive action. Notice that the above account is of categorical not hypothetical warnings. Most warnings are probably hypothetical: "If you do not do *X* then *Y* will occur." |

| Types of rule | *Greet* | *Congratulate* |
|---|---|---|
| Propositional content | None. | Some event, act, etc., *E* related to *H*. |
| Preparatory | *S* has just encountered (or been introduced to, etc.) *H*. | *E* is in *H*'s interest and *S* believes *E* is in *H*'s interest. |
| Sincerity | None. | *S* is pleased at *E*. |
| Essential | Counts as courteous recognition of *H* by *S*. | Counts as an expression of pleasure at *E*. |
| Comment: | | "Congratulate" is similar to "thank" in that it is an expression of its sincerity condition. |

[1] In the sense of "ask a question" not in the sense of "doubt".

Figure 9 : Tableau tiré de Searle, 1969 : 66, 67.

[71] Ces conditions forment l'ensemble des conditions de réalisation des actes illocutoires dans la théorie searlienne de 1969.

Suivant ces idées, il faut retenir que les intentions de communication sont constitutives des actes de langage. Les actes de langage sont effectivement des actions *intentionnelles* réalisées à travers le langage. Ainsi, l'interprétation des actes de langage nécessite normalement l'identification des intentions communicationnelles (il faut comprendre l'état psychologique de son énonciatrice ou énonciateur pour les comprendre complètement). Néanmoins, il est clair qu'il est possible de réaliser des énoncés qui sont formellement identiques à des actes de langage, mais qui n'ont pas d'intentions de communication comme ils devraient en avoir (on peut promettre sans avoir l'intention de respecter sa promesse et une machine peut utiliser les énoncés qui servent à ordonner, promettre et demander, sans réellement avoir quelque chose comme une intention communicationnelle véritable). Néanmoins, il faut bien voir que lorsqu'une machine dotée d'IA réalise un acte de langage ou manifeste une capacité à en comprendre, elle peut aussi paraître capable d'avoir des intentions de communication et de comprendre les intentions de communications des autres, car ces intentions sont normalement constitutives des actes de langage réussis. Il est alors plausible que nous soyons portés à attribuer la capacité à comprendre des intentions et à en avoir à des machines intelligentes, sur cette base. Effectivement, aussitôt que l'on considère que les énoncés qu'une machine interprète et réalise consistent en de véritables actes de langage, il est normal que nous soyons portés à considérer qu'elle peut aussi entretenir ou au moins comprendre les états psychologiques intentionnels qui sont normalement associés aux différents actes de langage.

La littérature scientifique en sciences informatiques est d'ailleurs truffée de tentatives de formalisation du traitement artificielle des actes de langage. On peut penser à l'exemple donné dans le chapitre 2 de ce mémoire à propos du traitement de la syntaxe du mot *may* en tant que verbe servant à effectuer une demande plutôt qu'en tant que nom (Wang et al., 2021). De plus, un article de Berg et al. (2011) présente un modèle théorique intéressant du traitement des actes de langage

à travers un robot conversationnel qui cherche à identifier les buts de ses utilisatrices et utilisateurs, puis à s'y adapter (Berg et al., 2011 : 4). Les catégories d'actes de langage auxquelles les chercheurs s'intéressent sont les suivantes : « *information seeking, information providing and action requesting*. » (Berg et al., 2011 : 4) Les auteurs présentent l'importance de ne pas les confondre : demander au système de chercher des restaurants (une recherche d'information), ce n'est pas lui demander de planifier un itinéraire jusqu'à ceux-ci ni de réserver une table sur internet (une demande de réalisation d'une action) (Berg et al., 2011 : 4). Ainsi, voici comment le système étudié peut en venir à comprendre l'intention d'un utilisateur à partir de l'exemple d'une demande de réalisation d'une action, en identifiant l'acte de langage réalisé, son contexte de réalisation et la forme syntaxique de l'énoncé : « *"Could you please open the window?" can be formalised as a quintuple: (concern, user, action request, question, smart room).* » (Berg et al., 2011 : 4)

### 4.3 La place du traitement du langage pragmatique dans une définition fonctionnelle de la compréhension

Nous voilà ainsi confrontés à trois catégories de situations où l'interprétation littérale du langage (la capacité à se représenter les significations des mots traités) est insuffisante si elle ne tient pas aussi compte de la dimension pragmatique du langage ordinaire. Ces situations présentent effectivement des moments clés où une interprétation correcte du discours nécessite de prendre en compte les contextes d'usages, les règles d'usages et les actions réalisées avec le langage.

Nous avons choisi, dans ce chapitre, de nous concentrer sur la notion d'intention de communication, pour deux raisons : la première est que nous souhaitons, dans le cadre de ce travail, dresser une sorte de panorama des analyses possible de la compréhension à partir des trois dimensions classiques du langage : syntaxique, sémantique et pragmatique. C'est donc une dernière analyse de la compréhension linguistique qui ressort des lignes précédentes. La

compréhension linguistique peut se manifester par la saisie des intentions de communication qui peuvent être implicites, parfois exprimées à travers la capacité à suivre des règles d'usages (comme le fait de participer à des jeux de langage) et, finalement, qui sont souvent exprimées par la réalisation d'actes de langage. Et de là suit la deuxième raison : le concept d'intention de communication est particulièrement pertinent pour aborder le traitement du langage pragmatique par l'IA, puisqu'il cadre parfaitement avec les tâches que remplissent les systèmes que nous connaissons. Ceux-ci servent effectivement différentes fonctions : ils nous informent, répondent à des demandes, emmagasinent des informations, apprennent, etc. De là, il faut constamment que ces machines soient en mesure d'identifier nos besoins pour pouvoir y répondre. Elles peuvent les comprendre si et seulement si elles identifient correctement les intentions derrière nos prises de paroles. À l'inverse, pour que l'on considère qu'elles peuvent comprendre ce qu'elles-mêmes disent, il paraît aussi clair qu'elles doivent être en mesure d'exprimer de telles intentions. Repensons à l'exemple de la promesse donné dans le chapitre 3 de ce mémoire : qui ferait confiance à un assistant vocal qui dirait : « je promets de vous réveiller à 7h00 demain matin », sans considérer au minimum que l'assistant exprime quelque chose d'au moins fonctionnellement équivalent à une promesse et donc qu'il exprime quelque chose d'équivalent à l'intention sincère de réveiller l'utilisatrice ou l'utilisateur à sept heures tapantes ?

Enfin, suivant ce qui a été présenté plus haut, il convient d'ajouter une dernière partie à notre définition fonctionnelle de la compréhension. Rappelons son contenu jusqu'à présent : la compréhension implique de se montrer capable d'identifier des relations et leurs directions respectives exprimées entre les mots et groupes de mots contenus dans des expressions, puis de respecter ces relations et leurs directions lors de la formulation de sorties. Ces relations peuvent être de différents types : relations spatiales, relations temporelles, relations agentielles, intentions de communication, etc. Les relations identifiées dans les expressions traitées peuvent d'ailleurs

référer à des relations réelles entre les référents des mots traités. Ensuite, puisqu'en tant qu'êtres humains nos concepts sont des contenus mentaux qui agissent comme des blocs nous permettant de construire nos pensées les plus complexes et de les partager aux autres, une machine dotée d'IA doit montrer qu'elle peut former et entretenir des représentations sémantiques adéquates et donc fonctionnellement équivalentes aux concepts que nous possédons. Pour ce faire, ces représentations sémantiques doivent représenter correctement le monde (elles doivent donc s'ajuster au monde, ainsi qu'aux différents contextes conversationnels et sociaux et être exemptes de biais propices à causer des situations de mécompréhensions), puis doivent permettre à la machine de remplir des fonctions cognitives équivalentes à celles que nous permet de remplir notre propre appareillage conceptuel. L'agent doit donc pouvoir référer à des entités et des objets réels, puis être en mesure de les décrire adéquatement. Il doit aussi pouvoir manifester une connaissance des relations sémantiques qui existent entre les significations des mots qu'il traite et utilise (notamment parce que c'est une capacité nécessaire à la compréhension de descriptions). Les représentations sémantiques de l'agent doivent aussi lui permettre de réaliser certains actes cognitifs utiles comme des inférences, des prises de décisions et des explications comportementales. Elles doivent finalement lui permettre de réaliser des actes de discrimination, de comparaison et de catégorisations d'objets linguistiques qui réfèrent à des objets réels, sur la base de leurs ressemblances et de leurs différences, puis sur la base de la possession ou de la non-possession de certaines propriétés.

Compte tenu de ce qui a été présenté dans ce chapitre, ajoutons ce qui suit : pour manifester une capacité à comprendre fonctionnellement équivalente à la nôtre, un agent artificiel doit pouvoir se montrer capable de comprendre ses intentions de communication et celles d'autrui. Une grande partie des expressions que doivent traiter les agents artificiels contemporains expriment effectivement des intentions de communication implicitement ou explicitement. Nous avons

présenté trois conditions que doit pouvoir satisfaire une machine dotée d'IA pour manifester une capacité à comprendre des intentions de communications : elle doit d'abord pouvoir se montrer capable de traiter les énoncés à partir de leurs contextes conversationnels pour réaliser des inférences qui montrent qu'elle comprend l'information implicite nécessaire à saisir pour bien comprendre ce que l'autre *veut dire*. Elle doit aussi pouvoir montrer qu'elle connaît des règles d'usages du langage et leurs impacts sur les significations des expressions qu'elle traite. Ultimement, cela montrerait qu'elle connaît les fonctions du langage et qu'elle peut donc reconnaître ce que les autres cherchent à faire avec celui-ci. Finalement, elle doit pouvoir réaliser et traiter des actes de langage. En manifestant une capacité à exprimer et à traiter des actes de langage, la machine se montre encore une fois capable de comprendre les intentions de communication de ses interlocutrices et interlocuteurs. Ultimement, puisque ces actes sont réalisés grâce à une capacité de l'agent qui les réalise à suivre des règles (comme la condition de sincérité, par exemple), il va de soi qu'une machine qui manifesterait de la compréhension devrait pouvoir montrer qu'elle connaît ces règles et comprend ce que leur respect implique. On peut d'ailleurs penser au fait qu'il est nécessaire qu'une machine qui réaliserait l'acte de « promettre » respecte réellement sa promesse pour que l'on soit portés à considérer qu'elle comprend vraiment ce qu'elle dit et donc qu'elle comprend l'intention de communication normalement associée à une promesse : un engagement intentionnel à l'obligation de faire ce qui est promis. Ces deux conditions (connaître les règles qui régissent la réalisation d'actes de langage et se montrer capable de comprendre ce que leur respect implique) se présentent comme nécessaires pour qu'une machine manifeste de la compréhension d'actes de langage, car elles remplissent les mêmes fonctions que la capacité à réaliser et interpréter de véritables actes de langage dirigés par des intentions de communication.

# CONCLUSION

C’est sur cette caractérisation exhaustive de la compréhension linguistique applicable à ces nouvelles machines que se clôt l’essentiel de notre réflexion. Nous souhaitons maintenant revenir brièvement sur les grandes étapes que nous avons franchies pour l’élaborer, mais aussi nous pencher sur une dernière question qui, depuis les premières lignes de ce travail, a guidé nos recherches : la question de savoir quel rôle joue l’attribution de la compréhension dans l’attribution d’états mentaux comme les attitudes propositionnelles. Ainsi, les prochaines lignes nous serviront à la fois à revenir sur le chemin que nous avons parcouru et à clarifier la relation entre l’attribution de la capacité à comprendre le langage et l’attribution d’états mentaux aux machines dotées d’IA. Pour ce faire, nous présenterons comment les principaux résultats de recherche de chacun de nos chapitres (et donc notre définition fonctionnelle de la compréhension) permettent de soutenir l’idée voulant que l’attribution de la capacité à comprendre aux machines dotées d’IA soit plus fondamentale que l’attribution d’états mentaux particuliers parce que l’attribution de la première peut nous pousser à attribuer des états mentaux.

D’entrée de jeu, le tout premier chapitre de ce mémoire a servi à mettre en lumière les notions fondamentales pour bien saisir le contexte de notre recherche et l’appareillage conceptuel important pour s’y introduire. Nous y avons aussi présenté le phénomène particulier qui a attiré notre attention et qui a motivé la rédaction de ce mémoire : l’attribution d’états mentaux et de capacités mentales/cognitives aux machines dotées d’intelligence artificielle. Pour faire un pas de

plus vers une compréhension complète de ce phénomène fascinant, nous avons présenté et expliqué notre choix de nous concentrer sur l'attribution d'une capacité en particulier : la capacité à comprendre le langage. Nous avons choisi de nous pencher sur cette capacité, car elle semble présupposée dans l'attribution de la capacité à avoir des états mentaux : il faut comprendre des propositions pour entretenir des attitudes cognitives à leur égard. De plus, nous avions soutenu que l'attribution d'états mentaux aux machines passe souvent par une expression antérieure d'états mentaux à travers des comportements verbaux, mais aussi que l'attribution de la capacité mentale à comprendre le langage à des machines dotées d'IA soit extrêmement fréquente et intuitive (plus que l'attribution d'états mentaux spécifiques).

Dans le chapitre 2 (*L'analyse syntaxique et la compréhension*), nous avons formulé une analyse de la compréhension fondée dans le traitement de la syntaxe des expressions des langues naturelles. Ce faisant, nous avons argumenté que le traitement adéquat de la syntaxe manifeste une capacité à comprendre des relations entre des mots. Fréquemment, nous avons fait allusion à l'idée voulant que cette capacité d'une machine à montrer qu'elle comprend des relations entre des mots puisse aussi potentiellement provoquer la perception, chez nous interlocutrices et interlocuteurs, d'une capacité à comprendre des relations entre des objets et entités réels (les référents des mots traités). Cela dit, l'attribution de cette capacité à comprendre des relations extralinguistiques présuppose encore une fois l'attribution de la capacité à comprendre des relations linguistiques, car c'est la démonstration du traitement adéquat de ces dernières, par l'IA, qui nous amène ensuite à inférer qu'elle possède aussi une capacité à comprendre des relations entre des choses réelles. En ce sens, la manifestation de la compréhension linguistique peut démontrer de la compréhension extralinguistique, car le langage est un moyen d'affirmation des relations réelles entre les choses du monde. Maintenant, si l'on attribue à une machine la capacité à comprendre non seulement les relations linguistiques entre des mots, mais aussi des relations réelles entre des entités, des objets

et des états, et bien il est normal que la capacité à traiter les relations linguistiques puisse lui permettre de manifester de la compréhension des expressions d'attitudes propositionnelles qui consistent elles-mêmes en des relations : *des attitudes à l'égard de propositions qui représentent la réalité telle qu'elle est ou telle qu'elle pourrait être.* La ligne est effectivement mince entre le fait de considérer qu'une machine peut comprendre la relation exprimée dans l'expression « X (la machine elle-même, par exemple) *a peur que* S (la situation : être éteinte par l'utilisateur) se produise » et le fait de considérer qu'elle peut comprendre qu'il existe une certaine relation réelle entre X (elle-même) et la situation réelle à laquelle S réfère (la situation où elle est éteinte). C'est d'ailleurs la même chose lorsqu'elle identifie correctement, grâce à l'analyse syntaxique, une relation agentielle comme la « demande » présente dans la structure syntaxique d'un énoncé. À ce moment-là, nous pouvons être amenés à considérer qu'elle peut réellement comprendre, non seulement l'expression d'une relation de nature linguistique, mais aussi qu'il existe une relation réelle de ce type entre elle-même et l'utilisateur qui fait l'énoncé qui exprime un acte de demande. Bref, une machine qui manifeste une compréhension des relations linguistiques peut aussi sembler comprendre les relations extralinguistiques auxquelles elles renvoient.

Ensuite, dans le chapitre 3 (*Les représentations sémantiques et la compréhension*), nous nous sommes concentrés sur les notions de représentations sémantiques et de concepts et avons présenté une seconde analyse de la compréhension, cette fois fondée dans le traitement des significations. Nous avons alors soutenu que les machines dotées d'IA puissent manifester de la compréhension en se montrant capables de former et d'entretenir des représentations sémantiques fonctionnellement équivalentes à nos concepts. Deux éléments sont à mentionner à ce propos. D'abord, comme nous l'avons explicitement présenté, les représentations sémantiques adéquates d'un système peuvent lui permettre de réaliser certains actes cognitifs fonctionnellement équivalents aux nôtres comme des inférences, des prises de décisions, des explications

comportementales, ainsi que des actes de discrimination, de comparaison et des catégorisations d'objets linguistiques. En ce sens, il paraît clair que l'attribution de la capacité à comprendre le langage semble plus fondamentale que l'attribution de ces autres capacités mentales, car c'est d'abord la capacité à avoir des représentations adéquates des significations des mots et expressions qui procure aux systèmes les données (les significations) sur lesquelles s'exécutent ces opérations (les actes cognitifs). Et de là suit le deuxième élément à mentionner : ultimement, puisque les mots et expressions peuvent référer à des objets, entités et états (pour les expressions qui réfèrent à des attitudes, par exemple), il est normal que nous soyons aussi portés à considérer que les machines qui manifestent une capacité à comprendre des significations comprennent aussi ce à quoi elles renvoient dans le monde. De là, en ce qui a trait aux attributions d'attitudes propositionnelles aux machines dotées d'IA, il semble effectivement que de se montrer capable de comprendre les concepts qui réfèrent aux attitudes, ainsi que les significations des mots qui réfèrent aux porteurs de ces attitudes, puis les significations des propositions qui sont objets d'attitudes semble suffisant pour que l'on considère que la machine peut entretenir des états fonctionnellement équivalents aux états mentaux que nous exprimons de la même manière. Si une machine peut montrer qu'elle a une bonne représentation de la signification de la croyance (équivalente à notre concept de croyance), par exemple, et donc qu'elle comprend ce qu'implique d'exprimer et d'entretenir une croyance (comme les inférences réalisables à partir d'elle, les engagements ontologiques qui viennent avec elle, etc.), en plus de montrer qu'elle se représente adéquatement le porteur de la croyance (elle-même) et en quoi consiste l'objet de sa croyance, il est normal que nous soyons portés à considérer qu'elle peut avoir un état mental comme la croyance.

Enfin, dans le quatrième chapitre *(Le langage pragmatique et les intentions de communication*), nous avons formulé une dernière analyse de la compréhension fondée dans le traitement du langage pragmatique en mettant l'emphase sur la notion d'intention de

communication. Ainsi, nous avons mis en lumière l'importance de la capacité à exprimer des intentions de communication et à en identifier dans le discours des autres pour montrer que l'on comprend un langage. Ici, le lien entre l'attribution de la capacité à comprendre le langage et l'attribution d'états mentaux à l'IA est clair et net : les intentions de communication sont analysables sous la forme d'attitudes propositionnelles. Lorsqu'on comprend qu'une personne réalise une fonction du langage en suivant une règle d'usage donnée, on comprend qu'il *veut que* la fonction soit exécutée. Dans le même sens, l'analyse d'une assertion est en partie l'analyse d'une *croyance que* quelque chose est le cas. L'analyse d'une promesse contient *l'intention que* la promesse soit tenue. De plus, lorsqu'on comprend de l'information implicite, on comprend que l'autre *implique que* (*imply that*) telle ou telle chose. Suivant ces idées, en considérant qu'une machine peut exprimer et comprendre des intentions de communication, il est normal d'être amenés à considérer qu'elle peut avoir et comprendre des états comme ceux exprimés par les attitudes propositionnelles, car elles sont les outils par excellence que nous utilisons lorsque nous inférons et décrivons les intentions de communications d'autrui.

Suivant ce qui vient d'être présenté, le principal résultat de ce mémoire est une définition fonctionnelle de la compréhension qui rend compte de notre tendance à attribuer cette capacité mentale aux machines dotées d'IA, mais aussi du rôle de la capacité à traiter les langues naturelles dans l'attribution d'états mentaux à ces machines. Elle peut être résumée brièvement ainsi : une machine/un système a une compréhension fonctionnellement équivalente à la nôtre d'un énoncé reçu ou formulé si et seulement si la machine/le système se montre capable d'identifier adéquatement et de respecter les relations entre les mots et groupes de mots contenus dans l'énoncé,

ainsi que leurs directions respectives OU[72] la machine/le système se montre capable de former et d'entretenir des représentations sémantiques des mots contenus dans l'énoncé qui représentent la réalité adéquatement (par rapport aux concepts auxquels réfèrent, pour nous, ces mots) et qui lui permettent de réaliser les mêmes fonctions cognitives OU la machine/le système se montre sensible aux intentions de communication qui influencent la signification de l'énoncé en s'adaptant aux contextes conversationnels, en se montrant capable de respecter les règles d'usages qui accompagnent l'énoncé et en traitant (ou réalisant) adéquatement les actes de langage qui sont constitutifs de l'énoncé.

La définition proposée est englobante, du fait qu'elle s'intéresse aux trois dimensions du langage. En ce sens, elle est constituée de conditions nécessaires à l'attribution de la capacité à comprendre le langage à des agents artificiels. Maintenant, est-ce que ces conditions sont suffisantes ? Il est clair que les différents contextes conversationnels qui ont lieu dans la vie quotidienne influencent ce qu'il est suffisant de savoir faire et de pouvoir faire avec le langage pour montrer qu'on le comprend. Nous sommes d'avis que les exemples de situations concrètes d'interactions humains-machines présentées dans ce mémoire témoignent de la complexité de cette question. Néanmoins, nous avons tâché de montrer les conditions qui devaient être remplies pour que nous soyons portés à percevoir de la compréhension dans les situations que nous avons étudiées. De là, notre définition présente des conditions que nous pouvons juger comme nécessaires et suffisantes dans le cadre des contextes spécifiques ici étudiés.

[72] Le terme « OU » est, dans la définition, utilisé comme en tant que disjonction inclusive, au sens où une machine/un système manifeste de la compréhension en se montrant capable de réaliser un seul niveau ou plusieurs niveaux de compréhension décrits dans la définition (la machine/le système fait preuve de maîtrise syntaxique OU de maîtrise sémantique OU de maîtrise pragmatique).

Nous souhaitons maintenant utiliser les dernières lignes de ce mémoire pour conclure cette réflexion en abordant brièvement une dernière problématique qui se présente devant nous comme l'éléphant dans la pièce : est-ce qu'une machine peut vraiment comprendre ? Notre définition fonctionnelle de la compréhension cherche à montrer que certaines machines dotées d'IA peuvent traiter le langage de façons équivalentes aux êtres humains et que l'exécution de ces traitements du langage peut manifester, chez la machine, une capacité à comprendre. En ce sens, nous attribuons cette capacité à une machine lorsqu'elle sait adéquatement traiter le langage : lorsqu'elle *manifeste* de la compréhension. Suivant ces idées, notre définition fonctionnelle de la compréhension ne soutient pas que les machines dotées d'IA comprennent comme le peut un être humain, mais décrit plutôt comment se manifeste la compréhension linguistique chez ces agents artificiels. La définition montre ainsi quels traitements sont fonctionnellement équivalents à la compréhension naturelle et propose que, puisque ces agents peuvent réaliser ces traitements, il soit normal que l'on considère, dans certains contextes, qu'elles comprennent le langage.

Enfin, puisque la définition fonctionnelle de la compréhension présentée dans ce mémoire se veut un outil permettant de mieux comprendre comment se manifeste la compréhension linguistique et comment nous la percevons chez les agents artificiels, nous sommes d'avis que sa pertinence est à la fois théorique et pratique, du fait qu'elle peut alimenter nos réflexions morales, sociales et politiques qui ont trait aux domaines d'usages de ces nouvelles technologies autonomes. L'effervescence du développement de l'intelligence artificielle provoque effectivement des vagues d'engouement par rapport à ses usages potentiels. Au moment d'écrire ces lignes, on parle constamment de l'intégrer à nos milieux de vie et de travail, pour remplir des tâches liées au service à la clientèle, pour rédiger des articles de journaux ou même pour fournir de la compagnie aux personnes vivant de l'isolement social. Le choix d'intégrer ou non l'IA dans les sphères professionnelles et intimes de nos vies mérite effectivement d'être bien réfléchi et une réflexion

rationnelle à ce propos passe nécessairement par une étude sérieuse des perceptions que nous avons, au quotidien, des capacités et des limites de ces systèmes « intelligents ». Nous espérons que la définition de la compréhension attribuable aux machines dotées d'IA ici formulée contribuera de façon positive à cette réflexion.

# LISTE DES RÉFÉRENCES


AIRENTI, Gabriella. « The development of anthropomorphism in interaction: intersubjectivity, imagination, and theory of mind », *Frontiers in Psychology*, vol. 9, novembre 2018, pp. 1-13.

APPEL, Jana, VON DER PÜTTEN, Astrid, KRÄMER, Nicole C. et GRATCH, Jonathan. « Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction », *Advances in Human-Computer Interaction*, août 2012, pp. 1-10.

AZAMFIREI Razvan, KUDCHADKAR, Sapna R. et FACKLER, James. « Large language models and the perils of their hallucinations », *Critical Care*, vol. 27, no. 120, mars 2023.

BAKER, Lynne R. *Explaining attitudes: a practical approach to the mind*, Cambridge, Cambridge University Press, 1995, 262 p.

BAR-ELLI, Gilead. « The context principle », dans *The sense of reference : intentionality in Frege*, Berlin, De Gruyter, 1996, pp. 108-132.

BENDER, Emily M., GEBRU, Timnit, MCMILLAN-MAJOR, Angelina et SHMITCHELL, Shmargaret. « On the dangers of stochastic parrots: can language models be too big? », dans le cadre de la conférence *ACM conference on fairness, accountability, and transparency*, événement virtuel, Canada, mars 2021, pp. 610-623.

BENETEAU, Erin, RICHARDS, Olivia K., ZHANG, Mingrui, KIENTZ, Julie A., YIP, Jason C. et HINIKER, Alexis. « Communication breakdowns between families and Alexa », dans le cadre de la conférence *Conference on human factors in computing systems*, Glasgow, mai 2019.

BERG, Markus, RAAB-DÜSTERHÖFT, Antje et THALHEIM, Bernhard. « Adaptation in speech dialogues : possibilities to make human-computer-interaction more natural », dans le cadre de la conférence *Fifth baltic conference Human-Computer-Interaction*, janvier 2011.

BEYSSADE, Claire. « Les implicatures conversationnelles », dans *Sous le sens: pour une sémantique multidimensionnelle*, Saint-Denis, Presses universitaires de Vincennes, 2017, pp. 45-74.

BICHO, Estela, ERLHAGEN, Wolfram, LOURO Luis, COSTA E SILVA, Eliana, SILVA, Rui et HIPOLITO, Nzoji. « Joint action, collaboration and communication », dans *New frontiers in human-robot interaction*, Advances in interaction studies, Dautenhahn, Kerstin et Joe Saunders (éd.), Amsterdam, John Benjamins Publishing Company, 2011, pp. 133-277.

BŁACHNIO, Wojciech. « Is the artificial intelligent? A perspective on AI-based natural language processors », *New horizons in English studies*, no. 4, 2019, pp. 19-34.

BLOCK, Ned. « Advertisement for a semantics for psychology », *Midwest studies in philosophy,* vol. X, 1986, pp. 615-678.

BRANDOM, Robert. « Inferentialism and some of its challenges », *Philosophy and Phenomenological Research*, vol. 74, no. 3, mai 2007, pp. 651–676.

BUTZ, Martin V. « Towards Strong AI », *Künstliche Intelligenz*, vol. 35, février 2021, pp. 91–101.

CAPPELEN, Herman et DEVER, Josh. *Making AI intelligible*, http://fdslive.oup.com/www.oup.com/academic/pdf/openaccess/9780192894724.pdf, 2021.

CAREY, Susan. « Conclusion II: Implications for a theory of concepts », dans *The Origin of Concepts*, Oxford, Oxford University Press, 2009, pp. 487- 538.

CARNAP, Rudolf. *Meaning and necessity*, https://archive.org/details/meaningandnecess033225mbp/page/n9/mode/2up, 1947.

CHOMSKY, Noam. *Aspects of the theory of syntax*, https://apps.dtic.mil/sti/pdfs/AD0616323.pdf, 1965.

CHOMSKY, Noam. *Le langage et la pensée*, traduction de Louis-Jean Calvet et Claude Bourgeois, Paris, Payot, (nouvelle édition augmentée) 2009, 325 p.

CHURCHLAND, Paul. « Le matérialisme éliminativiste et les attitudes propositionnelles », dans *Philosophie de l'esprit. Psychologie du sens commun et sciences de l'esprit, vol. 1* Traduction de Pierre Poirier, D. Fisette, P. Poirier, (éd.), Paris, Vrin, 2002, pp. 117-151.

CLARK, Herbert H. « Joint activities », dans *Using language*, Cambridge, Cambridge University Press, 1996, pp. 29-58.

CONWELL, Colin, ULLMAN, Tomer D. « Testing relational understanding in text-guided image generation », *ArXiv Computer vision and pattern recognition*, juillet 2022, pp. 1-11.

DASGUPTA, Ishita, GUO, Demi, GERSHMAN, Samuel J. et GOODMAN, Noah D. « Analyzing machine-learned representations: a natural language case study », *Cognitive Science*, vol. 44, no. 12, novembre 2020, pp. 1-31.

DAVIDSON, Donald. « The logical form of action sentences », dans *The essential Davidson*, Oxford, Oxford University Press, 2006, pp. 37–71.

DAVIS, Randall, SHROBE, Howard, SZOLOVITS, Peter. « What is a knowledge representation? », *AI Magazine*, vol. 14, no. 1, printemps 1993, pp. 17-33.

DENNETT, Daniel C. « De l'existence des *patterns* », dans *Philosophie de l'esprit. Psychologie du sens commun et sciences de l'esprit, vol. 1*. Traduction de Dominique Boucher, D. Fisette, P. Poirier (éd.), Paris, Vrin, 2002, pp. 153-193.

DENNETT, Daniel C. *The Intentional Stance*, Cambridge, MA, MIT Press, 1987, 388 p.

DREYFUS, Hubert L. *What computers still can't do : a critique of artificial reason*, Cambridge, MA, MIT Press, 1992, 354 p.

ELUGARDO, Reinaldo. « Fodor's inexplicitness argument », dans *The compositionality of meaning and content: vol. I, foundational issues*, M. Werning, E. Machery, G. Schurz, (éd.), Frankfurt, De Gruyter, 2005, pp. 59–85.

EPLEY, Nicholas, WAYTZ, Adam et CACIOPPO, John T. « On seeing human: a three-factor theory of anthropomorphism ». *Psychological Review*, vol. 114 no. 4, octobre 2007, pp. 864-86.

FEIGL, Herbert. *The "mental" and the "physical"*, https://hdl.handle.net/11299/184614, 1958.

FERREIRA, Fernanda, BAILEY, Karl G.D. et FERRARO, Vittoria. « Good-enough representations in language comprehension », *Current directions in psychological science*, vol. 11, no. 1, février 2002, pp. 11-15.

FEUILLARD, Colette. « À propos des fonctions syntaxiques », *La linguistique*, vol. 45, no. 2, 2009, pp. 93-114.

FODOR, Jerry A. *Concepts : where cognitive science went wrong*, New York, Oxford University Press, 1998, 174 p.

FODOR, Jerry A. « Fodor's Guide to mental representation », dans *Readings in Philosophy and Cognitive Science*, A. I. Goldman (éd.), Cambridge, MA, MIT Press, 1993, pp. 271-296.

FODOR, Jerry A. « A modal argument for narrow content », *The Journal of Philosophy*, vol. 88, no. 1, janvier 1991, pp. 5-26.

FODOR, Jerry A. *The Language of Thought*, New York, Thomas Y. Crowell, 1975, 214 p.

FODOR, Jerry A. *Psychosemantics : The problem of meaning in the philosophy of mind*. Coll. Explorations in cognitive science, 2, Cambridge, MA, MIT Press, 1987, 171 p.

FODOR, Jerry A. « Propositional attitudes », *The Monist*, vol. 61, no. 4, octobre 1978, pp. 501-523.

FODOR, Jerry A. et LEPORE, Ernest. *The compositionality papers*, Oxford, Clarendon Press, 2002, 212 p.

FORBES, Maxwell et CHOI, Yejin. « Verb physics: relative physical knowledge of actions and objects », *ArXiv Computation and language*, juillet 2017.

FREDE, Michael. « The origins of traditional grammar », dans *Historical and philosophical dimensions of logic, methodology and philosophy of science*, R.E. Butts et J. Hintikka, (éd), The University of Western Ontario series in philosophy of science, vol. 12, Dordrecht, Springer, 1975, pp. 51-79.

FREGE, Gottlob. « Letter to Jourdain », dans *Philosophical and mathematical correspondence*, traduction de Hans Kaal, Gottfried Gabriel (éd.), Chicago, Chicago University Press, 1980, pp. 78-80.

FREGE, Gottlob. « On sense and reference », dans *Meaning and Reference*, A.W. Moore (éd), Oxford, Oxford University Press, 1993, pp. 22-41.

GAYRAL, Françoise, KAYSER, Daniel et LÉVY, François. « Challenging the principle of compositionality in interpreting natural language texts », dans *The compositionality of meaning and content: vol. II*, *applications to linguistics, psychology and neuroscience*, E. Machery, M. Werning et G. Schurz, (éd.), Frankfurt, De Gruyter, 2005, pp. 83-105.

GEORGI, Geoff. « Propositions, representation, and truth », *Synthese*, vol. 196, mars 2019, pp. 1019-1043.

GILDEA, Daniel et JURAFSKY, Daniel. « Automatic labeling of semantic roles », *Computational Linguistics*, vol. 28, no. 3, 2002, pp. 245–288.

GLANZBERG, Michael. « Truth », *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (éd.), https://plato.stanford.edu/archives/sum2021/entries/truth/, été 2021.

GRAND, Gabriel, BLANK, Idan Asher, PEREIRA, Francisco, FEDORENKO, Evelina. « Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings », *ArXiv Computation and language*, mars 2018, pp. 1-18.

GRICE, Herbert Paul. « Logic and conversation », dans *Studies in the Way of Words*, Cambridge, MA, Harvard University Press, 1989, pp. 22-40.

GRICE, Herbert Paul. « Utterer's meaning, sentence-meaning, and word-meaning », *Foundations of language*, vol. 4, no. 3, août 1968, pp. 225-242.

HARMAN, Gilbert. « Deep structure as logical form », *Synthese*, vol. 21, no. 3/4, octobre 1970, pp. 275-297.

HARNAD, Stevan. « The symbol grounding problem », *Physica D*, vol. 42, 1990, pp. 335-346.

HARNAD, Stevan. « Other bodies, other minds: A machine incarnation of an old philosophical problem », *Minds and Machines*, vol. 1, février 1991, pp. 43-54.

HARRIS, Paul L. *The work of the imagination*, Oxford, Blackwell, 2000, 236 p.

HARRIS, Paul L. *L'imagination chez l'enfant : son rôle dans le développement cognitif et affectif*, traduction de P. Torracinta, Paris, Retz, 2007, 222 p.

HERMBERG, Kevin. *Husserl's phenomenology : knowledge, objectivity and others*, London, Continuum, 2006, 145 p.

HIGGINBOTHAM, James. « A note on phrase-markers », *Revue québécoise de linguistique*, vol 13, no. 1, 1983, pp. 147-166.

HIGGINS, Edward Tory. « Knowledge activation: Accessibility, applicability, and salience », dans *Social psychology: handbook of basic principles*, E. T. Higgins et A. W. Kruglanski, (éd.), New York, The Guilford Press, 1996, pp. 133-168.

HINDRIKS, Frank. « Restructuring Searle's making the social world », *Philosophy of the Social Sciences*, vol. 43, no. 3, 2013, pp. 373-389.

HOROWITZ, Alexandra C. et BEKOFF, Marc. « Naturalizing anthropomorphism: behavioral prompts to our humanizing of animals », *Anthrozoös*, vol. 20, no.1, 2007, pp. 23-35.

HORSTMANN Aike C., BOCK Nikolai, LINHUBER Eva, SZCZUKA, Jessica M., STRAßMANN Carolin et KRÄMER, Nicole C. « Do a Robot's Social Skills and Its Objection Discourage Interactants from Switching the Robot Off? », *Plos One*, vol. 13, no. 7, juillet 2018, pp. 1-25.

IBM CLOUD EDUCATION. « What are recurrent neural networks? », (page consultée le 21 août 2023), https://www.ibm.com/topics/recurrent-neural-networks.

IBM CLOUD EDUCATION. « What are neural networks? », (page consultée le 21 août 2023), https://www.ibm.com/uk-en/cloud/learn/neural-networks.

IRVINE, Andrew D. « Bertrand Russell's logic », dans *Handbook of the history of logic*, *vol. 5 : logic from Russell to Church*, D. M. Gabbay et J. Woods, (éd.), 2009, pp. 1-28.

JACKMAN, Henry. « Holism, relevance and thought content », dans le cadre de la conférence annuelle *Ohio Philosophical Association*, 1999, 140-151.

KAPLAN, Ronald M. et BRESNAN, Joan. « Lexical functional grammar a formal system for grammatical representation », dans *Formal issues in lexical functional grammar,* M. Dalrymple, R. M. Kaplan, J. T. Maxwell III et A. Zaenen, (éd.), 1995, pp. 1-102.

KAPLAN, David. « Afterthoughts », dans *Themes from Kaplan*, J. Almog, J. Perry et H. Wettstein, (éd.), Oxford, Oxford University Press, 1989, pp. 565-614.

KATZ, Jerrold J. et POSTAL, Paul. *An integrated theory of linguistic description*, Cambridge, MA, MIT Press, 1964, 178 p.

KATZ, Jerrold J. « La théorie du langage », dans *La philosophie du langage*, Coll. Bibliothèque Scientifique, J. J. Katz (éd.), Paris, Payot, 1971, pp. 87-157.

KATZ, Jerrold J. « Chomsky on Meaning », *Language*, vol. 56, no. 1, mars 1980, pp. 1-41.

KIM, Jaegwon. *Philosophie de l'esprit*, Coll. Philosophie, préface de P. Engel, traduit par D. Michel-Pajus, M. Mulcey (dir.) et C. Théret, Paris, Ithaque, 2008a, 371 p.

KIM, Jaegwon. « The supervenience argument motivated, clarified, and defended », dans *Physicalism, or Something near Enough*, Princeton, Princeton University Press, 2008b, pp. 32-69.

KRACHT, Marcus. « Referent systems and relational grammar », *Journal of logic, language, and information*, vol. 11, no. 2, printemps 2002, pp. 251-286.

LANDGREBE, Jobst et SMITH, Barry. « There is no artificial general intelligence », *ArXiv Computation and language*, novembre 2019, pp. 1-58.

LEE, Seungcheol Austin et LIANG, Yuhua. « Robotic foot-in-the-door: using sequential-request persuasive strategies in human-robot interaction », *Computers in human behavior*, vol. 90, janvier 2019, pp. 351-356.

LEE, Sangwon, LEE, Naeun et SAH, Young June. « Perceiving a mind in a chatbot: effect of mind perception and social cues on co-presence, closeness, and intention to use », *International Journal of Human-Computer Interaction*, vol. 36, no. 10, décembre 2020, pp. 930-940.

LEVESQUE, Hector J., DAVIS, Ernest et MORGENSTERN, Leora. « The Winograd Schema Challenge », dans le cadre de la conférence *Thirteenth international conference on principles of knowledge representation and reasoning (AAAI)*, 2012, pp. 552-561.

LI, Huao, NI, Tianwei, AGRAWAL, Siddharth, JIA, Fan, RAJA, Suhas, GUI, Yikang, HUGHES, Dana, LEWIS, Michael et SYCARA, Katia. « Individualized mutual adaptation in human-agent teams », *IEEE Transactions on human-machine systems*, vol. 51, no. 6, décembre 2021, pp. 706-714.

LIANG, Claire, PROFT, Julia, ANDERSEN, Erik et KNEPPER, Ross A. « Implicit communication of actionable information in human-AI teams », dans le cadre de la conference *Conference on human factors in computing systems*, Glasgow, mai 2019, papier 95.

LINSKY, Bernard et PELLETIER, Jeffry. « What is Frege's theory of descriptions? », dans *On Denoting: 1905-2005*, G. Imaguire et B. Linsky, (éd.), München, Philosophia, 2005, pp. 195-250.

LOCKWOOD, David G. « Syntax and semology » dans *Syntactic analysis and description : a constructional approach*, Coll. Open Linguistics Series, Londres, Continuum, 2002, pp. 310-328.

LUDWIG, Kirk. « Logical form », dans *Routledge Companion to philosophy of language*, G. Russell, D. Graff, (éd.), Londres, Routledge, 2012, pp. 29-41.

MANNING, Christopher D., CLARK, Kevin, HEWITT, John, KHANDELWAL, Urvashi, LEVY, Omer. « Emergent linguistic structure in artificial neural networks trained by self-supervision », *PNAS*, vol. 117, no. 48, décembre 2020, pp. 30046–30054.

MARCHESI, Serena, GHIGLINO, Davide, CIARDO, Francesca, PEREZ-OSORIO, Jairo, BAYKARA, Ebru et WYKOWSKA, Agnieszka. « Do we adopt the intentional stance toward humanoid robots? », *Frontiers in Psychology*, vol. 10, mars 2019, pp. 1-13.

MARTIN, Dorothea Ulrike, PERRY, Conrad, MACINTYRE, Madeline Isabel, VARCOE, Luisa, PEDELL, Sonja et KAUFMAN, Jordy. « Investigating the nature of children's altruism using a social humanoid robot », *Computers in human behavior*, vol. 104, mars 2020, article 106149.

MARTY, Anton. « Sur l'élaboration des moyens d'expression sous la forme d'un langage articulé », dans *Sur l'origine du langage*, Paris, Hermann, 2017, pp. 87-134.

MAVRINA, Lina, SZCZUKA, Jessica, STRATHMANN, Clara, BOHNENKAMP, Lisa Michelle, KRÄMER, Nicole et KOPP, Stefan. « "Alexa, you're really stupid": a longitudinal field study on communication breakdowns between family members and a voice assistant », *Frontiers in Computer Science*, vol. 4, janvier 2022, article 791704.

MILNE, Perte. « Frege's context principle », *Mind*, vol. 95, no. 380, octobre 1986, pp. 491-495.

MÜLLER, Vincent C. « Is there a future for AI without representation? », *Minds and Machines*, vol. 17, no. 1, 2007, pp. 101-115.

PAGIN, Peter. « Meaning holism », dans *The Oxford Handbook to the philosophy of language*, E. Lepore et B. Smith, (éd.), *The Oxford Handbook to the philosophy of language*, Oxford, Oxford University Press, 2008, pp. 213-232.

PAGIN, Peter et WESTERSTÅHL, Dag. « Compositionality I: definitions and variants », *Philosophy compass*, vol. 5, no. 3, mars 2010, pp. 250-264.

PARTEE, Barbara. « The garden of Eden period for deep structure and semantics », dans *50 years later: reflections on Chomsky's Aspects*, Á. J. Gallego et D. Ott, (éd.), Cambridge, MA, MIT Press, 2015, pp. 187-198.

PAYNE, Stephen J. « Mental models in human-computer interaction », dans *Human-computer interaction*, Coll. Human Factors and Ergonomics, A. Sears et J. A. Jacko, (éd.), New York, Taylor & Francis Group, 2009, pp. 39-52.

PERROTTA, Carlo, SELWYN, Neil, EWIN, Carrie. « Artificial intelligence and the affective labour of understanding: the intimate moderation of a language model », *New Media & Society*, février 2022, pp. 1-25.

PITT, David. « Mental representation », *The Stanford Encyclopedia of Philosophy,* E. N. Zalta, U. Nodelman, (éd.), https://plato.stanford.edu/archives/fall2022/entries/mental-representation/, automne 2022.

PRAKASH, Pandey Om. « Grammar: a historical survey », *IOSR Journal of humanities and social science*, vol. 10, no. 6, mai et juin 2013, pp. 60-62.

PULLUM, Geoffrey K. et HUDDLESTON, Rodney. « Prepositions and preposition phrases », dans *The Cambridge Grammar of the English language*, Cambridge, Cambridge University Press, 2002, pp. 597–661.

PUTNAM, Hilary. *Mind, Language, and Reality*, Coll. His Philosophical Papers, vol. 2, New York, Cambridge University Press, 1975, 476 p.

RAMESH, Aditya, PAVLOV, Mikhail, GOH, Gabriel, GRAY, Scott, VOSS, Chelsea, RADFORD, Alec, CHEN, Mark et SUTSKEVER, Ilya. « Zero-shot text-to-image generation », *ArXiv Computer vision and pattern recognition*, février 2021.

RICHARD, Mark. « What are propositions », *Canadian journal of philosophy*, vol. 43, nos. 5-6, 2013, pp. 702-719.

ROSCH, Eleanor et MERVIS, Carolyn B. « Family resemblances: studies in the internal structure of categories », *Cognitive Psychology*, vol. 7, no. 4, octobre 1975, pp. 573-605.

RUSSELL, Bertrand. « Introduction », dans *Tractatus Logico-Philosophicus*, https://www.gutenberg.org/files/5740/5740-pdf.pdf, 2010 (édition originale : 1922).

RUSSELL, Bertrand. « Sentences, syntax, and parts of speech », dans *The basic writings of Bertrand Russell*, L. E. Denonn, R. E. Egner et J. G. Slater, (éd.), Londres, Routledge, 2009a, pp. 90-102.

RUSSELL, Bertrand. *Problems of philosophy*, https://www.gutenberg.org/files/5827/5827-h/5827-h.htm, 2009b (édition originale : 1912).

RUSSELL, Bertrand. « La compréhension des propositions », dans *Théorie de la connaissance : le manuscrit de 1913,* E. R. Eames et K. Blackwell, (éd.), traduit par J-M Roy, Paris, Vrin, 2002, pp. 137-153.

RUSSELL, Bertrand. « On propositions: what they are and how they mean », *Aristotelian Society supplementary*, vol. 2, no. 1, juin 1919, pp. 1-43.

RUSSELL, Bertrand. « On denoting », *Mind*, vol. 14, no. 56, octobre 1905, pp. 479-493.

SEARLE, John R. *Speech acts : an essay in the the philosophy of language*, London, Cambridge University Press, 1969, 203 p.

SEARLE, John R. « A classification of illocutionary acts », *Language in Society*, vol. 5, no. 1, avril 1976, pp. 1-23.

SEARLE, John R. « Minds, brains and programs », *Behavioral and brain sciences*, vol. 3, no. 3, 1980, pp. 417-457.

SEARLE, John R. « The purpose of this book », dans *Making the social world : the structure of human civilization*, Oxford, Oxford University Press, 2010, pp. 3-24.

SEVERSON, Rachel L. et WOODARD, Shailee R. « Imagining others' minds: the positive relation between children's role play and anthropomorphism », *Frontiers in Psychology*, vol. 13, novembre 2018, article 2140.

SHARMA, Piyush., DING, Nan, GOODMAN, Sebastian et SORICUT, Radu. « Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning », dans le cadre de la conférence *56th Annual Meeting of the Association for computational linguistics*, Melbourne, juillet 2018, pp. 2556–2565.

SMITH, Barry. « Beyond concepts: ontology as reality representation », dans le cadre de la conférence: *FOIS International conference on formal ontology and information systems*, A. C. Varzi et L. Vieu, (éd.),Turin, novembre 2004, pp. 1-12.

SOWA, John F. « Semantic networks », dans *Encyclopedia of artificial intelligence*, S. C. Shapiro (éd.), New York, Wiley, 1992, 1724 p.

SPATOLA, Nicolas, MARCHESI Serena et WYKOWSKA, Agnieszka. « The intentional stance test-2: how to measure the tendency to adopt intentional stance towards robots », *Frontiers in robotics and Ai*, vol. 8, octobre 2021, article 666586.

STALNAKER, Robert. « Common ground », *Linguistics and Philosophy*, vol. 25, no.5, 2002, pp. 701-721.

STRAWSON, Peter Frederick. « On referring ». *Mind*, vol. 59, no. 235, juillet 1950, pp. 320-344.

SZABÓ, Zoltán Gendler. « Compositionality », *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, U. Nodelman, (éd.), https://plato.stanford.edu/archives/fall2022/entries/compositionality/, automne 2022.

TASCHEK, William W. « On Belief Content and That-Clauses », *Mind & Language*, vol. 10, no. 3, septembre 1995, pp. 274-298.

TURING, Alan Mathison. « Computing machinery and intelligence », *Mind*, vol. 59, no. 236, octobre 1950, pp. 433-460.

VAN DER VELDE, Frank. « Neural architectures of compositionality », dans *The compositionality of meaning and content: vol. II*, *applications to linguistics, psychology and neuroscience*, E. Machery, M. Werning et G. Schurz, (éd.), Frankfurt, De Gruyter, 2005, pp. 265-281.

WANG, Jixuan, WEI, Kai, RADFAR, Martin, ZHANG, Weiwei et CHUNG, Clement. « Encoding syntactic knowledge in transformer encoder for intent detection and slot filling », dans le cadre de la conférence *The thirty-fifth AAAI conference on artificial intelligence*, 2021, pp. 13943-13951.

WEBSTER, Jonathan J. « An introduction to Continuum Companion to systemic functional linguistics », dans *Bloomsbury Companion to systemic functional linguistics*, Coll. Continuum Companions, M.A.K. Halliday et J. J. Webster, (éd.), Londres, Bloomsbury, 2009, pp. 1-11.

WILLIAMS, Edwin. « Grammatical relations », *Linguistic Inquiry*, vol. 15, no. 4, automne 1984, pp. 639-673.

WITTGENSTEIN, Ludwig. *Tractatus Logico-Philosophicus*, https://www.gutenberg.org/files/5740/5740-pdf.pdf, 2010 (édition originale de 1922).

WITTGENSTEIN, Ludwig. *Recherches Philosophiques*, Coll. Bibliothèque de Philosophie, traduit par F. Dastur, M. Élie, J. L. Gautero, D. Janicaud et É. Rigal, avant-propos de É. Rigal, Paris, Gallimard, 2004, 367 p.

WITTGENSTEIN, Ludwig. *Notebooks 1914-1916*, G. H. von Wright et G. E. M. Anscombe, (éd.), traduit par G. E. M. Anscombe, New York, Harper and Brothers Publishers, 1961, 238 p.

YANG, Qian, STEINFELD, Aaron, ROSÉ, Carolyn et ZIMMERMAN, John. « Re-examining whether, why, and how human-AI interaction is uniquely difficult to design », dans le cadre de la conférence *Conference on human factors in computing systems*, Honolulu, avril 2020, papier 174.

YAO, Zijun, SUN, Yifan, DING, Weicong, RAO, Nikhil et XIONG, Hui. « Dynamic word embeddings for evolving semantic discovery », *ArXiv Computation and language*, février 2018.

YU, Jiahui, XU, Yuanzhong, YU KOH, Jing, LUONG, Thang, BAID, Gunjan, WANG, Zirui, VASUDEVAN, Vijay, KU, Alexander, YANG, Yinfei, KARAGOL AYAN, Burcu, HUTCHINSON, Ben, HAN, Wei, PAREKH, Zarana, LI, Xin, ZHANG, Han, BALDRIDGE, Jason et WU, Yonghui. « Scaling autoregressive models for content-rich

text-to-image generation », *ArXiv Computer vision and pattern recognition*, juin 2022, pp. 1-49.

ZŁOTOWSKI, Jakub. « Mind attribution: from simple shapes to social agents », dans le cadre de la conférence *25th IEEE International Symposium in robot and human interactive communication (Ro-Man)*, New York, août 2016, pp. 916-917.