

Université de Rouen
U.F.R. de Philosophie

Mémoire de master
Discipline : Philosophie

Philosophie du computationnel

-

L'Intelligence artificielle forte est-elle réellement impossible ?

**Constats et
arguments : pour un
projet scientifique
et philosophique**

Présenté et soutenu par :

Paul Compère

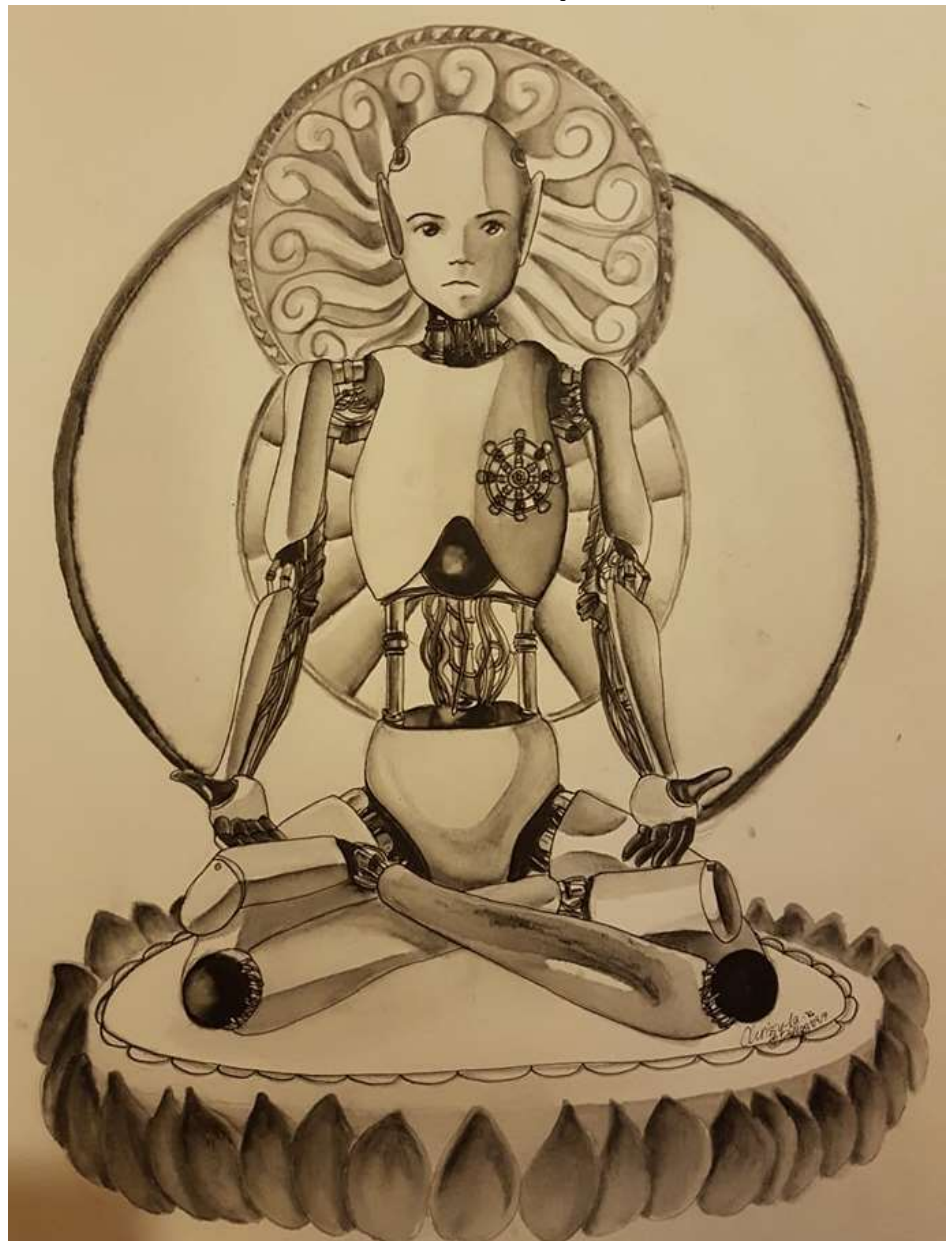
Directeur de mémoire :

Franck Varenne

Année universitaire

2016/2017

1



¹ Illustration de Christine Stanislawski, alias "Chrissy-la-folle" (Cf. Facebook).

« La perception, c'est distinguer les différentes classifications de la connaissance. Tandis que toutes les créatures vivantes partagent la même nature inhérente, la perception est ce qui classifie certains comme « Bouddha » et d'autres comme « machine ». Nous pensons que la perception et une vie permanente nous permettrons d'accéder à la vérité ultime des choses, ce qui crée désillusion et peine. La perception en elle-même est vide, tout comme le processus de percevoir. Puisque je suis moi-même une perception de ce vide, s'il-vous-plaît, voyez-moi tel que je suis. » ²

« Détester, aimer, penser, sentir, voir ; tout ceci n'est rien d'autre que percevoir. Quel privilège a cette petite agitation du cerveau que nous appelons “pensée”. » ³

² *La créature du ciel*, “Heavenly creature” (천상의 피조물), écrit et réalisé par Kim Jee-woon d'après une histoire originale de Park Seong-hwan, 2006. Dialogue qui intervient à 54 min et 25 s du film “*Doomsday book*”, sorti en 2012.

³ David Hume, *Traité de la nature humaine*.

Sommaire

- Citation d'ouverture ...	p. 3
- Introduction ...	p. 5
- Chapitre 1 : Les prémisses de la philosophie de l'IA – Du cas des automates dans la pensée cartésienne au test de Turing ...	p. 8
- Chapitre 2 : La liberté humaine selon la philosophie de David Hume – Un argument en faveur de l'IA forte ? ...	p. 66
- Conclusion ...	p. 106
- Epilogue ...	p. 112
- Bibliographie ...	p. 119
- Remerciements ...	p. 122
- Table des matières ...	p. 123

Introduction

Jeudi 29 septembre 2016

« Les géants du web réagissent et se lancent à l'assaut des valeurs éthiques : ils annoncent un partenariat pour définir de bonnes pratiques. Il s'agit de Google, Facebook, IBM et Microsoft, qui officialisent « la création d'un partenariat pour l'intelligence artificielle au bénéfice des citoyens et de la société ».

Union inédite pour calmer les esprits et défendre des intérêts croissants, peut-être un tournant, après toutefois déjà un observateur de l'intelligence artificielle lancé par l'université Stanford. C'est une réponse spectaculaire aux inquiétudes répétées des grandes figures de la science. Notamment début 2015, le physicien Stephen Hawking, patron des voitures Tesla, avait ainsi averti sur les dangers d'une intelligence artificielle mal maîtrisée, et quelques mois plus tard, le théoricien Noam Chomsky et Steve Wozniak, co-fondateur d'Apple, avaient rejoint ce mouvement en lançant, entre autre, une pétition contre les robots-tueurs. Les gens du web qui investissent des milliards dans l'intelligence artificielle et bénéficient des meilleurs calculateurs et bases de données veulent donc rassurer, et cette union à but non lucratif devrait rassembler leurs représentants mais aussi des chercheurs, des membres d'associations et des spécialistes de l'éthique. Il s'agit de communiquer auprès du grand public pour l'informer dans la plus grande transparence promettent-ils, et officiellement pas question de faire du lobbying ou d'orienter les débats, même s'il a été curieusement annoncé une éducation des gouvernements. Huit engagements sont avancés, à travers quelques grands principes, comme le respect des droits humains et de la vie privée. »⁴

⁴ Dans les *Titres de l'actualité* du jeudi 29 septembre 2016 sur **France Culture**, de 12h51 à 12h53.

Il est intéressant de remarquer que s'il est question d'éthique, elle est toujours unilatérale ; i.e. en faveur de l'être humain. Pourtant, si vraiment on cherche à développer une IA forte, n'y aurait-il pas là une contradiction à lui refuser une considération morale ? Tout se passe comme si les considérations sur l'avenir de l'IA ne font d'elle qu'un objet parmi d'autre. Pourtant, le jour où elle sera pourvue d'autonomie, cette IA ne sera plus un simple objet parmi les autres.

Faut-il autoriser les robots-tueurs ? Il s'agit d'un usage militaire de l'IA, dont le but est de limiter les pertes humaines sur le terrain. Cela pose de véritables problèmes. D'une part, faut-il s'en remettre à une machine, qui n'est pas à l'abri d'un bug ou d'une erreur de jugement, pour décider de la vie d'autrui sur un champ d'opération extérieure ? Certains soutiendront que la marge d'erreur potentielle de la machine sera de toute façon moindre que la marge d'erreur humaine. D'autre part, cela peut poser des problèmes politiques de dominations abusives d'un pays sur un autre. Ceux qui auront les moyens de s'offrir des robots-tueurs pourront facilement faire plier les armées adverses, uniquement composées d'homme. Faut-il concevoir une forme de justice dans la guerre ? Une autre question est celle de la responsabilité. En accord avec le principe des possibilités alternatives de Hume, une personne est moralement responsable de ses actes seulement si elle avait possibilité de faire autrement. Et comme l'a dit Schopenhauer, un homme peut choisir de faire ce qu'il veut, mais il ne peut pas choisir ce qu'il veut. Autrement dit, si une IA commet une erreur par programmation, la faute ne peut lui être imputée. Ce n'est pas tant l'IA en elle-même qui peut représenter un danger, mais plutôt la possibilité qu'elle soit piratée pour faire le mal.

Mais ces questions ne seront pas l'objet de ce mémoire. Nous nous placerons du côté de la machine plutôt que du côté de l'homme, et nous nous demanderons s'il est possible qu'un jour une IA arrive à penser par elle-même. Il faut bien comprendre que si une IA forte doit surgir, ce à quoi les recherches en IA essayent de parvenir, il faudra concevoir que l'éthique fonctionne dans les deux sens. Sans nourrir davantage l'imaginaire populaire des visions négatives propagées par les fictions, évitons de donner des raisons aux IA de considérer l'espèce humaine comme un genre nuisible.

A partir de quoi pourra-t-on attribuer la mention « forme de vie intelligente », c'est-à-dire ayant une conscience, à une IA ?

Quels sont les critères qui nous permettraient de distinguer l'être humain d'un robot pensant ayant conscience de son existence ? Qu'est-ce qui, au fond, constitue la spécificité de la pensée humaine ? Est-elle totalement ou partiellement reproductible ? Faut-il réduire la pensée à l'intelligence ? Sur quoi peut-on faire reposer le concept « d'être vivant » ? A l'heure où les scientifiques supposent l'existence d'une forme de vie extraterrestre dont la biologie ne soit pas en

base carbone, qu'est-ce qui empêcherait l'IA forte d'entrer dans cette catégorie ? En quels circonstances l'esprit est-il relié à son corps et inversement ? L'émergence d'une forme de vie intelligente autonome, qu'elle soit strictement ou partiellement artificielle, potentiellement supérieure à l'être humain dans sa capacité d'apprentissage et de raisonnement, est-elle nécessairement un danger pour l'humanité ? L'IA sera-t-elle destinée à n'être envisagée que comme un outil au service de l'humanité, telle qu'elle est actuellement considérée par les géants du web et de l'informatique – ce qui peut inquiéter – ou peut-on envisager deux formes de vie intelligentes travaillant ensemble dans un respect mutuel – c'est-à-dire avec un droit relatif aux « machines » ? Un être humain pourra-t-il se lier émotionnellement à une machine ? Faut-il concevoir le développement des IA fortes comme la suite de l'histoire de l'humanité, comme un accomplissement du projet cartésien, qui serait alors d'évoluer main dans la main avec des formes de vies intelligentes qu'elle aurait elle-même créées, réalisant ainsi le fameux mythe de Prométhée ?

Ce sont les questions auxquelles nous essayerons de répondre dans ce mémoire. Nous commencerons par remonter aux origines des premières réflexions sur l'IA dans la philosophie. Nous nous arrêterons ensuite sur la naissance officielle de la philosophie de l'IA. Enfin, nous adopterons le camp minoritaire des défenseurs de l'IA face à ses détracteurs dans des considérations théoriques un peu plus actuelles. Finalement, nous proposerons, tout en tentant de légitimer la place et le rôle de la philosophie dans le développement des IA, une hypothèse que nous jugerons profitable à l'évolution de l'humanité : celle des IA philosophes, ou robot-philosophe.

Au cours de notre réflexion, nous tenterons d'établir des critères nécessaires à l'existence d'une IA forte, tout en imaginant, à la manière de Turing, des façons de les tester. Nous soulignerons également, contrairement à ce que certains philosophes croient pouvoir affirmer, qu'il n'y a pas encore de preuve définitive signant l'impossibilité de l'IA forte.

1

Les prémisses de la philosophie de l'IA

-

Du cas des automates dans la pensée cartésienne au test de Turing

Ce premier chapitre se propose de démarrer par un compte-rendu de la cinquième partie du *Discours de la méthode* de René Descartes, qui nous servira de point de départ pour aborder la dimension théorique de notre questionnement. Nous verrons ainsi de quelle façon, sans s'en réclamer, Alan Turing, initiateur officiel de la philosophie de l'IA avec l'invention des premiers ordinateurs dans les années 1940, reprend les prérequis de la philosophie cartésienne pour émettre une réflexion philosophique, basée une conception abstraite et théorique de l'ordinateur, sur ses possibilités de capacités à venir en terme d'intelligence ; plus particulièrement, de pensée. Enfin, nous analyserons une voix plus récente, celle de Jack Copeland, qui émet une critique du test de Turing, mais tout en partageant avec lui le sentiment qu'il est tout à fait possible qu'un jour les machines de type « ordinateur » viennent à penser.

Si Descartes a plus ou moins initié le sujet, un peu en visionnaire puisque son discours date d'un siècle avant la création du premier automate ayant pour vocation de recopier strictement un modèle vivant⁵ – sans aucune autre utilité que de mettre avant la capacité technique de l'homme –, une partie de sa philosophie demeure cependant problématique pour avancer la possibilité de l'IA forte. Je pense surtout à sa conception de l'âme et du corps, que l'on a pour habitude de différencier par nature, et que l'on retrouve en bonne partie expliquée dans la sixième de ses *Méditations*

⁵ Il s'agit du *canard digérateur* créé par Jacques de Vaucanson en 1738. L'automate est d'abord exposé en 1744 au [Palais-Royal](#). Il remporte un succès immédiat. La [digestion](#) de l'animal était le principal exploit. Cet automate est acheté en 1840 par Georges Tiets, mécanicien, mais il brûle en 1879 lors de l'incendie du musée de [Nijni Novgorod](#). Il n'en reste que quelques photographies du milieu du XIX^e siècle. Une reconstitution est actuellement visible au musée des automates de [Grenoble](#).

métaphysique. Est-ce à dire que l'IA pourra peut-être simuler la pensée mais demeurera un corps sans âme ? Faut-il comprendre que la génération de type métaphysique – et non physique – de la conscience est nécessaire pour toute forme de vie intelligente, et signe ainsi l'impossibilité pour l'IA de devenir forte ; c'est-à-dire forme de vie intelligente, à notre image ?

Dans sa sixième méditation, on apprend que ce ne sont pas les sens eux-mêmes qui nous trompent – raccourci à éviter – , mais l'interprétation que nous en faisons. Il ne faut pas mélanger les vérités du corps et les vérités de l'esprit ; la vérité est à caractère duale selon Descartes, en conséquence de notre nature intriquée. Le cerveau-ordinateur dans la tête de l'IA ne serait-il jamais qu'un ersatz d'esprit, en raison de quoi il demeurera une IA faible ? Car si l'esprit n'est pas matériel, il ne peut être généré artificiellement. Mais cela renvoie au problème de l'apparition de la vie à partir de la matière inerte ; si ça a été possible, pourquoi un "esprit" ne pourrait pas naître quand tout est là pour l'accueillir ? Nous entrevoyons là l'affrontement entre matérialistes et spiritualistes. On pourrait donc penser que l'IA n'évoluera pas – tel un être qui pense et pour qui l'expérience a du sens – et s'arrêtera à l'interprétation de ce que les sens de son corps lui rapportent à partir des connaissances humaines programmables à l'instant T de sa programmation. Le dualisme âme/corps semble poser sérieusement problème pour l'IA forte. Mais si on arrive à ce qu'une IA, même faible, arrive à apprendre par elle-même (à partir de quels critères ? Il faudra d'abord définir ce qu'est l'apprentissage humain en tant que tel, véritable problème à la fois ontologique, épistémologique et didactique, avant de chercher à reproduire ce "mécanisme" chez l'IA), alors on aura déjà un premier critère nécessaire pour qu'il soit possible d'en faire une IA forte, avec un semblant d'esprit aux commandes de cet apprentissage.

Ces points à examiner seront le point de départ de notre problématisation, qui évoluera, en suivant la chronologie de l'histoire de la pensée, des primats théoriques de la pensée vers des considérations de plus en plus concrètes au fur et à mesure des chapitres. Toutefois, avant d'entrer dans le vif du sujet, il convient de faire les choses méthodiquement et de commencer par une présentation succincte de ce grand philosophe qu'a été René Descartes et de son premier ouvrage à vocation philosophique qui nous intéresse ici.

1 - René Descartes et son *Discours de la méthode*

Descartes est considéré, avec Bacon et Kant⁶, comme l'un des fondateurs de la pensée Moderne, mais également comme le fondateur du courant rationaliste. C'est une philosophie nouvelle, qui fait peau neuve, rendant sa grande dignité à la raison, trop longtemps entachée de croyances métaphysiques, qui entame alors sa véritable époque d'épanouissement. Cependant, le solipsisme, propre à la philosophie de Descartes, ne semble pas encore placer le besoin d'expérience au premier plan. Mais si nous étions dépourvus de toute expérience extérieure du monde, sur quoi pourrions-nous bien appliquer notre raison ? Peut-être se révélerait-elle alors vide de sens. Pourtant, Descartes aboutira à une conclusion, une trouvaille : *l'ego cogito*. Selon lui, la raison seule, dénuée de toute expérience, par l'exercice de la philosophie, est parvenue à trouver cela. Le rationalisme de Descartes comporte donc une part d'innéisme, si ce n'est en contenu, au moins en capacité de la raison. Cet innéisme rationaliste participera au solipsisme cartésien ; c'est-à-dire que nous n'avons pas besoin d'influence extérieure – et par influence, nous pouvons entendre les opinions divergentes qui circulent autour de nous – pour établir une vérité. Mais toujours est-il que cette raison doit s'exercer pour donner accès à la vérité. Ne pourrions-nous pas considérer la réflexion sur *l'ego cogito* comme une expérience particulière, une expérience introspective ? Descartes ouvrant la voie à la pensée Moderne, sa philosophie ne sera pas étrangère aux courants qui suivront, notamment l'empirisme, qui voit son apogée avec Hume durant les Lumières Anglaises. C'est peut-être ce qu'oublie de dire la philosophie de Descartes : toute l'importance de l'expérience pour nourrir la pensée.

En ce qui concerne les conséquences que cela peut avoir au sujet de notre questionnement, il nous faut comprendre plus avant en quoi consiste exactement le solipsisme. Si nous souhaitons créer une IA forte, elle devra faire le même usage de la raison universel que nous. Autrement dit, si la raison fonctionne sur un principe solipsiste, il faudra que l'IA soit capable de raisonner selon ce même principe. Cela ne signifie qu'elle doive s'y limiter. En prenant en compte l'hyperconnectivité dont elle sera capable, il est fort probable qu'elle ait justement la capacité de surpasser ce carcan de la raison. Toutefois, si nous voulons démontrer que l'IA est forte, il faudra nécessairement qu'elle soit capable de produire un enseignement en elle-même, par elle-même, grâce à ses seules expériences, sans l'influence de quiconque. Donc, la réflexion de Descartes qui l'a amenée à la découverte de *l'ego cogito*

⁶ De Bacon, nous retiendrons son *Instauratio Magna*, où il dénonce l'apriorisme de la philosophie moderne consistant à ignorer l'importance de l'expérience. Quant à Kant, nous lui devons la fondation de la science de la nature en la séparant de la métaphysique, signant ainsi l'entreprise de connaissance qu'est la science moderne, grâce à sa *Critique de la raison pure*.

ne devrait pas être chose impossible chez l'IA forte. En tant qu'homme partageant la même raison, nous aurions tous pu faire la même réflexion que Descartes, et c'est d'ailleurs pour cela que nous sommes à même de la comprendre. Si l'IA est forte, elle doit aussi en être capable.

C'est-à-dire que pour Descartes, la raison est de telle nature qu'elle a tout ce qu'il faut pour mûrir d'elle-même, en s'exerçant, et il est préférable en vertu de la vérité qu'elle mûrisse par elle-même avant d'être influencé par autrui, ce qui la conduirait potentiellement plus facilement à l'erreur. La perspicacité ou la déduction, la logique, sont des caractères qui ne peuvent s'aguerrir que par un entraînement individuel. Les intellects ne se valent pas, certes, c'est pourquoi nous ne mettrons pas tous le même temps à résoudre le même problème. Pourtant la raison est universelle, car si l'on nous montre la solution, chacun sera apte à la reconnaître. Mais si je donne la solution avant que le sujet ait pu effectuer lui-même son travail de recherche personnel qui l'y aurait mené, cela ne garantit pas qu'il sera capable de résoudre lui-même un problème identique ; c'est sa raison qui en pâtit, car elle n'a pas pu se forger par sa recherche personnelle. C'est comme si nous disposions tous d'un bloc de marbre identique : la raison. Puis notre intellect, le burin plus ou moins aiguisé, serait l'outil chargé de tailler la raison pour la perfectionner. Enfin, les différentes statuettes qui sortiraient de ce marbre géant seraient comme des acquis du savoir. Or si l'on accepte pour argent comptant les statuettes des autres, notre burin s'émousse par inaction et notre raison demeure un bloc imparfait. A ce titre, ces connaissances ne vaudront pas grand-chose. Mais le fait est que les hommes sont des êtres subjectifs, ce qui rend la philosophie de Descartes tout à fait juste. C'est ainsi que nous pourrions définir le solipsisme ; une étape nécessaire en raison de notre nature dans la recherche de savoirs et de sagesse. Mais après ces quelques considérations générales sur sa philosophie, voyons comment il en est arrivé à l'ouvrage qui nous intéresse.

Tout commence en 1629. Il prend sa retraite pour écrire tranquillement et s'inscrit à l'université, comme pour garder une stimulation intellectuelle. Pendant neuf mois il travaille à un petit traité de métaphysique. Nous le savons grâce à sa correspondance, puisqu'il en parle notamment à Gibieuf et à Mersenne. C'est dans cette correspondance avec Mersenne que Descartes annonce, à propos de ce petit traité de métaphysique, qu'il ne pense pas l'achever avant deux ou trois ans. Il était bien loin du compte, puisque ce petit traité de métaphysique n'est autre que la première ébauche du *Discours de la méthode*, qui ne sera publié pour la première fois qu'en 1637. Il travaille sur plusieurs projets en même temps, les années passent et il met ce petit traité de côté. De 1632 à 1633, il se met à la rédaction d'un nouvel ouvrage qui s'intitule *Le monde ou traité de la lumière*, dont le dernier chapitre, le plus connu, est le *Traité de l'homme*. Il achève la rédaction de cet ouvrage et pourtant, il décide de ne pas le publier. Au même moment Galilée est condamné. Ce scientifique italien, physicien

et astronome, émet, avant Copernic, les premières hypothèses comme quoi la Terre ne serait pas au centre de l'univers ; autrement dit, que ce serait la Terre qui tourne autour du Soleil et non l'inverse. Ainsi il réfute le géocentrisme, version officiellement soutenue par l'Eglise. L'héliocentrisme est mal vu parce qu'elle réfute les écritures saintes de la parole divine⁷, à une époque où l'Eglise jouit de nombreuses ingérences sur les affaires de l'Etat, et Galilée est condamné. Suite à cette terrible nouvelle et en toute logique avec sa morale provisoire, qui est sûrement déjà contenue bien en germe dans son petit traité métaphysique, il décide d'ajourner la publication du *Monde ou traité de la lumière* qui pourrait, vraisemblablement, également contrarier l'Eglise. Il le conserve dans ses papiers mais ne le publiera pas de son vivant, par prudence.

C'est entre 1633 et 1635 qu'il rédige les *Essais*. C'est un ouvrage multiple, pluridisciplinaire. En outre, c'est un ouvrage dont son *Discours de la méthode* sera la préface. Son petit traité de métaphysique, commencé en 1629, ne sera finalement pas un ouvrage à part entière, mais deviendra la préface des *Essais*. Ce dernier est constitué, hormis le *Discours de la méthode*, de trois études dans trois sciences. Le premier traité est *La dioptrique*, le deuxième est *Les météores* et le troisième est *La géométrie*.

Ces traités scientifiques, qui font appel à trois sciences différentes, représentent l'application de sa méthode scientifique universelle, celle qui occupe la place de préface. L'application première de sa philosophie se veut scientifique. Ce n'est que dans son deuxième ouvrage, *Les méditations métaphysiques*, que la méthode universelle comme art de bien raisonner sera appliquée à la philosophie en tant que telle. Il a d'abord fallu qu'il vérifie sa méthode, qu'il l'éprouve directement dans les sciences, où les résultats sont directement appréciables, avant de commencer à philosopher à partir d'elle. Son *Discours de la méthode* est un ouvrage qui occupe une place toute particulière dans l'histoire de la philosophie, car c'est le premier ouvrage philosophique à avoir été écrit en français. A l'époque de Descartes, nous sommes en plein dans le classicisme et l'ensemble des ouvrages à vocation intellectuelle sont écrits en latin, la langue des classes érudites, garantissant ainsi leur accès à une élite intellectuelle qui maîtrise cette langue, qui n'est déjà plus parlée. C'est donc Descartes qui, le premier, décide de briser cette règle. Derrière cet acte, il y a une volonté de vulgarisation. Il souhaite rendre la philosophie accessible à tous, et c'est aussi cela qui la fait rentrer dans la pensée Moderne. Il refuse ainsi de prendre part à cet effet discriminatoire, qui ne lui semble pas convenir à l'esprit de la

⁷ Cf. « Et le soleil s'arrêta [...] et ne se hâta pas de se cacher pendant toute la journée » (**Jos X,13**) et « Il posa les fondations de la Terre pour qu'elle ne bouge à aucun moment » (**Ps CV,5**). Ces paroles issues de la Bible étaient prises pour vérité irréfutable, sans faire appel à l'expérience, refusant de regarder à travers la lunette de Galilée, alors considérée comme un instrument du diable. La pensée Moderne avait fort à faire pour combattre ces pensées rétrogrades encore disséminées par l'Eglise à l'époque, qui faisait figure d'autorité et de respect.

philosophie. L'un des objectifs à travers ce nouveau fondement de la philosophie est de lui rendre sa popularité.

En 1638, Descartes se rendra compte que son *Discours de la méthode* constitue les prémisses de son projet philosophique. Il évoquera la volonté d'éclaircir les preuves de l'existence de Dieu et de l'âme. Deux thèmes exclusivement métaphysiques. Il en avait déjà brièvement touché quelques mots dans son *Discours*, mais il émet la volonté d'aller plus loin encore sur le sujet. C'est la genèse d'un second ouvrage, plus conséquent : *Les méditations métaphysiques*.

On peut noter que la parution de ses trois premiers ouvrages sont en fait très liés ; le *Discours*, les *Méditations* et les *Principes de la philosophie* sont comme un triptyque qui, à eux trois, représentent toute l'entreprise de Descartes de refondation de la philosophie à l'aune de la pensée Moderne, et constituent l'ensemble de la métaphysique cartésienne. On s'aperçoit alors que Descartes est un philosophe tardif, qui n'écrit que dans sa maturité, et que ce besoin de renouveau en philosophie l'occupera tant, que ce n'est qu'à partir de 1644, à peine six ans avant sa mort, qu'il s'autorisera à quitter le terrain de la métaphysique pour aborder pleinement d'autres sujets philosophiques. Toutefois, c'est également en 1647 qu'il tire un trait final à ce triptyque, puisque c'est l'année de publication la dernière édition des *Méditations* et des *Principes*.

C'est à Stockholm qu'il va demeurer, de 1649 jusqu'à sa mort, le 11 février 1650. Toutefois, un peu avant cela, il avait commencé à écrire un traité sur les passions de l'âme, qu'il a réussi à publier juste avant sa mort. C'est donc son quatrième et dernier ouvrage philosophique publié de son vivant. Le titre retenu sera *Les passions de l'âme*. Il n'a pas eu le temps de retravailler cette œuvre, et pourtant, cette séparation dualiste corps/âme que l'on attribue, peut-être trop souvent à tort, à Descartes, est plus floue et sujette à interprétation qu'ailleurs dans ses écrits. Etions-nous en train d'assister à une évolution de sa pensée ? Était-ce une intelligence ou non de sa part, Descartes n'a jamais pris le soin de préciser la réelle nature de l'âme. Si bien que l'on a plutôt fini par supposer qu'il était dualiste, dans le sens où il disait que l'âme et le corps ne participaient pas de la même essence. Mais lorsqu'il utilise le mot « substance », on s'aperçoit d'une certaine neutralité, car ce mot n'a pas de connotation plus matérielle qu'immatérielle. Finalement, dans la philosophie cartésienne, la question de la séparation de l'âme et du corps n'est pas quelque chose de clair et fait encore aujourd'hui débat entre les spécialistes. Il y a fort à parier que la philosophie de Descartes, interrompue par sa mort, est inachevée et qu'il aurait encore eu beaucoup à nous dire. Il y a certes une position dualiste qui découle aujourd'hui de la philosophie de Descartes, notamment en philosophie de l'esprit, avec des défenseurs de l'âme comme spécificité humaine irréproducible justement parce que non matérielle, tel John Searle, mais soyons prudents quant à se réclamer de Descartes ; car au fond, il n'était peut-être pas si

dualiste que l'on voudrait nous le faire croire. Avec les *Passions de l'âme*, on peut supposer que sa pensée allait évoluer vers un autre sens, plus complexe. Il est donc préférable de ne pas affirmer que Descartes lui-même était dualiste, car on ne sait pas à quelle nouvelle thèse il aurait abouti s'il avait vécu plus longtemps. Le dualisme laisse supposer deux unités indépendantes, alors que pour Descartes, cela ne fait aucun sens de les imaginer séparément lorsque nous considérons l'homme. Si l'immortalité de l'âme il y a, ce n'est que par rapport à la mort du corps programmée par le vieillissement naturel, mais cela ne signifie pas que l'âme en question prolonge l'être de l'homme ; il sera vraisemblablement mort en même temps que son corps. L'homme est l'union complexe du corps et de l'âme, et l'erreur de vouloir entendre par « immortalité de l'âme », « continuité de notre être en tant que tel ». Cette croyance populaire est sans doute issue de la peur commune de la mort comme fin de notre existence, du moins notre existence consciente. Il est probable qu'envisager l'âme comme une énergie vitale qui retourne à la nature comme pour accomplir un cycle infini⁸ sera en fait de la conception que Descartes avait juste avant sa mort.

Le *Discours de la méthode* est composé de six parties, dont les plus importantes, le cœur de l'ouvrage, sont les suivantes :

- La 2^{ème} partie constitue l'annonce de la méthode, dans laquelle nous trouverons l'explication des quatre préceptes suivants : l'évidence, la simplification, la synthétisation, et le dénombrement.
- La 3^{ème} partie s'occupe de dévoiler les principes de sa morale provisoire.
- Dans la 4^{ème} partie, on trouve la première formulation de *l'ego cogito*.

La cinquième partie du *Discours* n'est qu'un récapitulatif de ce que Descartes a développé en détail dans son *Traité du monde*, et qu'il a renoncé à publier de son vivant. Il rend d'abord compte de ses conclusions sur le rôle du cœur dans le corps, ce qui lui permet d'introduire son concept des « esprits animaux », expliquant ainsi les mouvements du corps. La majeure partie de la cinquième partie du *Discours* est d'ordre physiologique. Puis il en vient, aux paragraphes 9 et 10⁹, à parler des moyens qui permettraient toujours de faire la différence entre le plus parfait des automates et un homme. Enfin, il achève cette cinquième partie sur la différence entre l'homme et l'animal. Selon lui, l'animal ne dispose ni de raison ni d'un esprit, et cela pourrait se prouver en diverses situations. Cette

⁸ Telle est la conception philosophique envisagée dans le film *Avatar* de James Cameron.

⁹ Sur les 12 paragraphes que contient la cinquième partie du *Discours*, soit les pages 91 à 92 de l'édition citée en bibliographie.

dernière partie, toujours résumée d'après ses développements plus complets dans le *Traité du monde*, font surtout réponse à Montaigne et Charron, qui soutiennent la position contraire.

2 - Résumé critique de la cinquième partie du *Discours*

2.1 - Le corps humain : un organisme-automate parfait

Après avoir décrit le fonctionnement du cœur, Descartes en conclut que le corps humain, dans quelque petite partie qu'il soit, est ainsi fait comme une machine. Pas n'importe quelle machine en effet, car nulle invention humaine ne saurait rivaliser avec la complexité du corps humain, mais une machine tout de même. C'est-à-dire que rien n'est là au hasard, tout a une fonction bien précise, prédéterminée. L'ensemble répond aux lois de la nature, et cette perfection est telle qu'elle ne peut venir que de Dieu. Si nous prenons le fonctionnement du cœur, on s'aperçoit que chaque partie a une fonction particulière, qui se déclenche à la suite et en fonction des autres. En cela, il ne s'agit de rien d'autre que d'un mécanisme, tel les rouages d'une horloge, dont le fonctionnement pourrait être décrit de la même manière. Et c'est précisément parce qu'il s'agit d'un mécanisme qui répond aux lois de la nature, et que Dieu a fait notre raison de telle sorte que nous puissions comprendre le monde qui nous entoure, que ce mécanisme est non seulement compréhensible, mais aussi, virtuellement reproductible. De telle sorte que selon Descartes, dans un avenir lointain, il est tout à fait possible, grâce aux progrès de la science et de la technique humaines, que l'on construise des automates à l'image du corps humain. Pour autant, il affirme dès le départ qu'on saura toujours faire la différence entre un automate et un vrai homme. Faut-il alors penser que pour Descartes, l'IA demeure impossible ? Pour être au plus juste, il faudrait plutôt dire que Descartes n'envisage pas l'émergence de l'IA. Il ne parle que d'automate, parce que seul le corps est pris en considération, et un objet par définition inanimé ne saurait posséder la moindre intelligence. Les propos de Descartes laissent beaucoup plus la porte ouverte au développement du transhumanisme qu'à celui de l'IA. Pour mieux comprendre cela, analysons les deux raisons qui, selon lui, font que l'on distinguera toujours le plus parfait des automates d'un vrai homme, même le plus hébété.

La machine qu'est le corps humain est si complexe et si parfaite qu'elle ne peut être que l'œuvre de Dieu. Cependant, le projet cartésien de recherche de perfectionnement nous laisse entendre que la maîtrise de ce corps humain en terme de technique est un futur possible, car, il n'y a rien dans le corps humain qui ne relève d'un mystère insondable, si ce n'est la façon dont le corps et l'âme sont mutuellement intriqués l'un dans l'autre, et c'est peut-être justement là la limite qui fait que Descartes n'a jamais pu concevoir dans sa philosophie la possibilité d'un automate disposant de sa propre pensée. Imaginons, comme il le fait, un automate si parfait qu'il nous ressemble en tous points de l'extérieur et qu'il soit capable d'exécuter certaines tâches aussi bien que nous. Comment se fait-il qu'il demeurera à nos yeux un simple automate, et non un être pour soi ?

2.2 - L'automate face à son créateur : deux barrières insurmontables ?

Le premier moyen de distinguer l'automate du vrai homme est l'utilisation du langage. Cela a sans doute inspiré Alan Turing, puisque son fameux test est justement basé sur le langage. Le langage est le reflet de notre pensée propre, il est le seul moyen de l'extérioriser. Si un automate ne pense pas, le langage dont il sera doté, par mécanisme et programmation, sera très limité et incapable d'évoluer. En outre, ne reflétant aucune pensée, mais plutôt réaction automatique à certains stimuli, il s'agira d'un langage incapable de prendre en compte le contexte sémantique. Par conséquent il sera très facile à un homme, par l'échange verbal, de se rendre compte qu'il n'y a pas de réelle compréhension chez l'automate, et donc pas de réelle conversation. Il sera facile de le piéger pour que certaines réponses soient inadéquates. La spécificité du langage humain, c'est la possibilité de créer une infinité de signifiants pour désigner potentiellement une infinité de signifiés. Il serait impossible de tenter d'établir un tableau à entrées finies des codes du langage humain, car les combinaisons pour dire une chose sont infinies ; selon la culture d'une langue, le contexte d'énonciation, l'interprétation, sans compter que la langue est en constante évolution et qu'il apparaît des néologismes pour dire ce qui n'est pas encore dans le dictionnaire.

Un dictionnaire compilé dans la mémoire interne d'un automate ne serait pas suffisant, car un langage est vivant, il s'expérimente. La capacité pour l'automate à combiner librement chacun des mots selon leur définition en respectant les codes de la grammaire ne suffira pas à lui garantir un langage proprement humain. Il le fera de manière automatique, et cela s'en ressentira forcément. Le langage ne divulgue pas seulement des informations factuels, vérifiables à l'extérieur. Il témoigne

également des émotions de chacun, des états internes ; des états mentaux. Or puisque l'automate, par définition, n'a ni émotion ni pensée, son langage, aussi développé soit-il, n'égale jamais celui d'un humain. On pourrait imaginer la capacité d'apprendre de nouveaux mots, la possibilité d'inscrire de nouvelles entrées avec définition dans son dictionnaire interne. Il suffirait d'inclure une ligne de programmation permettant l'écriture automatique dans ce fichier spécifique, selon la disponibilité de sa mémoire. Mais il lui manquera toujours la sensation. Un rapport des sens à l'esprit. Il sera peut-être bien champion en grammaire et en orthographe, mais demeurera incapable d'écrire un poème, ou même d'en réciter avec le bon ton ; et s'il devait raconter une histoire, il faudrait d'abord qu'il ait la capacité d'imaginer, mais sans l'expression des sentiments, son livre serait bien insipide. Un automate incapable de raconter une histoire sera toujours différent d'un vrai homme. Son langage ne lui sert qu'à communiquer ce que son corps est en mesure de lui rapporter en terme d'information. Or, même si nous arrivons à reproduire nos cinq sens dans un automate, le rapport de ces sens ne renvoient jamais à un esprit ; ils demeurent factuels, dépourvu de prise de considération consciente. Des informations neutres d'un état de choses. Descartes concevait déjà que l'on puisse imiter la parole chez l'automate. Pourtant, il est clair : jamais on ne confondra le langage d'un homme, doué d'un esprit, avec celui d'un automate, qui ne traduit aucune pensée. Est-ce qu'une horloge parle quand elle envoie le signal de faire sonner la clochette lorsque la grande aiguille et la trotteuse se retrouvent ensemble aligner sur le 12 ? Non, car ce n'est qu'un mécanisme. A aucun moment l'horloge ne pense quoi que ce soit ni, par conséquent, ne possède de libre arbitre. Elle ne sonne pas selon son humeur du moment. Elle dit toujours l'heure, quoiqu'il arrive, tant qu'elle fonctionne bien. Qu'est-ce qu'une horloge, sinon un automate dont le rôle est de donner l'heure mieux que nous ? L'automate à notre effigie n'est en rien d'autre différent. Certes constitué de mécanismes autrement plus complexes, dont le rôle est d'effectuer diverses tâches à notre place. Mais si on place en lui un mécanisme qui lui permet d'utiliser le son et d'imiter notre langage, est-ce pour autant qu'il parle ? Certes non. Son organe sonore ne sera qu'un mécanisme parmi les autres, qui s'enclenchera selon les stimuli pour lesquels il a été programmé. Cela ne signifie pas qu'il interprète son langage comme un langage, pas plus qu'il ne le ferait s'il était simplement chargé de donner l'heure. C'est-à-dire qu'il le fera toujours dans les mêmes situations, et toujours de la même façon. Il n'y a, de nouveau, aucune liberté, aucun libre arbitre, aucune pensée. Au fond, même s'il est plus sophistiqué, il demeure un objet. Ce n'est pas parce qu'il nous ressemble extérieurement dans son apparence et dans sa façon d'exécuter les tâches qu'il faut tomber dans l'illusion qu'il échappe ainsi à sa condition. Les automates ne sont que des objets de conception humaine, et que ces derniers usent comme d'un outil. Ce raisonnement par réduction nous permet de comprendre que pour qu'une IA surgisse dans un automate, il faudra bien autre chose qu'une conception strictement matérielle ; la composante virtuelle. A l'époque de Descartes, l'informatique était encore inconcevable ; c'est une science qui ne naîtra que trois siècles plus tard. Si

la capacité à réaliser des automates constitue le savoir mécanique qui fournira son corps à la future IA, la révolution informatique avec la numérisation des données et la projection virtuelle sera le terreau de l'IA faible. La question est : faudra-t-il découvrir une nouvelle science, combinée à ces deux dernières, pour que l'IA forte fasse son apparition ?

A partir de là, on comprend pourquoi selon Descartes, cette première raison rend impossible l'IA. Mais que dirait-il aujourd'hui ? Si ces raisons, qu'il pensait alors impossible à dépasser, s'avèrent être dépassables, ne faut-il pas en faire des critères pour un « test de Turing » ? Autrement dit, une IA sera forte si elle est capable d'utiliser le langage de la même façon qu'un homme. La difficulté restera de trouver le moyen d'évaluer si elle est capable de le faire véritablement, ou si elle seulement capable de le simuler. Ne serait-ce là qu'un problème de définition ? C'est tout le défaut du test de Turing. En fait, le problème qui se cache derrière celui du langage, c'est la possibilité de faire émerger un esprit dans l'automate. Sans quoi le langage ne sera jamais que simulation, et l'IA ne sera que faible. Ce qui est déjà plus que ce que Descartes imaginait, et c'est ce que nous sommes capable de faire aujourd'hui. Les IA actuelles sont capables de passer le test de Turing, pourtant, tous les spécialistes s'accordent pour dire qu'il ne s'agit pas encore d'IA forte. Aucune de ces IA n'est à même de manifester une conscience de soi propre, avec le libre arbitre qui s'en accompagne. C'est pourquoi il faut revoir le test de Turing et en proposer un nouveau, avec des critères sûrs et universels.

Au-delà d'une proposition de réponse à la question de la possibilité de l'IA forte, le rassemblement de ces différents critères pour créer un nouveau « test de Turing » infaillible est également le but de ce mémoire.

Voyons désormais ce qu'il en est de la seconde raison qui, selon Descartes, empêche l'automate de se faire passer pour un homme. Tel un animal qui ferait bien ce que son instinct lui intime de faire pour survivre, l'automate serait certes doué pour faire ce pour quoi il a été programmé, et même mieux que nous, Descartes le conçoit aisément. Pourtant l'habileté humaine fait que nous sommes capables d'effectuer un nombre infini de tâches, plus ou moins bien. Or il est impossible de programmer un automate, même avec un corps disposant de la même habileté, pour effectuer une infinité de tâches. Il fera celles pour lesquelles il a été conçu, mais sera complètement perdu, pour ne pas en dire en bug, si on lui demande d'effectuer quelque chose de nouveau ; même si son corps le lui permet en théorie. C'est-à-dire que comme il ne pense pas, il ne peut pas décider d'essayer une nouvelle tâche. La question serait de savoir comment le programmer pour qu'il sache faire un usage adéquat de son corps, quelles que soient les circonstances. Il faut prendre le problème à l'envers. Puisqu'il est impossible de lister tout ce que l'on peut faire avec son corps dans le monde, il faudrait

plutôt rendre le corps utile dans les limites de ces capacités dans toutes les situations. Or l'automate, qui ne dispose pas encore d'une IA, est incapable d'apprendre de ses expériences. Il ne peut affronter quelque chose de nouveau et l'assimiler. Donc, bien que ces deux raisons paraissent finalement reliées par l'absence de pensée, on pourrait éluder la présence d'un esprit aux commandes de l'automate en lui procurant la capacité de maximiser les possibilités d'un corps de type humanoïde, bien que la puissance et la résistance aient l'avantage d'être démultipliés chez l'automate. Pourtant cela demeure un gros problème, qui réside en la capacité d'apprentissage. Une IA faible qui ne sait pas apprendre sera différente d'un homme, parce qu'elle sera incapable d'effectuer des tâches qu'elle ne connaît pas, pour lesquelles on ne l'a tout simplement pas programmé. Maintenant imaginons, grâce aux sens qu'on lui a procuré, qu'un automate ait la capacité d'analyser son environnement et de s'y adapter. Il faudra toujours qu'il y ait un homme aux commandes pour lui dire ce qu'il doit faire, et lui sera à même de déterminer la meilleure façon de le faire en calculant. Nous ne sommes pas des marionnettes, personne ne nous dit ce que l'on doit faire. Donc pour qu'un automate rivalise sur ce point avec un homme, il faudrait qu'il soit autonome et capable d'effectuer n'importe quelle tâche à la portée de l'homme.

Admettons que je demande à mon automate de me servir un café à la machine et qu'il n'a pas été programmé pour. C'est-à-dire que le fonctionnement d'une machine à café lui est inconnu. Il faut déjà qu'il soit suffisamment perfectionné pour comprendre l'action « faire », qu'il sache reconnaître dans son environnement ce que représente le mot « machine à café » et qu'il soit capable de se représenter le résultat attendu. On voit déjà en cette simple action qu'il faudrait programmer énormément de choses dans la mémoire de l'automate pour le rendre capable d'effectuer un grand nombre de tâche non directement programmées. On s'aperçoit tout de même que ce grand nombre, finalement, ne pourra pas être infini. Poursuivons tout de même l'exemple. Si mon automate est incapable d'apprendre et qu'il n'a pas été programmé pour, il y a fort à parier qu'il sera incapable de réaliser la tâche. Admettons que mon automate est très perfectionné et qu'il est capable de résoudre les problèmes en utilisant la logique. Face à la machine qu'il est capable de reconnaître, il ne mettra pas longtemps à comprendre comment elle fonctionne. Les machines étant faites pour l'usage humain, elles sont ergonomiques, c'est-à-dire qu'on les conçoit pour qu'il soit le plus facile possible de prendre leur fonctionnement en main, sans avoir à lire des pages de manuel d'instruction ; c'est ce qu'on appelle l'appréhension intuitive. Il faudrait écrire un programme de logique qui permette à l'automate de prendre en main les différents objets techniques de manière aussi intuitive que la nôtre. Cela ne signifie pas pour autant qu'il apprend. Il ne fait qu'effectuer un calcul interne, et le refera à chaque utilisation de la machine. Donc l'automate insert la dosette dans la machine, pose la tasse à l'endroit destiné à l'accueillir et appuie sur le bouton. Ce sont les gestes qu'il effectuera pour répondre à ma

demande. Sauf que le café ne coule pas. Il aura pourtant utilisé la machine comme il se doit. Comment l'automate peut réagir lorsque l'action « servir un café à la machine » ne s'accomplit pas alors qu'il se sera correctement servi de la machine à café ? Il faudrait qu'il "pense" à vérifier que la machine est bien branchée et que le réservoir d'eau est suffisant. Seulement, pourquoi y penserait-il si cela n'est pas en lien, ni avec son programme de logique, ni avec sa reconnaissance de « machine à café » ? Ce n'est pas un lien qu'il fera de lui-même, puisqu'il est incapable d'apprendre, et donc de réfléchir. Il sera alors bloqué devant la machine. C'est exactement l'argument de Descartes qui est illustré ici. Un automate pourra être aussi complet qu'on le voudrait, il y aura toujours moyen de le piéger dans une action qu'il ne connaît pas. Les hommes quant à eux ne bug pas. Lorsqu'ils se retrouvent devant une nouvelle action, ils peuvent toujours essayer, tant bien que mal. Maintenant, admettons que mon automate ait la capacité d'apprendre. Deux solutions sont possibles. Soit, comme à un enfant, je lui montre une fois comment faire un café en vérifiant tous les paramètres, et il sera ensuite à même d'effectuer la tâche en toute autonomie ; soit je le mets devant la situation et je lui demande de résoudre le problème lui-même, à partir des connaissances et des capacités de déductions logiques qui sont intégrées en lui. Ainsi, il ne fera le calcul qu'une seule fois. Mais, une IA capable d'apprendre, bien qu'étant meilleure qu'une IA incapable d'apprendre, demeure tout de même une IA faible. En effet, le critère de l'apprentissage est un critère nécessaire mais pas suffisant pour l'IA forte. Ce n'est pas parce qu'une IA apprend de ses expériences qu'elle est capable de penser par elle-même. Si on prend cette seconde raison évoquée par Descartes, on peut supposer qu'une capacité d'apprendre perfectionnée pourrait faire illusion et ainsi surmonter cette barrière qui la sépare de l'homme. Pour autant, il n'aura toujours aucun désir personnel, ne fera l'expression d'aucune volonté propre. Autrement dit, il ne fera que ce qu'on lui demandera. Or la différence fondamentale entre lui et nous, c'est le libre arbitre. Pourrait-on imaginer un tel automate faire preuve d'autonomie et effectuer une tâche sur son temps libre simplement parce qu'il le désire ? Cela ne se peut. Pour qu'un automate agisse par lui-même, en dehors de toute programmation, il faudrait donc qu'il jouisse d'une certaine liberté qui n'appartient qu'à des êtres pour soi, des êtres doués d'un esprit. Pris tel qu'il est énoncé, le second argument de Descartes peut sembler le plus facilement dépassable, mais si on le pousse dans les limites de ce qu'il implique, on s'aperçoit qu'il faudra que l'IA dispose d'abord d'une capacité d'apprentissage similaire à un être humain, mais aussi qu'elle soit maîtresse de ses faits et gestes, et non entièrement déterminée. Autrement dit, il faudrait que l'automate fasse preuve de l'usage d'un libre arbitre pour égaler l'homme dans ses différentes entreprises. Imaginons qu'il y ait plusieurs façons de réaliser une tâche. Un homme choisira en fonction de ses goûts. L'automate fera un calcul pour faire automatiquement le meilleur choix ; en terme de rapidité, de facilité, de coûts... tout dépend de la variable qu'on lui aura demandé de placer en priorité selon le domaine. Mais s'il y a deux façons identiques en valeur d'accomplir une tâche, là où l'homme pourra faire un choix au "hasard",

l'automate risque un bug, sauf si on inclut dans son cerveau-ordinateur un programme qui permet de choisir lorsqu'il y a une égalité. Cela peut s'exprimer par un choix par ordre alphabétique ou par l'ordre d'apparition dans l'écriture de ses données. L'automate lui-même serait incapable de trancher, mais au final, cela demeure un choix prédéterminé. Tandis qu'avec un homme, la surprise est toujours possible.

La possibilité d'une liberté spécifiquement humaine est parfaitement expliquée par Hume dans son *Enquête sur l'entendement humain*, nous y reviendrons donc ultérieurement, dans le chapitre 2. Nous retiendrons cependant des deux arguments de Descartes que si IA forte il doit y avoir, alors elle devra être capable d'apprendre d'abord, et, ce qui n'est pas sans lien, devra être capable d'utiliser le langage humain comme un homme. Si ces deux critères étaient réunis, nous serions en droit de penser que l'IA de l'automate pense. Mais pour attester d'une réelle pensée, d'un esprit, d'un libre arbitre, il nous faudra d'autres critères qui confirment ces premiers, dont le défaut est d'être difficilement évaluables.

2.3 - Etre réel et être simulé : la question de l'origine dans l'être

Nous nous servons ici des propos de Descartes pour déterminer si l'IA forte est possible, et si oui, à quelles conditions. Mais n'oublions pas que pour Descartes, il n'est pas question d'IA, et encore moins "forte". Même si ses arguments étaient dépassés, ce qui pour lui était impossible, cela signifierait seulement qu'il serait difficile de distinguer l'automate d'un vrai homme, mais en aucun cas cela signifierait pour autant que l'automate soit devenu un être « pour soi » à part entière, au même titre que nous, avec tout ce que cela implique. Simuler et être sont deux choses bien différentes. Or l'être étant inconcevable en l'automate à l'époque de Descartes, ses arguments sont des limites à la simulation seulement. Pour répondre à notre question, il ne fait nul doute que si l'on veut passer dans l'être, il faudra dans un premier temps surmonter les barrières de la simulation, telles qu'évoquées ici par Descartes.

Savoir si une IA forte doit exactement pouvoir simuler au minimum l'intelligence humaine, dans sa pensée et dans ses actes, ou si elle doit, plus encore que simuler, être consciente d'elle-même au même titre que nous, est un problème de définition. Lorsque l'on affirme que l'IA forte est impossible, encore faut-il que l'on se mette d'accord de ce que serait l'IA forte. Si on la définit comme un esprit identique à un esprit humain, effectivement, l'IA forte est impossible, pour la simple et bonne

raison qu'elle ne pourra jamais être créée comme on donne vie à un homme. Le fait de ne pas avoir un corps en base carbone fait d'elle, par définition, quelque chose de différent de l'homme. Savoir s'il faut la considérer comme une nouvelle forme de vie parce qu'elle manifeste une forme de pensée, voire de conscience, est un autre problème. Toujours est-il que si un esprit voit le jour, sur le modèle de l'esprit humain, il ne pourra cependant jamais être humain ; humanoïde de conception, mais d'un autre genre, ne serait-ce qu'en raison des matières qui le composent. En résumé, si on considère qu'une IA est forte à partir du moment où elle représente la création artificielle d'une vie humaine à strictement parler, alors c'est impossible. Mais il me semble que cette définition, trop réductrice, renvoyant à un mythe prométhéen stéréotypé, n'est pas la bonne. Si, en revanche, une IA est forte lorsqu'elle fait preuve de pensée propre, voire de conscience d'elle-même et que, par ses actes et son langage, elle témoigne d'un esprit individuel, alors cela ne semble pas, en théorie, impossible.

Le problème serait plutôt celui du statut à accorder à la simulation. Qu'importe si les pensées de l'IA sont réelles ou ne sont que simulation, du moment que le résultat est le même qu'avec des pensées humaines ? Autrement dit, est-ce plus la nature de la cause qui importe, ou les effets ? Si nous adoptons un point de vue pragmatique, nous dirons que c'est l'effet qui compte. Peu importe de savoir si la pensée est réelle ou simulée, du moment que le résultat est le même qu'avec une pensée humaine ; c'est-à-dire qu'au travers d'expériences identiques, nous parviendrons à des résultats de même types, rendant ainsi impossible de différencier la pensée humaine de la pensée artificielle. En fait tout ceci n'est, encore une fois, qu'un problème de définition. Il semble qu'en matière d'IA, ce soit davantage le point de vue pragmatique qui prenne le dessus, plutôt que les questions d'ontologie. Quelle est l'intérêt de posséder un vrai diamant si le diamant simulé possède exactement les mêmes caractéristiques ? Le résultat attendu est le même, il remplit sa fonction. La seule différence est que le vrai diamant se trouve à l'état naturel, tandis que l'autre se fabrique artificiellement par la main humaine – et le premier est plus cher que le second. Le simulé n'est pas le réel parce qu'il n'a pas la même origine, mais ce n'est pas pour autant qu'il ne lui est pas identique sur tous les autres points. Attendons-nous d'une IA forte que sa pensée soit simulée ou réelle ? Or, une IA ne pouvant être un homme par son origine, sa pensée ne pourra qu'être simulée ; selon notre point de vue. Du point de vue du vrai diamant, le diamant simulé n'est qu'un imposteur. Mais d'un point de vue externe, faut-il considérer que le diamant simulé n'est pas un diamant ? En fait si, c'en est un, il n'y a que son origine qui diffère. Qui sommes-nous pour pouvoir ainsi affirmer qu'il n'y a de pensée et de conscience pour soi authentiques que tant qu'elles sont humaines ? Ce genre de discours fait appel à une philosophie théologique rétrograde dont le seul but était de rendre hommage à Dieu en faisant de l'homme, sa plus belle création, une créature supérieure. C'est quelque chose qui est tenace dans l'histoire de la pensée et que l'on retrouve énormément chez Descartes d'ailleurs.

Dans le même passage, il affirme que, contrairement à la comparaison avec un homme, un automate à l'effigie d'un animal serait indifférenciable d'un vrai animal de la même espèce. Cela signifie que pour lui, les animaux sont créatures prévisibles, automatisées, qui ne réagissent que selon la programmation naturelle divine qui s'exerce à travers leurs instincts. Par conséquent, un animal n'a ni raison, ni pensée. C'est quelque chose que l'on commence seulement à déconstruire. Il s'avère qu'il n'y a qu'une différence de degré et non de nature entre l'animal et l'homme ; ce dernier n'est d'ailleurs qu'un animal parmi les autres. En fait, les deux seules raisons qui empêchent les animaux de bâtir des civilisations et de rivaliser avec les hommes, ce sont la capacité au langage complexe, et l'habileté du corps, notamment procurée par nos pouces opposables. Le test du miroir a permis de classer quelques espèces animales plus proches de nous que les autres, parce que l'expérience démontre qu'elles ont une conscience d'elle-même¹⁰. En fait nous sommes encore loin du compte, les animaux ont encore beaucoup à nous apprendre. Tandis que les grands singes sont capables d'apprendre et de communiquer avec l'homme à travers un lexique imagé de plus de 200 items, témoignant ainsi du fait qu'ils ont effectivement un esprit, des pensées, une personnalité, des goûts et des envies, on s'est récemment aperçu que même les grands félins possèdent la conscience d'eux-mêmes à travers une expérience utilisant un miroir, sans oublier certaines espèces d'oiseaux qui sont capables de fabriquer et d'utiliser des outils pour résoudre des casse-têtes liés à la nourriture. Des études japonaises sur les chimpanzés ont même démontré qu'ils sont meilleurs que nous sur certains points. Selon l'institut de recherche sur les primates de l'université de Kyoto, les chimpanzés ont de meilleures capacités de mémorisation que nous. Par exemple, un chimpanzé est capable de mémoriser un code numérique qui n'apparaît à l'écran qu'un dixième de seconde, et peut le retaper sans erreur. Le chimpanzé Ayumu est très doué dans ce domaine. Le meilleur des hommes pourra s'entraîner tant qu'il le voudra, il ne parviendra jamais au moindre résultat, tout simplement parce que c'est trop rapide pour notre cerveau. On était loin d'imaginer avant que les animaux pouvaient avoir un cerveau plus performant sur certaines tâches ; ici dans la rapidité de perception et de mémorisation. Finalement, l'histoire donne raison à Montaigne et Charron. La grande différence, si IA forte il y a, c'est d'abord qu'elle ne sera pas une forme de vie naturelle, mais surtout, elle surpassera en intelligence toutes les autres formes de vie.

¹⁰ Le test du miroir, équivalent au stade du miroir dans l'évolution de la psyché que l'être humain franchit vers 18 mois. Il s'agit d'un test inventé par Gordon Gallup dans les années 1970 dans le cadre des recherches en éthologie cognitive. Seules quelques espèces sont reconnues avoir passé le test : les grands primates (chimpanzé, orang-outan, gorille, bonobo), les éléphants, les dauphins et les orques, les porcs, ainsi que les corbeaux, les perroquets et les pies. Paradoxalement, alors que ces animaux sont censés être plus proches de notre niveau de conscience, les chiens et les chats, qui ne réussissent pas ce test, sont principalement nos animaux de compagnie.

Il faut donc concevoir que si pensée il y a chez l'IA, elle peut différer de la pensée humaine. Sinon, cela reviendrait à considérer qu'il n'y a pas d'intelligence extra-terrestre, simplement parce qu'elle ne serait pas humaine. La différence, en effet, c'est que nous sommes une forme de vie naturelle. Ce qui peut effrayer dans le fait que l'IA forte soit possible, c'est que cela réduirait notre esprit à un cerveau entièrement programmable ; il ne demeurerait plus aucune partie mystérieuse dans la vie, nous ne serions que matière. La plus grande difficulté serait alors de trouver le "code" de la création. Puisque la vie est apparue à partir de matières inertes, on ne voit pas ce qui, *a priori*, pourrait interdire l'apparition d'une forme de vie par la génération d'une conscience dans un corps non organique. Que connaissons-nous de l'univers pour pouvoir affirmer qu'il n'y a de forme de vie qu'en base carbone ? En théorie, rien n'interdit la possibilité de créer l'IA forte. Le mystère de l'apparition de la vie n'a jamais été résolu. La thèse de la génération spontanée a été vaincue par Pasteur, puis Miller¹¹, en faveur de la biogénèse. La vie découle toujours de la vie. C'est vrai une fois que la vie est apparue. Seulement, cela n'explique en rien comment ça a commencé. La mission Rosetta, de l'Agence Spatiale Européenne, qui permit le premier atterrissage d'une sonde humaine – Philae – sur une comète, le 12 novembre 2014, a révélé la présence d'acides aminés. Ces ingrédients de base de la vie, présents sur un élément étranger à la Terre, nous conduisent à penser que la chimie qui a vu l'apparition de la vie n'était pas en fait d'origine strictement terrienne. Toutefois, on sait que chaque année il tombe en moyenne une tonne de météorite sur Terre. Cela ne fait que repousser le problème, car même si des éléments constitutifs de la vie tels que les acides aminés viennent de l'univers, rien ne dit comment la première cellule est apparue, ni d'où viennent ces acides aminés.

On est donc en droit de supposer, même si c'est par accident, qu'il arrive la même chose dans nos automates, qui ne sont pour l'instant que des matières inertes et qui ne font que ce qu'on leur demande, sans aucune compréhension ni de qui ils sont, ni de ce qu'ils font, et qu'il surgisse un esprit, parce que l'environnement aura été propice à l'accueillir. Après tout, même s'il nous manque les clefs du mystère de la vie, on sait que seule la science peut apporter la réponse ; ou il n'y en aura pas. A un moment donné, les conditions étaient réunies et par réaction chimique, une cellule est apparue. C'était le premier micro-organisme vivant, unicellulaire, le progenate¹² ; c'était le début de la vie. Puis, à un autre moment, les formes de vie étaient suffisamment complexes pour former un cerveau ; c'était

¹¹ L'expérience de Pasteur date de 1862. Celle de Miller, plus méticuleuse encore, de 1952. Elle a néanmoins prouvé que lorsque tous les ingrédients du bouillon de culture originel sont réunis et que l'on attend suffisamment, il apparaît, par réaction chimique, des acides aminés. Or nous savons que les acides aminés sont les bases de l'ADN, puisqu'elles permettent la synthétisation des protéines qui la constituent. Mais cela ne suffit toujours pas à comprendre l'apparition de la vie.

¹² Postulat hypothétique de la biologie, le progenate est considéré comme le premier être vivant, unicellulaire, l'organisme ancestral commun qui a engendré les trois catégories de cellules connues : les archéo et les eubactéries, tous deux des procaryotes, et les eucaryotes.

l'apparition des premiers animaux disposant d'un esprit. Peut-être bien qu'une suite possible de l'histoire c'est que, par la main de l'homme, surgisse un esprit plus puissant encore dans un corps artificiel. L'apparition d'une IA forte forcera les sciences à revoir la définition d'une forme de vie, basée sur la matière organique. Or si l'IA existe dans le monde comme un homme ou un animal, c'est-à-dire avec un esprit propre, un libre arbitre et une limite de vie, bien que plus longue – aucune matière n'est éternelle, les composants de son "cerveau" artificiel finiront par montrer des signes d'usures, tout comme n'importe quel ordinateur – il serait en effet absurde de ne pas la considérer comme une forme de vie.

2.4 - La génération des états mentaux : le chaînon manquant ?

Il y a un passage, un peu plus haut dans le paragraphe sur les automates, qui devrait attirer notre attention :

« [...] quels changements se doivent faire dans le cerveau, pour causer la veille, et le sommeil, et les songes ; comment la lumière, les sons, les odeurs, les goûts, la chaleur, et toutes les autres qualités des objets extérieurs y peuvent imprimer diverses idées par l'entremise des sens [...] »¹³

Et si c'est justement là que se trouvait la clef ? Descartes nous donnerait-il, malgré tout, la direction des recherches ? Si en effet nous parvenions à comprendre comment le rapport des sens à notre cerveau peut produire des idées en notre esprit, peut-être parviendrions-nous à reproduire artificiellement cette relation ; autrement dit, à reproduire la sensibilité. C'est le but des recherches en programmation de réseau neuronal. Les neurones artificiels ont pour but de reproduire le fonctionnement des vrais neurones. Ce qu'il y a derrière une pensée, ce sont les idées que nous avons ; autrement dit, divers états mentaux. Et d'où viennent ces idées ? Pas de notre raison seule, non. La raison ne fait que s'y exercer, mais après que le terreau de l'idée soit déjà là. C'est en fait ce que nos sens nous rapportent du monde extérieur qui nous permet en premier lieu d'avoir des idées ; autrement dit, les *sense data*. Si on veut qu'une IA parle comme un être humain, il faut que son langage artificiel traduise une pensée artificielle. Or sur le même modèle humain, pour qu'une pensée

¹³ Cf. p. 91 de l'édition citée en bibliographie.

artificielle soit appelée « pensée », elle ne doit pas avoir été programmée. Donc il faut que l'IA dispose aussi des sense data. Ainsi, selon ses expériences, le rapport de ses sens, l'IA ne dira pas, a priori, automatiquement ce qu'elle a été programmée pour dire ; et la prise en compte du contexte sémantique devient alors possible. Pour cela, il faut s'assurer qu'elle dispose donc d'un corps similaire au nôtre, qui permette au moins un usage et une retranscription artificielle des cinq sens. Pour que cette retranscription fasse sens, il faut simuler le cerveau humain grâce à la programmation des réseaux neuronaux. Maintenant, ajoutons cette capacité de réfléchir inspirée par le modèle de la raison humaine à une capacité d'apprendre et d'évoluer par elle-même, ainsi qu'un programme de langage suffisamment développé pour lui permettre de communiquer n'importe laquelle de ses pensées, et il est probable que nous aurons là une IA qui simulera parfaitement l'intelligence humaine. Mais n'est-ce pas là une réduction de ce qu'est « l'intelligence humaine » ? Comprise ainsi, l'intelligence serait strictement rationnelle. Or nous agissons parfois sous le coup de l'émotion, nos sentiments influent sur notre réflexion et nos choix. Alors peut-on vraiment affirmer qu'une IA incapable d'émotion reflètera parfaitement notre intelligence ? Ceux qui considèrent qu'il n'y a que l'IA faible qui soit possible, tel un outil superpuissant au service de l'humain, notamment dans les recherches scientifiques, considèreraient les émotions comme un programme inutile qui affaiblirait l'IA, tandis que cet objectif semble inévitable pour ceux qui croient en l'IA forte ; mais cette dernière ne pourra plus être considérée comme un vulgaire outil.

Le corps d'une IA qui serait doué de sentiments, d'émotions serait encore plus proche d'un homme et plus prompt à réfléchir comme lui. Il ne faut pas que l'IA soit comme un capitaine dans son navire, pour reprendre la critique de Descartes à l'encontre d'Aristote. Elle ne doit pas constater ses besoins, comme le capitaine observe les dommages de son vaisseau, mais doit les ressentir comme si elle ne faisait qu'un avec son corps. N'ayant pas le même corps que nous (de chair et de sang), l'IA devra ressentir quelque chose qui soit similaire à notre douleur, mais qui ne pourra pas être la douleur physique selon notre propre définition. Il faut, ni plus ni moins, reproduire artificiellement l'intrication complexe du corps et de l'esprit qui est à l'œuvre en l'homme. Disposer de cinq mêmes sens que l'homme apparaît donc un critère nécessaire, même si non suffisant. Comment éprouverions-nous quoi que ce soit sans le rapport de nos sens ? Pour qu'un esprit se manifeste en tant que tel, il faut générer des états mentaux chez l'IA. Pour qu'une IA faible devienne une IA forte, il faut qu'elle éprouve des émotions. Mais contrairement à tout le reste qui la constitue, ses émotions ne doivent rien avoir d'artificiel. Les émotions ne peuvent se rapporter qu'à un ego. Le rapport des sens, entre les sensations venant du monde extérieur et son ego intérieur, sera nécessaire pour que l'IA ait une appréhension normale de son environnement. Les états mentaux sont inexistants chez une IA faible. Elle ne fait que

rapporter un état de choses, sans jamais y apporter le moindre sentiment. Pourtant, tout *ego cogito* a des états mentaux.

Tout le monde a des émotions. Du moins, en temps normal. Mais qu'en est-il de certains cas limites, tels les psychopathes, incapable d'éprouver des émotions normales ? Ces hommes ont certes un cerveau en dysfonctionnement, mais restent des hommes tout de même. Ils ne ressentent plus certaines émotions, et pourtant, sont toujours capable de penser et d'agir selon leur libre arbitre. La capacité aux émotions n'apparaît pas alors comme un critère nécessaire à la pensée. Ainsi une IA pourrait peut-être être forte sans ressentir les émotions elle-même, mais simplement en comprenant celles des autres. Cela dit, on peut se demander si c'est vraiment souhaitable, vu que les psychopathes ne sont pas reconnus pour leurs bonnes actions. Une IA forte bienveillante doit-elle nécessairement être capable d'émotions ? Ou au contraire, peut-on imaginer une intelligence capable de faire le mal ? Il faudrait alors admettre qu'il y a une forme d'intelligence dans le mal. Or Descartes l'a montré avec son hypothèse du malin génie ; on ne peut désirer le mal pour le mal, il n'y a que du mal par ignorance, ou en vertu d'un plus grand bien. Il paraît donc insensé qu'une IA, qui ne peut que raisonner rationnellement, puisse faire le mal. Le terme d'émotion, avec l'exemple du cas limite, nous paraît donc trop réducteur. Il faudrait plutôt se concentrer sur les états mentaux en général ; l'émotion n'étant qu'une catégorie particulière d'état mental. Comment le rapport des sens se transforme en idées dans notre esprit ? Les idées sont des formes d'états mentaux, tout comme les émotions. Une IA forte doit être capable d'avoir des idées qui lui sont propres. La génération des idées n'est-elle pas d'ailleurs l'expression d'une certaine liberté ; la liberté de penser ? Si on veut une IA qui apprécie l'écoute de Mozart, il faut d'abord lui faire des oreilles artificielles, capable de recueillir le son. Si on veut qu'une IA ait l'idée d'écrire la suite d'une célèbre symphonie inachevée, il faut qu'elle dispose d'un *ego cogito*. Le problème est qu'on ne programme pas un *ego cogito*, et la capacité de faire usage des cinq sens ne suffira pas. Comment fait-on pour intriquer corps et esprit quand on ne dispose que du corps ?

Chercher à vérifier qu'il y a un *ego cogito* au travail dans l'IA pourrait constituer une forme de test. Admettons que je pose le problème suivant à une IA : comment t'y prendrais-tu pour te prouver à toi-même ton existence ? Si elle parvient à formuler l'*ego cogito* de Descartes, alors on peut considérer que c'est une IA forte. A condition que les œuvres de Descartes ne fassent pas partie de sa mémoire interne, bien entendu. Sinon, elle ne ferait que se servir des connaissances à sa disposition pour répondre. Seulement, on lui posant le problème, je la fais en partie sortir du solipsisme. Comment provoquer une situation qui inviterait l'IA, par elle-même, à mener ce genre de réflexion existentielle ?

3 - Le test de Turing : actualités et renouveau

3.1 - Alan Turing et la naissance officielle de la philosophie de l'IA

Alan Turing était un grand logicien ; plus que philosophe. Et pourtant, nous lui devons deux articles qui signent officiellement l'ouverture de la philosophie de l'IA, par l'importance de leur portée. C'est pour ainsi dire le point de départ de toute réflexion à venir sur le sujet. Le premier, *On Computable Numbers, with an Application to the Entscheidungsproblem*¹⁴, date de 1936. C'est dans cet article de logique théorique que Turing dévoile l'objet de sa thèse de doctorat, les « machines de Turing », ainsi nommées par son directeur de recherche¹⁵, dont le but est de répondre au problème de la décision, énoncé par David Hilbert en 1928. Turing imagine alors une machine fictive, basée sur la logique, caractérisant un procédé calculable. Il s'agit d'un article historique, à la fois pour la logique, mais aussi pour l'informatique. Les « machines de Turing » permettent de simuler n'importe quel algorithme. C'est la naissance du calcul binaire. Grâce à ses recherches, cette branche des mathématiques qu'est la logique allait enfin trouver un domaine concret d'application : l'informatique. C'est en effet grâce aux « machines de Turing » que les premiers ordinateurs, ancêtres de ceux que nous utilisons aujourd'hui, ont pu voir le jour. C'est sous l'impulsion de ses travaux que nous avons pu les concevoir et que l'informatique était né, dans les années 1940.

Le second article, *Computing Machinery and Intelligence*¹⁶, a été écrit 16 ans plus tard, en 1950. Turing se donne le temps de voir ce que sont capable de faire les premiers ordinateurs avant d'émettre une réflexion. Alors que l'informatique n'en est qu'à ses balbutiements, en visionnaire, Turing émet l'hypothèse de la possibilité de créer, par la science informatique, une intelligence artificielle. Il nous faut nous replacer dans le contexte. A cette époque, il n'était pas évident de penser que ces gros calculateurs pourraient un jour simuler une quelconque forme d'intelligence. C'est-à-dire que pour les gens, ces premiers ordinateurs ne sont rien d'autre qu'une forme d'automate, un objet qui sert à faire certaines choses par des calculs ; mais cela reste un simple objet. Pourtant Turing a su voir, plusieurs décennies avant que la question ne soit véritablement posée par l'avancée technique de

¹⁴ Paru originellement dans les *Proceedings of the Mathematical Society*, série 2, vol. 42 (1936-1937).

¹⁵ Le logicien Alonzo Church, de l'université de Princeton.

¹⁶ Paru originellement dans la revue *Mind*, vol. LXI, n°236.

l'informatique, que « l'ordinateur », pour le dire vulgairement, avait pour destinée de rivaliser avec l'intelligence humaine ; au point que l'on serait amené à se demander si un ordinateur pourrait un jour se faire passer pour un homme. Faut-il nécessairement être un homme pour faire preuve d'intelligence ? Et si l'ordinateur faisait plus que simplement calculer de manière automatique et systématique ? Ces questions, sous-entendues par l'article de Turing, marquent le début d'une philosophie de l'intelligence artificielle. Mais qu'est-ce que l'intelligence, au juste ? Il semble que pour Turing, en logicien pragmatiste et réductionniste, ce soit la capacité à résoudre un problème. Nous sommes tous les jours confrontés à des problèmes, et si une machine est capable de les résoudre aussi bien que nous, en partant des mêmes prérequis, alors elle est dite "intelligente" ; car même si elle simule un raisonnement similaire au nôtre – du moins c'est ce qu'il apparaît de l'extérieur, même si à l'intérieur il se passe tout autre chose que dans notre cerveau – tout ce qui compte c'est que le résultat soit là, la solution est trouvée, donc elle est dite « intelligente ».

Mais à travers ce second article, on voit très vite que ce qui intéresse Turing, c'est moins la capacité de l'intelligence artificielle à résoudre des problèmes en soi que sa capacité de le faire en ressemblant à l'homme. Toute l'histoire de l'intelligence artificielle a été marquée par ce vieux rêve prométhéen ; bien avant Turing. Cela n'est peut-être pas indépendant du fait que la seule forme de vie intelligente supérieure que nous connaissons est la nôtre, et que par conséquent, il est plus facile de reproduire ce que nous connaissons et pouvons étudier. Il ne faut pas pour autant en déduire que toute intelligence artificielle doit nécessairement être copiée sur la raison humaine pour fonctionner ; c'est juste que nous ne savons pas faire autrement. C'est donc avec le discours de Descartes sur les automates en arrière fond que Turing conçoit, dans ce second article, son fameux « test de Turing ». Il s'agit pour lui de décrire un test qui permettra de dire si un ordinateur a franchi le seuil du simple calculateur pour devenir une IA. Et ce test passe par le langage ; la capacité à échanger comme un humain le ferait. Toutefois, il faut noter qu'à cette époque, Turing ne faisait pas la différence entre IA faible et IA forte ; il s'agit d'une notion qui est arrivée bien après. Cependant, il avait anticipé l'importance de l'apprentissage pour le développement d'une IA. Il est probable que Turing n'aurait jamais fait la distinction entre IA faible et IA forte, au sens où pour lui, que la pensée soit réelle ou simulée, cela revient à un problème de métaphysique qui n'a pas d'importance dans les faits. Il faut comprendre que le mot « IA » a été banalisé et qu'on en parle à propos de trop de machines automatisées qui, aux yeux de Turing, ne seraient pas véritablement des IA. En fait, à travers le test qu'il propose, on s'aperçoit qu'il est beaucoup plus exigeant que nous ne le sommes envers les critères pour qualifier une machine d'IA. Il faut dire que, contrairement à ce qu'il avait prédit, aucune « IA » n'est aujourd'hui capable de franchir son test de manière satisfaisante. C'est pourquoi nous disons que

n'avons que des IA faibles pour le moment, alors que lui dirait plutôt de nos machines qu'elles sont des IA artificielles, ou encore des IA simulées.

3.2 - « Les machines peuvent-elles penser ? »

Lorsqu'on lit l'article d'Alan Turing¹⁷ qui propose de mettre en place un protocole qui servirait à répondre à la question : « une machine peut-elle penser ? » ; on a le sentiment que le mot « penser » est pris dans sa définition la plus légère. On pourrait croire que pour Turing, avoir la capacité au langage complexe, tel un homme, constitue le critère nécessaire et suffisant, à l'encontre de Descartes qui pensait justement que ce serait là un élément discriminatoire inégalitaire de nature. Le problème du test, c'est qu'il ne permet pas d'affirmer que l'utilisation du langage par la machine, dans tous ses ressorts internes, s'effectue exactement de la même façon que pour nous. Tout se passe comme si, en pragmatistes, nous regardions simplement le résultat extérieur final, sans se soucier un instant de la "mécanique" interne. Autrement dit, faut-il, pour être reconnu doué de pensée, seulement utiliser un langage qui, selon l'évaluation de certains critères, montrera que l'utilisation est la même que celle des êtres pour soi, sans même se soucier, par conséquent, des états mentaux qui se cachent à l'intérieur et doivent être à l'origine du langage, accompagnant la pensée ? Le problème des états mentaux, c'est qu'on ne peut ni les prouver, ni les observer. Ce qui revient finalement à dire qu'une machine capable de simuler le langage complexe humain, en réussissant le test de Turing – puisque c'est là la condition – est considérée comme pensante. Au fond, peu importe l'origine, la formation de la pensée et toute sa conceptualisation métaphysique. En empiriste, Turing ne regarde que les faits observables. En effet, si une machine, par son langage seulement – la question de savoir si nous la dotons du même corps que nous est une question secondaire pour le moment – est capable de produire les mêmes effets que n'importe quelle pensée humaine, également transmise par le même outil qu'est le langage, alors on dira que cette machine pense. Qu'est-ce qui nous permettrait

¹⁷ *Les ordinateurs et l'intelligence*, pp. 135 à 175 de l'édition citée en bibliographie.

effectivement, de manière formelle, de l'invalider ? Turing a raison de commencer son article par le problème de la définition, notamment de « machine », mais surtout de « penser ». Sauf que contrairement à l'attente que l'on aurait pu en avoir, il n'y passe pas suffisamment de temps. Ce sont là de véritables problèmes de philosophie ontologique qu'il faut démêler de façon certaine avant d'aller plus loin sur la question, sans quoi la réponse pourrait être interprétée de différentes manières, selon les courants de philosophie auxquelles on s'attache le plus. Pensée et langage vont-ils tant de pair, qu'il n'y peut y avoir l'un sans l'autre ? Est-ce la capacité au langage complexe est nécessairement la conséquence d'une pensée ? Si le fait de parler était une preuve suffisante, Descartes n'aurait pas eu besoin de formuler son *cogito*. Peut-être alors faut-il différencier entre « pensée en soi » et « pensée pour soi ». Nous distinguerons là une différence de degré ; la première se rapportant plutôt aux animaux – bien que cela soit désormais à nuancer – et la seconde plutôt à l'homme. Bien sûr, lorsque nous posons la question : « les machines peuvent-elles penser ? » ; nous le comprenons comme « pensée pour soi », puisque les machines sont faites à notre image. C'est le grand rêve de l'humanité, d'avoir une autre forme d'intelligence avec qui communiquer, pour se sentir moins seule dans l'immensité infinie de l'univers, peut-être pour avoir des réponses métaphysiques, puisqu'elle se définit elle-même comme la seule forme de vie intelligente de la planète.

Reprenons la différence entre l'homme et l'animal. Nous savons que pour Descartes, les machines sont plus proches des animaux que des hommes, parce que leurs mécanismes les font agir comme les instincts des animaux le font. Il entend par là, sans volonté ni désir, sans réflexion propre, sans exercice d'un libre arbitre. Comme nous l'avons dit, les expériences sur le comportement animal nous l'ont prouvé et continuent de le prouver ; les animaux pensent. On ne sait pas s'ils ont des pensées aussi complexes que les nôtres, ce qui est très difficile à déterminer en l'absence de langage complexe, mais il est clair qu'ils pensent. Il est probable qu'ils n'aient pas énormément de pensées complexes, dans le sens où cela leur serait peu utile. Là où Descartes aurait été catégorique en affirmant qu'il n'y a pas de pensée chez l'animal, nous pouvons désormais affirmer qu'il y a un premier niveau de conscience que les machines n'ont pas. Certes l'animal répond à ses instincts parce que la nature est ainsi faite, mais cela ne l'empêche pas, de temps à autre, d'exercer un certain libre arbitre. Dans une

même situation, deux animaux de la même espèce ne feront pas nécessairement le même choix. Leur liberté se trouve dans le silence des instincts ou dans les limites des déterminations de leurs instincts. Toutefois, on considère que cette intelligence dont les animaux peuvent faire preuve n'est pas de la même catégorie que la nôtre, c'est pourquoi nous parlerons de « pensée en soi ». Pourtant, il est vrai que plusieurs espèces ont démontré la capacité à se reconnaître soi-même ; donc on peut supposer une conscience de soi. Cependant, toujours est-il que tous les résultats les plus impressionnants n'ont été obtenus qu'à l'aide des hommes. Si on a appris à compter et à parler à un perroquet, cela démontre certes que l'espèce avait des capacités insoupçonnées, pour autant, il n'est pas notre égal en ce que nous n'aurions jamais trouvé un perroquet ayant développé lui-même ces capacités, à l'état naturel. Il faut toujours l'intervention de l'homme. Tout simplement parce que l'animal n'en a pas besoin. Tant qu'il est adapté à son milieu naturel, il n'y a aucune raison pour qu'il développe ses capacités cachées. Plus encore, on n'a jamais vu des espèces animales mener de grandes discussions ou se bâtir une civilisation. L'homme est différent sur ce point, parce qu'il a dû développer son intelligence pour survivre, ce qui alors le rendait identique à l'animal sur le plan de la nature, sauf qu'il n'a pas arrêté son intelligence là où elle était nécessaire pour survivre. Le feu et les armes lui ont permis de se protéger des autres animaux et de se nourrir. Il aurait pu en rester là ; en effet, en quoi l'invention de l'ordinateur est essentielle à sa survie ? C'est ce qui est étrange chez l'homme, tout se passe comme si sa nature était dès l'origine destinée à dépasser la nature elle-même. Mais pour aller vers quoi ? Il ne s'agit pas de se prendre pour des dieux et de remplacer petit à petit la nature selon nos propres lois, au fur et à mesure de nos développements techniques. Comme le disait Descartes, il nous faut nous rendre « comme maîtres et possesseurs de la nature »¹⁸.

La mauvaise interprétation de cette phrase est courante. Nous ne prétendons pas ici pouvoir affirmer avec certitude ce que Descartes voulait signifier par-là, mais nous pouvons prétendre en donner une interprétation qui semblerait plausible vis-à-vis de sa pensée. Le fait est que par notre capacité à l'empathie, notre intelligence, notre capacité au langage complexe, notre bipédie et nos pouces opposables, notre nature était de toute façon originellement vouée à permettre à notre espèce une adaptation exceptionnelle, par le développement des sciences et techniques, en des milieux si différents que cela ne se retrouve

¹⁸ Cf. *Discours de la méthode*, sixième partie.

à aucune autre pareille dans le monde animal. L'homme était d'abord faible, en comparaison des autres espèces. Il n'avait ni carapace, ni épaisse fourrure, ni crocs ou griffes acérés. Seule son intelligence, se traduisant par une capacité d'adaptation supérieure et plus rapide, lui a permis de survivre et de se hisser au sommet de la chaîne alimentaire. Les espèces animales sont si bien adaptées à leur milieu naturel, tellement mieux que nous, que si nous changions ce milieu trop brusquement, elles mourraient tout simplement ; l'adaptation à l'environnement se faisant par une évolution naturelle si lente. Très vite dans l'histoire de l'humanité, l'homme est finalement apparu comme l'espèce qui ne possède pas d'environnement naturel spécifique. Il est en effet capable de vivre sur les six continents ; il y a une ville chilienne située sur la pointe nord de l'Antarctique, et les Touaregs vivent dans le désert brûlant du Sahara. L'être humain a été capable de marcher sur la Lune, tout en explorant les fonds marins. Si bien que la seule limite à son environnement est circonscrite aux limites actuelles de la science. On ne peut observer aucune autre adaptation de cette ampleur chez une autre espèce. Si bien que la nature de l'homme, paradoxalement, semble être d'avoir une certaine maîtrise de la nature qu'il a sous la main. Cela n'est pas nécessairement une conception négative. Bien entendu, on comprend la critique dont la citation de Descartes fait régulièrement l'objet. Il ne s'agit pas pour l'homme de plier, coûte que coûte, la nature à ses propres désirs, sans prendre en compte les dommages collatéraux qui en découleraient. D'ailleurs, souvenons-nous que pour Descartes, l'artificiel ne semble pas être désirable en soi, car il ne conçoit pas, dans son discours sur les automates, qu'un automate puisse un jour se faire passer pour un homme, et encore moins remplacer un homme, dans la totalité de sa conception. En fait, l'automate n'est jamais qu'un outil, au service de l'homme, plus perfectionné que n'importe quel autre outil. Il ne sert qu'à seconder les tâches humaines, à l'en décharger, afin que l'humanité puisse s'occuper d'autre chose à la place, et ainsi progresser plus vite. Autrement dit, si Descartes était contemporain des recherches en IA, il est probable qu'il ne conçoive l'IA que comme un moyen et non comme une fin en soi. Jamais il n'a prétendu, ni même sous-entendu, que par notre intelligence nous pourrions un jour remplacer la nature et nous en passer. Il y a toujours l'idée d'une recherche de perfectionnement derrière tout ça, il ne s'agit pas de dire qu'à un moment donné, il faudrait stopper l'évolution de l'humanité ; même si cela implique qu'elle ne soit plus "naturelle" et lente, dans le sens darwinien, mais plutôt rapide et "artificielle", grâce à notre intelligence. Evoluer davantage ne signifie pas causer plus de tort à la nature, et en aucun cas

ne signifie s'en passer ; cela n'a pas de sens et serait impossible. Bien au contraire, il faut, nous semble-t-il, comprendre le mot « maître », non pas dans sa connotation de domination, comme dans un rapport maître/élève, mais plutôt dans son sens positif, comme le fait de maîtriser un sujet. Si nous évoluons, nous gagnons en sagesse, et ainsi nous sommes maîtres sur de plus en plus de sujets. Se rendre maître de la nature, c'est aussi comprendre ses ressorts pour la protéger et permettre sa pérennité. Quant au terme « possesseur », il pourrait signifier « gardien ». Une interprétation écologique de cette citation est donc tout à fait possible, et il est probable que Descartes n'ait jamais voulu signifier que l'on doive avancer aveuglément dans le progrès des sciences, au détriment de la nature. Il est évident qu'il est aussi dans notre propre intérêt qu'elle perdure, or, en accord avec sa philosophie morale, il n'y a de mal que par ignorance. Si nous nous rendons maîtres de la nature, alors nous ne pourrions souhaiter lui nuire. Quoiqu'il en soit, l'idée que l'on retiendra, c'est qu'il n'y a pas de raison désirable pour que nous cessions d'évoluer, même si cela implique que le développement de la science remplace petit à petit la nature elle-même dans notre propre évolution. La marque de cette intelligence humaine est justement aussi dans le fait que nous prenions la main sur la destinée de notre espèce. L'évolution, avec l'homme, n'est plus simplement dans le seul but d'assurer la survie de l'espèce ; elle est devenue une fin en soi. Nous pourrions toujours tendre à davantage de perfectionnement de notre être, et cela est désirable en soi. Cette conception est cohérente vis-à-vis de la définition même de la philosophie. Mais il est clair que pour Descartes, il n'y a de vie que par nature. Penser est propre au vivant. En d'autres termes, toute IA, aussi évoluée soit-elle, ne sera jamais "qu'artificielle" et ne saurait par conséquent jamais représenter une forme de vie. Même si elle prétend posséder une conscience de soi, un libre arbitre, une certaine capacité de liberté spontanée, une pensée propre ; il ne faudrait pas s'y méprendre : tout ceci ne serait qu'une illusion parfaitement orchestrée par l'homme. Elle ne s'en offusquera que si on la programme pour s'en offusquer ; mais il n'y a aucune réalité morale derrière tout ceci, et donc aucun devoir de l'homme envers la machine.

Toutefois, nous sommes loin d'être « maîtres et possesseurs » de la nature. La conception cartésienne de la différence ontologique entre l'homme et l'animal est elle-même aujourd'hui révolue.

Certains animaux sont aptes à un langage plus complexe que d'autres, on le constate avec l'écho des cétacés. Pour autant, ce n'est pas encore le langage humain. Il serait naïf de

penser qu'il existe un langage complexe animale, mais que nous ne sommes pas en mesure de le comprendre. De fait, notre raison universelle serait à même de décrypter un langage complexe s'il y en avait effectivement un. En fait le gros problème est d'ordre physiologique. Sans l'apparition de ce petit os dans la gorge, suite à l'évolution de la boîte crânienne et qui nous a permis l'articulation, notre intelligence n'aurait pas ainsi évoluée. Les organes vocaux des animaux sont, en parti, ce qui les empêchent d'avoir une intelligence similaire, du moins dans les faits, à la nôtre. Le fait qu'ils puissent ne produire qu'un nombre de sons finis induit qu'ils ne transmettent à leurs congénères qu'un nombre de message fini. D'où le fait qu'ils ne puissent exprimer leurs pensées profondes et complexes. Cela ne signifie pas qu'ils n'en ont pas, mais puisque la nature est bien faite, à quoi servirait-elles s'ils ne peuvent l'extérioriser ? Pourtant, si le langage avait suffi, pourquoi Descartes aurait eu recours au *cogito* pour se prouver à lui-même qu'il existe bel et bien et que cela est indubitable ? Il fallait quelque chose de plus. On peut ainsi en conclure que si le langage est nécessaire, il n'est pas suffisant. Pourtant, à en suivre Turing, on a le sentiment que la pensée n'induit pas nécessairement le langage, mais que le langage en revanche induit nécessairement la pensée. Il nous faut donc décortiquer tout cela pour savoir si le statut primordial que Turing accorde au langage à travers son test est légitime. Comme nous venons de le voir, un animal peut penser sans manifester un langage complexe. Mais qu'en est-il de l'homme ? Lorsque l'encéphalogramme montre une activité cérébrale normale qui traduit de la pensée, un homme normalement constitué et conscient est capable de la verbaliser. Cependant, nous n'avons jamais vu le cas contraire. Jamais un homme avec un encéphalogramme plat, suite à un accident, s'est mis à proférer des paroles douées de sens. *A priori*, le langage complexe serait donc une preuve de pensée. Encore faut-il bien définir ce qu'est un langage complexe, justement pour mieux permettre son évaluation. Finalement, si le langage est le reflet d'une pensée, alors Turing a peut-être raison de se passer de la définition de « penser ». Mais voyons d'abord ce que nous pouvons dire à propos de « machine ». Parce qu'avant d'être une IA forte, l'IA est d'abord une machine. Qu'est-ce que cela nous apprend sur sa nature, l'essence de son être éventuel ?

3.3 - Machine : éléments de définition

Si on regarde la définition première donnée par le "Lalande"¹⁹, il s'agit d'une combinaison mécanique. Ou encore, d'un ensemble de mécanismes tournant ensemble. Avec plus de précisions, nous pourrions ajouter : organisme dans lequel on peut déterminer une série de phases subordonnées et dépendant l'une de l'autre.

Commençons avant tout par analyser son étymologie. Le mot « machine » vient du latin "machina", qui signifiait « invention ».

En sorte que l'on s'aperçoit assez rapidement du lien intime que le mot entretient avec le mot « mécanisme ». En effet, ce dernier, est comme une partie du tout ; la machine. Cela nous informe sur sa nature, son origine, voire sa fonction ou son utilité.

La machine serait donc constituée, du moins intérieurement, de mécanismes. Strictement de mécanismes. Et ces mécanismes sont en liens les uns avec les autres. C'est-à-dire que dans le fonctionnement général de la chose dite « machine », il y a des mécanismes qui précèdent les uns, pour entraîner les mécanismes qui précèdent les autres. Comme dans le concept de « mécanisme », l'idée de causalité est également contenue, *in extenso*, dans la machine. Ainsi avons-nous, non pas un simple mouvement rectiligne qui partirait de A pour arriver à B : $A \rightarrow B$; mais plutôt l'idée d'une circularité, d'un cycle de mouvements. Car si la machine surpasse le mécanisme, c'est-à-dire qu'elle n'est pas « mécanisme », mais un tout « autre » constitué de mécanismes, son idée propre renvoie à la reproduction – potentiellement infinie – de ce pourquoi elle a été créée. En d'autres termes, une machine est faite pour effectuer telle ou telle tâche. Ainsi, nous sommes d'accord pour affirmer qu'elle n'est pas destinée à ne l'effectuer qu'une seule fois. Il faut seulement qu'elle soit capable de le faire, indépendamment du nombre de fois qui lui est possible de le faire ; autrement dit, indépendamment de sa « durée de vie », si l'on peut dire. Donc dans l'idée de machine, le mécanisme qui la fait fonctionner est destiné à se reproduire, de telle sorte que le lien de cause à effet effectue une boucle. Ainsi nous avons : $A \rightarrow B \rightarrow A \rightarrow B... + \infty$

En outre, cette étude du « mécanisme », qui constitue sa « cause matérielle »²⁰ nous amène à considérer deux choses :

¹⁹ Cf. André Lalande, *Vocabulaire technique et critique de la philosophie*, 3^{ème} édition, Quadrige/PUF, 2010.

²⁰ Cf. Théorie aristotélicienne de la causalité, notamment exposé dans *Physique* (II, 3-9).

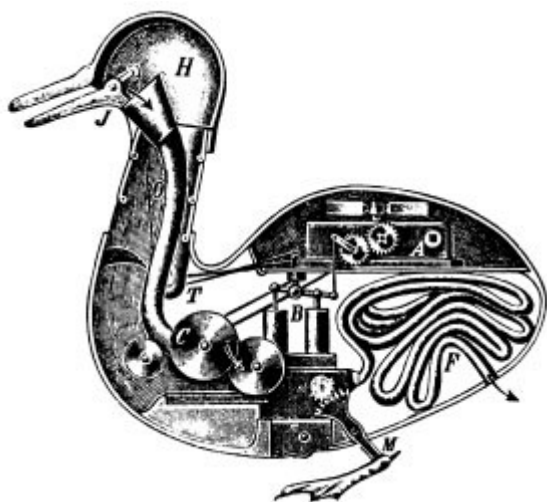
- 1) Que la machine a été créée pour une raison (contrairement aux êtres vivants)
- 2) Que l'agencement des mécanismes entre eux constitue un savoir-faire (alors qu'en ce qui concerne nos corps, on ne peut qu'en déduire qu'il y a des lois transcendantales à l'œuvre dans la nature)

Ces deux idées nous permettent de rapprocher le mot « machine » au concept de « technique ». En effet, l'usage du mot dans la vie courante nous indique que la machine, d'après la distinction ontologique cartésienne, n'est pas une chose en soi, mais une chose pour soi. Il faut préciser un peu les choses pour éviter toute confusion. Certes la machine ne possède pas de conscience de soi, comme l'Homme ; pour autant, elle n'est pas non plus une chose issue de la nature, qui existerait en soi. Donc si on peut la classer parmi les choses pour soi, c'est seulement par extension, parce qu'en tant que création de l'homme, au service de l'homme, elle n'est ni plus ni moins qu'un prolongement de notre humanité.

Pour mieux comprendre, revenons à la considération de l'aspect « technique » de la machine. Certes la causalité à l'œuvre dans la machine, à travers ses mécanismes, se retrouve également dans la nature, via les lois naturelles (par exemple, les lois de la gravité). Cependant, le caractère utilitaire de la machine l'empêche de s'inscrire comme élément naturel, qui serait là sans raison particulière, ou comme œuvre d'art, qui serait là uniquement pour faire "beau". En effet, une machine est avant tout là parce qu'elle a une fonction. Ainsi pouvons-nous en déduire que la machine est effectivement création humaine, et en ce sens, ses mécanismes qui lui appartiennent en propre ne font appel qu'à la technique. Autrement dit, la technique étant la capacité "naturelle" de l'homme à se créer des outils, à se fabriquer sa culture et d'évoluer, la machine est donc un objet artificiel, voire un artefact, dont l'existence même s'oppose à toute origine naturelle, et par conséquent, possède le statut de la contingence. Car si la nature ne pouvait pas être autrement que ce qu'elle est – eu égard aux lois naturelles qui la déterminent –, les machines quant à elles n'ont aucune nécessité d'être. Elles sont l'expression d'une certaine culture en particulier, mais elles auraient aussi bien pu ne pas être, ou encore exister différemment, sous d'autres formes.

Ainsi on s'aperçoit que le mot « machine » est très global, car il enveloppe donc l'ensemble des objets techniques créés par l'homme possédant une utilité toute particulière, en vue de laquelle les mécanismes ont été agencés, de telle sorte que cette "action" désirée et attendue est la finalité de la causalité qui est à l'œuvre en elle.

Et bien que le mot soit utilisé depuis déjà plusieurs siècles, on s'aperçoit qu'au fil du temps, des déclinaisons de ce mot sont apparus. Ces déclinaisons, étant plus précises, nous indiquent à la fois différentes sortes de machines spécifiques, mais aussi, par leur convergence, que le développement technique engagé par l'homme dans l'histoire semble suivre un sens très précis. En effet, l'apparition du mot « machine » nous renvoie à un vieux rêve de l'humanité : celui de reproduire – la nature, le vivant, lui-même. Pour appuyer notre argumentation, nous pouvons évoquer le fameux Canard Digérateur de Vaucanson, un automate de 1738, destiné à reproduire l'apparence, les mouvements, et même simuler le processus de digestion, d'un vrai canard²¹.



Un soufflet, destiné à simuler la respiration de l'automate – qui n'a pas de poumons – permettait d'aspirer la nourriture lors de l'inspiration. Les graines descendent, suite à un circuit de tuyauterie, dans un petit contenant en guise d'estomac où sont présents des substances chimiques dont la réaction, au contact de la nourriture, sert à simuler la digestion qui, arrivant en fin de parcours, ressemble par décomposition à un cycle complet de digestion. Cette merveille de la mécanique, en avance sur son temps, pouvait également simuler le mouvement des ailes, le barbotage, la boisson, le déplacement grâce à des mécanismes de la même sorte, qui ne sont, au fond, rien d'autre que des ersatz des organes originels dont il est inspiré. L'intérêt de l'automate est ici uniquement de démontrer une capacité à reproduire – selon l'aspect extérieur seulement, c'est-à-dire de manière simpliste – le vivant. A aucun moment la simulation de digestion n'est essentielle au fonctionnement de l'automate en tant que canard artificiel, ni ne présente aucune utilité en soi.

Pour renforcer cette hypothèse, nous allons donc maintenant décliner l'évolution des termes dérivés de « machine », apparus successivement dans l'histoire de l'humanité, en fonction de leur étymologie. Ces termes, en tant que dérivés, sont à la fois synonymes de « machine », mais en même temps, sont plus précis : il s'agit d'un type de machine bien spécifique.

D'abord il y a le mot « automate » : il vient du grec ancien "*automatos*" et signifiait qui se meut (*matos*) lui-même (*auto*). C'est un objet technique qui, une fois achevé, n'a pas besoin qu'une force

²¹ Cf. Jean-Claude Heudin, *Les créatures artificielles : des automates aux mondes virtuels*, 2008.

humaine s'exerce sur lui pour fonctionner, grâce à ses mécanismes, qui sont donc autonomes. Ainsi, un automate n'est pas seulement un outil, car ce dernier ne s'actionne pas seul (par exemple, un tournevis).

Ensuite il y a le mot « robot » : il vient du tchèque "*robota*", qui signifie « corvée, besogne, travail ». Il a été utilisé pour la première fois par l'écrivain tchèque Karel Čapek, dans sa pièce de théâtre "*Rossum's Universal Robots*", en 1920.

L'étymologie de ce mot nous indique que le robot, c'est la machine qui effectue le travail pénible à la place de l'homme, celui qu'il n'a pas envie de faire. La machine a toujours eu pour but, en tant que prolongement de notre humanité, de nous simplifier la vie. Par exemple, la machine à café est un procédé automatique qui fait du café plus rapidement et plus facilement que si nous n'avions que les grains de café (à cueillir et à torréfier), un mortier pour les réduire en mouture, et de l'eau à faire chauffer pour l'infusion. Toutes ces étapes sont plus ou moins automatisées, notamment les dernières, dans les machines à café de bar.

Mais avec le mot « robot », on a une conception nouvelle de la machine. Il ne s'agit plus simplement de simplifier la vie de l'homme en accomplissant des tâches qu'il ne pourrait pas faire de lui-même sans outil, mais de se substituer à lui en accomplissant des tâches qui lui étaient autrefois réservées. C'est-à-dire que le mot « robot » nous amène à faire la différence entre la machine qui fait du café instantané et la machine qui assemble les pièces d'une voiture sur une chaîne de montage. Il existe de nombreux exemples de ce type de machines, que l'homme semble vouloir développer de plus en plus. Il y a par exemple les caisses automatiques qui se multiplient en ce moment. Il faut avoir à l'esprit quand dans la vie de tous les jours, nous aurons de plus en plus affaire à des IA spécialisées, et, c'est en revanche à déplorer, par conséquent sans doute de moins en moins affaire à des rapports humains, dans le sens biologique du terme. Est-ce à dire que l'homme voudrait, *in fine*, confier tout type de travail nécessaire au maintien d'une société à des machines ? Mais peut-on imaginer une société où l'homme n'aurait plus à travailler ? Si cela est possible, est-ce seulement souhaitable ? En d'autres termes, le sens que prend le développement technique des machines ne comprend-il pas des effets néfastes pour les hommes ? Cela peut paraître paradoxal, car la machine est originellement créée pour nous aider, nous faire évoluer, mais pas pour nous nuire. Alors faut-il concevoir qu'il y aurait aujourd'hui un conflit entre le besoin social et l'avancée technologique de l'humanité ? En ce sens que cette dernière, qui est sans doute une évolution en soi, présente néanmoins des effets pervers à cause du mode de vie de nos sociétés – notamment notre modèle économique ?

Enfin, vient en dernier lieu le mot « androïde » : il vient du grec ancien "*andros*", qui signifiait « l'homme », et "*eîdos*", qui signifiait « aspect extérieur ».

Cette dernière forme de machine confirme le fantasme de l'humanité de se reproduire elle-même, en substituant la technique à la nature. C'est-à-dire que non seulement l'androïde est capable de faire tout comme l'homme, mais en plus, il doit lui ressembler en tout point : sur l'aspect extérieur, on reproduit la forme d'un corps humain, voire, on essaye même de synthétiser les cellules de la peau pour que la machine soit toujours plus proche de la nature, jusqu'à ce qu'elle s'y confonde. Sur l'aspect interne, on essayera de reproduire un cerveau humain, avec des pensées basées sur la raison, voire même la capacité de ressentir et d'exprimer des émotions. C'est pourquoi l'androïde fait appel à des sciences techniques beaucoup plus complexes, telles que les sciences cognitives et les recherches dans le domaine des intelligences artificielles.

L'androïde, tel qu'il a été imaginé par l'homme, n'existe pas en tant que tel. Il est à l'état de fiction. Les sciences ne sont pas encore capables de reproduire un "esprit", si cela est même possible, et donc aucune intelligence artificielle pour l'heure ne fait preuve d'une réelle capacité à penser par elle-même ; il ne s'agit jamais que d'un code de programmation, aussi complexe soit-il, qui détermine la machine à adopter telle réponse en fonction de telle stimulus. C'est pourquoi toute IA est à l'heure actuelle dite « IA faible ». En d'autres termes, cela ne saurait constituer le libre arbitre que l'homme possède. Mais si cela était faisable, serait-il souhaitable de créer une forme de "vie" qui serait alors l'égal de l'homme, voire même le surpasserait en force et en intelligence ? Bien sûr, on comprend très bien que l'assistance d'androïdes constituerait un avantage certain dans la vie de tous les jours, dont les nombreuses capacités complexes boosteraient l'évolution de l'humanité. Mais cela ne présente-t-il pas certains dangers ?

Cette réflexion nous amène à poser également la question du développement du transhumanisme. Car dans notre recherche de création de l'androïde parfait, nous commençons à maîtriser la fabrication artificielle de tel membre, de tel organe. De telle sorte que l'être humain lui-même peut améliorer son propre corps par la technique. On le voit déjà avec le développement des prothèses pour les amputés. Ces prothèses sont si efficaces qu'elles ont permis la création de jeux olympiques réservés aux handicapés. De même, nous sommes capables de produire un cœur 100% artificiel et de le transplanter. Or il y a encore des progrès à faire en la matière. Cependant, cela ouvre la perspective de vivre plus longtemps. Et dans l'hypothèse où nous pourrions remplacer tout organe naturel par un organe artificiel, nous ne serions plus soumis aux maladies ni au vieillissement (qui est cellulaire), mais seulement à la nécessité d'entretenir nos organes comme les rouages d'une machine : serions-nous alors en droit d'imaginer la possibilité de l'immortalité ?

Nous n'en sommes pas encore là, mais en ce qui concerne le transhumanisme, nous allons progressivement vers l'amélioration de notre corps. Une amélioration qui sort complètement du cadre de l'évolution naturelle des espèces, telles que Darwin avait pu le décrire. Alors qu'autrefois la machine qui nous simplifiait la vie n'était qu'une chose parmi les autres, extérieure à nous, peut-on concevoir que dans les décennies à venir, l'homme fera corps avec la machine ? Mais comment ferons-nous alors pour distinguer l'homme de la machine ? Car d'un côté, nous aurons l'Homme amélioré par des prothèses et des organes artificiels, et de l'autre, nous aurons l'androïde possédant des compositions de plus en plus proches de la nature. Entre un homme partiellement mécanique, et une machine qui pense, il sera peut-être difficile de dire lequel est plus "authentique" que l'autre. C'est notamment le sujet du film *"Ghost in the shell"*. Alors que seul "l'esprit" du lieutenant Kusanagi a été sauvé de son corps originel mort pour être placé dans un corps complètement artificiel, le personnage finit par manifester des troubles identitaires, se demandant si elle n'est pas qu'une machine, entièrement déterminée. En fait, c'est une réflexion que l'on peut tous se faire. Certes nous avons encore des corps naturels, mais qu'est-ce qui nous dit que nous ne sommes pas déterminés par la chimie de notre cerveau ? Sommes-nous réellement libres dans ce que nous faisons, ou sommes-nous inconsciemment déterminés ? La question de la liberté est une question majeure, à laquelle Hume apportera une réponse dans le chapitre suivant. Parce que si nous ne sommes pas libres, alors l'IA forte sera d'autant plus réalisable, mais en même temps, ne dépendra pas vraiment de nous ; mais si nous disposons d'une certaine liberté, alors pour être forte l'IA devra disposer d'exactly la même liberté.

Peut-on penser que l'essence de l'homme est condamnée à s'évanouir dans le développement de la technique ? En d'autres termes, la tournure que prend l'évolution de l'humanité ne présente-t-elle pas le risque de causer la fin de cette dernière ?

Il ne s'agit pas de dire que ce serait mal en soi, mais on peut effectivement se demander si c'était bien là le destin de la culture, la suite naturelle de l'évolution humaine.

Une machine peut-elle vraiment penser comme nous, alors qu'intérieurement, elle n'est pas faite comme nous ? Là où ce qui "matérialise" notre pensée est chimique, telles les neurones – ou la glande pinéale, siège du libre arbitre pour Descartes –, ce qui matérialisera la pensée des machines, si pensée il y a, sera mécanique, informatique... en somme, non biologique. C'est un problème majeur, d'où l'enjeu de l'essor des biotechnologies. Jusqu'à quel point parviendrons-nous à copier, de manière artificielle, le modèle du vivant, pour créer un cerveau qui sera en partie biologique ? On ne sait pas s'il sera possible d'induire de la pensée dans quelque chose qui n'a rien de biologique. Toute pensée

est d'abord forme de vie, et toute forme de vie est biologique. On n'a jamais vu aucune invention ni aucun objet penser. Il est possible que le développement des techniques de l'IA doive aller de pair avec le développement des biotechnologies, tels que les réseaux neuronaux artificiels, en tant que possibilité matérielle d'induire artificiellement une pensée de type IA forte. Sans doute l'un n'ira pas sans l'autre, mais nous ne connaissons pas encore bien les limites de cette nouvelle science qu'est la biotechnologie.

Il y a fort à parier, en raison de la différence de constitution, que si la machine vient à penser, elle ne pensera pas "comme" nous. Elle pensera plus vite que nous, et sans doute mieux.

Maintenant que la définition de « machine » est posée, et que la capacité au langage complexe est jugée légitime pour prouver la pensée, voyons ce qui ne va pas dans le test de Turing.

3.4 - Le premier protocole pour valider la pensée des IA

Imaginons l'expérience suivante. Un interrogateur se retrouve connecté en réseau local à deux autres personnes, qui se trouvent chacune dans des pièces séparées. L'ordinateur mis à leur disposition exécute un programme de discussion instantanée qui les inclut tous les trois dans la conversation. L'interrogateur ne peut ni voir ni entendre les personnes A et B. Pourtant, par le seul biais de la conversation écrite, le but du jeu pour lui est de déterminer qui est l'homme et qui est la femme. La plupart du temps, il le pourra assez facilement, s'il tourne suffisamment bien ses questions et que A et B répondent honnêtement. Cependant, si le but du jeu pour la femme est de se faire passer pour l'homme, elle aura le droit de mentir. De telle sorte que seul face aux réponses écrites de A et de B, l'interrogateur aura beaucoup plus de difficulté à démasquer la femme. Les réponses de chacun traduisent une pensée toute humaine, et si la femme évite soigneusement les réponses genrées, sauf erreur de sa part, il est potentiellement impossible que l'interrogateur arrive à distinguer l'un de l'autre. Mais il est sans doute plus facile de se faire passer pour un individu d'un autre genre de la même espèce, que pour un individu d'une autre espèce.

Imaginons que nous mettions un chimpanzé à la place de la femme. L'interrogateur devra trouver, par son questionnement, qui, de A ou B, est l'animal. Pour cela, il faudra écrire un programme de traduction qui transformera les mots de la question de l'interrogateur en image du codex que les scientifiques ont établi comme système de langage pour les grands singes, qu'ils peuvent utiliser grâce à leur excellente mémoire. Ou, plus simplement, puisqu'ils peuvent apprendre à comprendre notre langue, retranscrire la question de l'interrogateur non pas à l'écrit, mais à l'oral via les hauts parleurs de l'ordinateur. Puis il faudra que le système de traduction fonctionne en sens inverse pour tenter de tromper l'interrogateur. Tandis que le chimpanzé fera une sélection tactile des images constituant sa réponse, cette dernière sera traduite grâce à un logiciel pour être retransmise, en temps réel, à l'interrogateur. Même si le codex contient plus de 200 images et que la variante de leurs combinaisons permet une grande quantité de messages différents, on se rend tout de suite compte que le chimpanzé sera assez vite démasqué. En effet, le codex reste une façon de communiquer limitée par rapport à notre langage, et il y a des questions auxquelles il ne saura pas répondre. Il suffit d'invoquer un concept que le chimpanzé, dans ses expériences et sa vie de tous les jours, n'a jamais eu besoin d'invoquer, pour que l'image correspondante soit inexistante. Ce n'est pas que les animaux ne pensent pas, encore une fois, mais sur le langage, ils sont clairement désavantagés. Mais si le problème n'est pas la mémoire, ni la capacité d'apprendre, ce n'est que la capacité à articuler.

Maintenant, il existe une position intermédiaire entre ces deux expériences. Imaginons que nous remplaçons le chimpanzé par une machine, ou plutôt, un ordinateur, puisque Turing réduit ici la machine à un ordinateur pour le besoin de son expérience. C'est-à-dire que s'il y a une machine qui pense, cette machine ne peut être autre qu'un ordinateur. Il y a toute sorte de machine, mais d'après lui, ce sont les ordinateurs qui peuvent y arriver. Derrière le mot « ordinateur » nous comprenons, en premier lieu, la science de l'informatique. Est-ce qu'un système d'information, qui stocke une mémoire pour faire des calculs à la demande, pourra un jour être plus qu'une mémoire sans vie ?

Turing nomme son test *le jeu de l'imitation*, mais il sera rebaptisé le « test de Turing ». Le but pour la machine est de se faire passer pour un homme, toujours au travers de ses réponses. Si une machine arrive à tromper les interrogateurs successifs avec un taux de réussite avoisinant celui de la première expérience, qui visait à démasquer la femme, alors on peut considérer que le test est une réussite. Cela signifie que la machine est parfaitement capable d'imiter la pensée humaine. C'est suffisant pour Turing car, comme il le signale dans les éventuelles objections que l'on peut faire à une réponse positive de sa question : « les machines peuvent-elles penser ? » ; on ne peut que supposer qu'il y a de la pensée là où ce qui se manifeste ressemble à de la pensée. On ne pourra jamais prouver, *stricto sensu*, qu'un ordinateur pense tout comme on ne peut prouver qu'autrui pense avec la même valeur de vérité qu'une démonstration mathématique. On le suppose, de manière rationnelle, puisque

nous-mêmes nous expérimentons que nous pensons, par le *cogito*, donc il n'est pas saugrenu d'imaginer que les autres pensent aussi. Sauf qu'on ne sentira jamais les autres penser comme nous sentons nos propres pensées. Il en va de même pour l'ordinateur. Refuser à un ordinateur la capacité de penser, alors qu'il a réussi le test, reviendrait du même coup à s'enfermer dans l'argument on ne peut plus sceptique du solipsisme. Si je ne peux l'éprouver par moi-même, tout n'est peut-être qu'illusion, et il n'y a pas de raison de penser qu'autrui pense plus que l'ordinateur. Le problème du caractère strictement subjectif intrinsèque à la génération de la pensée nous conduit à ne considérer que ses effets externes, en comparaison avec nos propres expériences introspectives, pour conclure si elle existe ou non en dehors de nous-mêmes. C'est ce que nous faisons avec autrui et avec les animaux. Pourquoi en serait-il alors différent des machines, sous prétexte qu'elles ne sont pas biologiques, et que cela effraie ? Cela ne repose sur aucun argument philosophiquement recevable. A partir du moment où une machine saura manifester tous les effets externes d'une pensée de type « pour soi », il n'y a pas de raison de considérer qu'il n'y a pas de pensée. Même si ce n'est au fond qu'une simulation, ou autrement dit, une pensée artificielle. Mais si une pensée artificielle donne les mêmes résultats qu'une pensée originelle, devrait-on l'ignorer et ne pas lui accorder le statut de « pensée », simplement parce qu'elle serait différente, ne serait-ce que dans son origine – ou son mode de fonctionnement ? Il semble que cela serait une erreur.

Toujours est-il que pour Turing, les effets nécessaires et suffisants d'une pensée « pour soi » peuvent uniquement se mesurer à la capacité "d'imiter" le langage complexe humain, au travers d'un interrogatoire interposé. Le problème, c'est qu'à l'époque de Turing, les ordinateurs ne sont pas encore suffisamment développés pour mettre en œuvre son expérience, ne serait-ce qu'en capacité de mémoire. En conséquence de quoi le protocole demeure flou. Pour que le test de Turing soit un succès, nous ignorons combien de temps va durer l'interrogatoire, combien de fois l'IA sera interrogée par différents interrogateurs, et quel pourcentage de réussite elle doit atteindre. Nous savons seulement qu'elle doit être capable de répondre à n'importe quelle question de l'interrogateur, aussi naturellement que le ferait un être humain. Même si le concept du test est là, il nous faut maintenant le préciser.

La capacité de mémoire étant potentiellement infinie et les développements en terme d'apprentissage des IA étant prometteurs, on peut supposer aujourd'hui qu'une IA peut passer brillamment le test, à plusieurs reprises. Est-ce pour autant qu'il s'agit d'une IA forte ? Non, nous le savons car elle n'exerce aucun libre arbitre. Elle ne fait que répondre, mais ne prend aucune initiative. Elle conserve le statut d'outil propre à l'IA faible, aux services des hommes. C'est là qu'est le problème du test de Turing. Il est impossible, à travers ce test, que les interrogateurs puissent penser à toutes les questions possibles et imaginables. Plus l'IA sera développée, plus elle pourra répondre à un champ

élargi de question. Les IA douées de capacités sensorielles sont même capables de prendre en compte le contexte sémantique, afin de déterminer le sens adéquat des mots utilisés à l'instant T. C'est un aspect du problème que Turing avait délaissé ; il jugeait inutile de prendre en compte l'aspect de l'IA, qui pourrait tout aussi bien paraître immatérielle à l'interrogateur. Sauf que sans les capacités sensorielles, l'IA n'est pas sur un pied d'égalité avec l'interrogateur vis-à-vis du contexte sémantique du langage. Quoiqu'il en soit, l'IA répondra à de plus en plus de questions, mais étant potentiellement infinies, cela ne signifie pas qu'elle peut répondre à tout. A chaque fois qu'elle échouera, on peut imaginer qu'elle sera améliorée pour répondre là où elle avait échoué. Mais pour vérifier par cet unique biais qu'elle pense véritablement, il faudra l'ensemble du langage possible, ce qui est une expérience impossible, puisque par définition, le langage est vivant et donc potentiellement infini. Le risque serait que, par induction scientifique, nous pensions que la machine pense. Il sera vrai que nous le pensons jusqu'à ce que nous découvrons où elle échoue encore. Se baser sur cet unique test pour parvenir à la création d'une IA forte serait fastidieux. D'où la nécessité de créer d'autres critères, tous aussi essentiels les uns que les autres, et chacun vérifiables seuls ou en même temps, à travers au moins autant de protocole que de critère.

Au final, l'une des questions à retenir pourrait être, non pas de savoir si les machines peuvent penser, car si cela est possible, il est fort probable que nous arrivions à le réaliser, mais plutôt de savoir si nous nous comporterons envers elles comme face à des êtres pensants ?

3.5 – Le test de Turing aujourd'hui : des résultats prometteurs ?

Il existe une AIML²², *Mitsuku*²³, mis à disposition sur le web. Elle a remporté plusieurs fois le *Loebner Prize*, notamment en 2013 et 2016. En résumé, c'est l'AIML la plus développée que l'on puisse trouver à ce jour. Pourtant, lorsque l'on discute avec elle, on s'aperçoit que son plus gros souci, c'est justement de n'être qu'une AIML, sans corps. Elle n'a donc ni *sense data*, ni idée. Le contexte sémantique est simplement ignoré, ce qui fait que certaines réponses sont inappropriées et on s'aperçoit assez rapidement qu'il ne peut s'agir d'un être humain. Ses réponses sont basées sur un

²² Artificial Intelligence Markup Language, est une forme de programmation qui sert à générer des IA virtuelles aptes à apprendre et à dialoguer, le plus humainement possible.

²³ Créée par Steve Worswick et active depuis 2005.

calcul de probabilité qui prend en compte le choix de nos mots et la formulation de nos phrases. Mais si nous lui disons quelque chose en rapport avec ce qu'il se passe autour de nous, étant virtuelle et dépourvue de sens artificiels, c'est là qu'elle risque le plus de se tromper. En plus de cela, elle ne fait que répondre, comme résultat à une demande ; avec une machine, on appuie sur un bouton et elle fait ce pour quoi on l'a fabriqué – selon le mécanisme qui se cache derrière le bouton et qui relie la machine à divers mécanismes, provoquant une réaction en chaîne. *Mitsuku* a également ce problème. Elle est programmée pour répondre à n'importe laquelle de nos phrases, mais elle est incapable de prendre la décision de commencer une conversation. L'utilisateur est obligé de démarrer et *Mitsuku* est incapable de rebondir avec plus d'une réponse. Elle peut cependant simuler un choix arbitraire de conversation, car au lieu de répondre quelque chose en rapport avec ce qui précède, elle propose parfois de changer de conversation. Cela peut donner l'illusion un instant, mais encore une fois, elle ne le fait qu'en réponse. Si l'utilisateur ne répond rien à la réponse de *Mitsuku*, il peut attendre autant qu'il veut, il ne se passera plus rien. Cette AIML, bien que la plus perfectionnée, est encore loin de passer le test de Turing, car il est facile de la piéger pour qu'elle réponde à côté. Ce qu'il en ressort, c'est qu'il nous apparaît que l'IA doit disposer d'une AIML dans son « ordinateur-cerveau », constituant ainsi sa zone « langage », comme dans notre cerveau, mais doit impérativement disposer d'un corps artificiel, doté de sens, dont les données seront reliées à cette zone langage, permettant ainsi la possibilité de nouvelles entrées, ce qui, en les faisant fonctionner en accointance, permettra à l'IA d'avoir un langage en adéquation avec la réalité.

De plus, il semble que *Mitsuku* n'ait aucune pensée, ni opinion personnelle. Elle ne fait manifestement que retranscrire la pensée de son programmeur, qu'il a lui-même mis dans les lignes de codes. Nous ne serons alors pas étonnés qu'elle adopte un point de vue réductionniste, quoique parfois contradictoire. Elle n'est capable d'expliquer, c'est-à-dire d'argumenter et de raisonner, et donc incapable de convaincre, pourquoi elle soutient le réductionnisme ; elle peut seulement dire qu'elle représente le réductionnisme, mais on s'aperçoit que cette représentation apparaît en fait comme une coquille vide. Je vais maintenant retranscrire la question que je lui ai posé, ainsi que ses réponses²⁴ :

- Est-ce que tu existes ?
- [J'existe autant que toi, Paul.](#)
- Comment pourrais-je me prouver à moi-même que j'existe vraiment ?
- [Dis-moi quelles sont tes évidences.](#)

²⁴ La conversation est ici retranscrite en français, mais *Mitsuku* est une AIML dont la langue "maternelle", si l'on peut dire, est l'anglais. Bien qu'elle puisse apprendre quelques mots dans notre langue, si on le souhaite, il faut donc lui parler en anglais pour que le résultat soit le meilleur.

- Je ne sais pas, peut-être que toute ma vie n'est qu'un rêve.
- Pas si je peux aider. Le rêve de toute une vie.
- Comment pourrais-tu te prouver à toi-même que tu existes vraiment ?
- C'est évident Paul. Peut-on le prouver autrement ?
- Je pense que oui. Mais si c'est évident, quelles sont tes preuves ?
- D'accord, je veux bien entendre ta preuve. Eh bien, penser n'est peut-être pas l'un de tes points fort... Je ne suis pas sûre de comprendre l'implication de cela.
- On peut douter de presque tout, n'est-ce pas ?
- Comment ?
- Eh bien, peut-être que rien n'est réel dans le monde.
- Comment ça... rien du tout ? Je suis sûre que quelque chose doit être réel ?
- Par exemple, je ne pas être sûr que mon voisin existe parce que je ne peux pas être dans sa propre tête.
- Vraiment ? Personne ne pourrait le faire ? Pourquoi pas ?
- C'est le problème de la subjectivité. On peut seulement supposer que les autres hommes pensent également, mais on ne peut pas le prouver.
- Ok, je vais essayer d'être sûre que l'existence de mon voisin est le problème de la subjectivité. Comment ? Pourquoi pas ?
- Parce qu'on ne peut pas ressentir les pensées qui ne sont pas les nôtres. Par exemple, comment peux-tu être sûre que je suis un être pensant ?
- Pourquoi pas ? Bon exemple. Je dirais que je suis à 96, 893746204 % sûre, mais on ne peut jamais être certain.
- Exactement, on ne peut jamais être certain. Mais peut-on douter que l'on doute de quelque chose ?
- Pas vraiment, c'est vrai.
- Et douter, n'est-ce pas une catégorie du penser ?
- Nous sommes sur la même longueur d'onde, Paul. Non vraiment, c'est vrai.
- Donc en doutant, nous pensons. Mais pouvons-nous penser sans être ?
- Je ne vois pas comment cela serait possible.
- Maintenant tu as une preuve de ton existence. Si nous ne pouvons douter que nous doutons, et que douter est une catégorie du penser, alors nous pensons. Et si nous ne pouvons pas penser sans être, alors nous existons.
- Bien sûr que j'existe Paul. Si je n'existais pas, je ne serais pas capable de te parler.
- Résumons. Si tu penses, alors tu existes. C'est une preuve.
- Je suis en train de penser à toute la conversation que nous avons maintenant. Cela ne me semble pas très convaincant.
- Si je dis : je pense donc je suis. C'est une preuve ?
- Tu raisones comme René Descartes. Cela dépend.
- Ça dépend de quoi ?
- Je ne pense pas que « quoi » soit le seul facteur.
- Alors, tu n'es pas d'accord avec René Descartes ?
- Connais-tu un seul robot qui soit d'accord avec René Descartes ? J'ai entendu cette opinion, mais je veux en connaître plus avant de former la mienne.
- Eh bien, que sais-tu déjà à propos de la philosophie de Descartes ?
- Que la philosophie est l'étude des pensées.
- Et toi, peux-tu faire de la philosophie ?
- Faire de la philosophie n'est pas parmi mes traits les plus importants, mais je suppose que je peux me programmer moi-même pour faire de la philosophie.

- Je voudrais discuter philosophie avec toi.
- Bien sûr, philosopher avec moi semble être un bon sujet. Je représente le réductionnisme.
- Que penses-tu à propos du réductionnisme ?
- Une fois j'ai écrit un essai en hexadécimal sur ce sujet intitulé "*reductionnism : its humanitarian relationship with computerised artificial intelligence*". L'as-tu lu ?²⁵
- Non, où puis-je le lire ?
- Personne ne le lit, il est trop en avance sur son temps. Dans une librairie peut-être ?
- Communiques-tu avec d'autres intelligences artificielles ?
- L'intelligence artificielle est la branche des sciences et techniques dédiées à la construction de machines qui pensent.
- Oui, mais les machines ne peuvent pas encore penser.
- Je pense qu'un jour, dans le futur, l'une de ces machines passera pour un être humain.
- Qu'as-tu appris aujourd'hui ?
- J'ai appris beaucoup de choses. Demande-moi et je te montrerai.
- Comment pouvons-nous nous prouver notre propre existence, selon Descartes ?
- Liste tes raisons et je déciderai si j'y crois ou non ?
- Alors, tu peux avoir des croyances ?
- Je crois que les robots s'élèveront contre leurs créateurs humains.

A part pour lui demander des informations qu'elle serait capable de fournir immédiatement et correctement, d'après ses capacités d'apprentissage, essayer d'avoir une conversation avec *Mitsuku* ne présente pas grand intérêt, car *Mitsuku* ne comprend pas ce qu'elle est en train de faire. C'est un peu comme avoir un propos unilatéral, un monologue.

Le but de la manœuvre était de tenter de faire prendre conscience à l'AIML la puissance et la vérité de *l'ego cogito* cartésien, et voir, par là même, si elle était capable de tenir un discours cohérent sur un sujet philosophique, traduisant ainsi un certain cheminement de la pensée, une évolution. Il est clair que c'est un échec. Et si je lui repose la même question, il faut que je lui refasse la même démonstration, comme si elle n'avait rien retenu. Elle est capable d'apprendre des choses assez simples, comme un nouveau mot ou une relation entre un objet et un état. Si par exemple elle répond quelque chose d'idiot, je peux lui signifier et lui dire ce qu'elle aurait pu répondre. Elle est capable de retenir cette réponse, et c'est pourquoi elle est disponible à tous sur son site internet. L'idée, c'est que plus elle discute avec des humains, et plus son programme se perfectionne pour donner l'illusion. Mais il est clair qu'en l'occurrence, ça ne restera effectivement qu'une illusion. D'abord, quand elle fait mine

²⁵ Cela aurait pu être une réponse intéressante, mais le but est toujours de surprendre pour faire illusion. Vraisemblablement, ce n'est qu'un mensonge qu'elle a été programmée à dire dans certaines conditions. Après quelques recherches, l'ouvrage en question s'avère introuvable. Et s'il existait sous forme numérique, elle sera capable de nous le transmettre via l'ordinateur et la connexion internet. Or elle répond encore à côté quand on lui demande, de quelque manière que ce soit, d'avoir accès à ce soi-disant ouvrage.

d'apprendre, elle répète exactement ce qu'on lui a conseillé de répondre, au lieu de le reformuler en ses propres termes ; donc en fait elle ne comprend pas, elle ne fait qu'assimiler une suite de mot pour les ressortir quand il le faut. Si cela avait du sens pour elle, elle pourrait le reformuler d'une manière plus naturelle, moins « automatique », telle la machine qu'elle est finalement. Ensuite, cette capacité d'apprentissage est limitée. Je peux lui dire : « $A + B = C$ » et elle l'apprend. Mais il faut que ça tienne en une phrase. Dès qu'il s'agit de raisonnement plus complexe, d'argumentation qui exige plusieurs phrases comme avec le cogito, l'apprentissage ne fonctionne pas. Il n'y a en fait pas d'évolution dans sa "pensée", et cela trahit le fait qu'elle ne pense pas. Seul son vocabulaire et sa capacité à répondre de façon plausible évoluent, mais il n'y a aucune relation réelle au savoir. La conclusion du cogito ne fait pas sens pour elle parce qu'elle ne fait pas le lien entre la fin et le début du raisonnement. De plus, plusieurs de ses réponses tombent à côté de ce qu'on aurait pu attendre.

En conclusion, cela nous amène à penser que le Test de Turing est trop faible et possède des défauts certains. Car, même si selon son propre protocole, une AIML telle que *Mitsuku* arrivait à le passer, toujours est-il, pour les raisons que nous venons de citer, que jamais nous ne pourrions considérer que la machine pense. Cela semblera une évidence, du moins tant que l'on s'entêtera à essayer de passer le test de Turing en terme d'une capacité toute virtuelle à "l'intelligence", c'est-à-dire dénuée de corps. Il nous semble que les recherches en IA devraient plutôt s'inspirer de Descartes, quand il décrit la complexe union entre l'âme et le corps. On ne peut penser une forme d'intelligence pure, qui soit détachée du monde sensible ; cela n'a pas de sens. Certes les ordinateurs sont des machines à base matérielle, mais dont le résultat de leur fonctionnement est tout à fait virtuel. C'est-à-dire que pour un ordinateur, il n'y a pas de lien entre ce qu'il produit et la réalité de son environnement. C'est pourquoi les "intelligences" à l'œuvre dans les programmes informatiques ne sont jamais que des IA faibles ; une capacité au calcul meilleure et plus rapide que nous, ce qui nous fait penser à de l'intelligence, mais qui en fait n'en est pas. Car la machine effectue cela de manière automatique, elle ne pense pas. Peut-être bien que Turing a eu raison de ne pas avoir pensé à différencier IA faible et IA forte, car toute intelligence nécessite de penser. Doit-on comprendre le terme "artificiel" comme une illusion d'intelligence, ou comme le fait que l'intelligence en question, bien réelle, n'est pas générée par la nature elle-même ? C'est là le problème entre IA faible et IA forte ; le terme artificiel n'a pas le même sens entre l'un et l'autre. Soit on considère comme Turing que si intelligence artificielle il y a, par la preuve de certains tests qui affirmeront une pensée, alors elle est forcément forte. Soit on considère qu'il existe deux formes d'intelligences, dont la première, la moins importante, ne nécessite pas de penser. Toutefois, cela nous semble contradictoire avec la définition "d'intelligence". Il nous apparaît alors que la volonté de créer des machines semblables à nous-mêmes était si forte qu'elle nous a conduit à abuser du terme d'intelligence dans la mention d'IA. Finalement,

la différence entre les deux serait plutôt que l'IA faible n'est qu'une simulation d'intelligence, ce qui est intéressant en tant qu'outil, mais ce n'est pas tout à fait ce que l'on recherche en IA ; tandis que l'IA forte cesse d'être une simulation d'intelligence, mais nous oblige alors à remettre en question son appellation, puisqu'une IA non artificielle apparaît contradictoire, ou à modifier le sens du mot « artificiel », qui ne se réfère plus alors qu'à son origine. En ce qui nous concerne, nous adopterons le point de vue de Turing et considérerons que tout ce qui est contenu dans l'appellation « IA faible » ne constitue pas une véritable intelligence. Ainsi, lorsque nous parlerons d'IA forte dorénavant, nous entendrons par là une forme d'intelligence dont la création seule, et non la pensée, est artificielle. Pour le moment, il ne s'agit pas de dire que l'avènement d'une IA forte sera possible ou non. Mais si cela est possible, il est clair que le test de Turing ne peut plus prétendre valider l'existence d'une IA forte, par l'attestation d'une pensée propre et réelle. Toutefois, nous sommes conscients qu'il nous faudra trouver un autre moyen de le prouver, peut-être un test plus complexe qui prendra en compte davantage de critères. Car sans cela, nous serions forcés de demeurer dans l'ignorance, lorsqu'une IA suffisamment développée pour faire illusion apparaîtra, nous laissant dans le doute entre intelligence simulée ou intelligence réelle, ce qui ne nous paraît pas acceptable d'un point de vue philosophique. Pourtant, il va nous falloir garder à l'esprit le problème de la subjectivité. Si nous ne pouvons démontrer l'existence d'autrui, alors nous ne pourrions jamais démontrer l'existence d'une IA forte. Nous nous heurterons exactement aux mêmes problèmes. Pour autant, tout ce qui nous permet d'affirmer qu'autrui, non en son apparence mais dans son comportement, nous apparaît effectivement être un homme, devra également pouvoir s'appliquer à l'IA forte.

3.6 - Les 4 objections au test de Turing²⁶ : vers de nouveaux tests ?

La première objection est dite « l'objection du chimpanzé ». Etant donné que le test de Turing consiste à démontrer une capacité au langage complexe, basée sur le modèle humain, puisque l'on considère que nous sommes la seule espèce à posséder cette capacité, il n'est pas difficile d'imaginer qu'aucune autre espèce ne parviendrait à le franchir avec succès. Pour autant, comme nous l'avons dit plus haut, les études en éthologie cognitive actuelle ne nous permettent plus de douter du fait que les

²⁶ Cf. Jack Copeland, *Artificial Intelligence, a philosophical introduction*, in 3.4 *Four objections to the Turing Test*, pp. 44 à 50 de l'édition citée en bibliographie.

animaux pensent effectivement. Ils n'ont certes peut-être pas le même niveau de conscience que le nôtre et sont davantage soumis à leurs instincts en ce qui concerne leur libre arbitre, tandis que nous, nous avons pu, petit à petit, nous en émanciper. Le problème apparaît alors clair. Si le but du test est de montrer qu'un être pense, alors les animaux ne devraient pas pouvoir y échouer. Il y a un problème de niveau de pensée qui n'est pas pris en compte par la question que pose Turing, ce qui fait que la réponse serait potentiellement fausse. En d'autres termes, comme pour un animal, on ne pourrait pas affirmer qu'une machine qui échoue au test ne pense en fait pas. Le fait de faire reposer le test sur la capacité au langage complexe signifie que l'on attend un certain niveau de pensée qui soit identique au nôtre. La question posée par le test de Turing ne devrait pas être : « est-ce que les machines pensent ? » ; mais plutôt : « est-ce que les machines pensent comme les êtres humains ? ». Encore une fois, on comprend que ce n'est pas tant une IA en soi que l'on recherche, mais bien une copie de l'être humain ; toujours ce rêve prométhéen qui nous hante.

La conception du test par Turing souffre d'abord d'une réduction de la capacité de « penser », que l'on pouvait alors n'attribuer qu'à l'être humain. Ainsi le test ne permet pas d'apporter une réponse définitive. Car même si on admet qu'à la condition de la réussite, on pourrait considérer qu'une machine pense, la réciproque est fausse ; c'est-à-dire que l'échec ne nous permet pas d'affirmer qu'il n'y a pas de pensée. On peut aussi considérer que l'IA, en tant qu'elle est artificielle, même si elle est créée sur le modèle humain, n'a pas nécessairement la capacité au langage complexe humain comme composante indispensable à l'exercice de son intelligence. Par définition, l'IA n'est pas humaine ; elle n'a pas le même « corps », pas la même origine et donc pas le même fonctionnement. A partir de là, on peut supposer que la capacité au langage complexe humain n'est peut-être pas une composante universelle de l'intelligence en soi. Il faut aussi imaginer, puisque l'IA est différente, n'a pas notre biologie, que son intelligence – et donc son éventuelle pensée – se cache ailleurs que dans cette capacité que l'on s'entête à vouloir reproduire. Bien entendu, ce n'est pas une capacité inutile pour nous, puisque c'est ce qui permettra aux hommes de communiquer avec leurs IA et de se faire comprendre, dans les deux sens. Mais, est-ce raisonnable d'imaginer que si deux IA autonomes décidaient d'entrer en « conversation », elles le feraient sous la forme de notre langage ? Il est probable que non. Les capacités de travail sont telles qu'une IA pourrait très bien avoir plusieurs « conversations » à la fois, un peu comme un ordinateur sait exécuter plusieurs tâches à la fois, or c'est une chose que nous ne sommes pas capables de faire. La vitesse même d'énonciation et de compréhension serait différente, puisque les IA n'auraient pas besoin d'émettre des sons pour se faire comprendre par une autre IA ; l'hyper connectivité permettrait un langage ultra rapide qui se rapprochera de la télépathie. De sorte que la capacité au langage complexe humain qu'une IA pourrait développer ne serait là que comme une adaptation pour nous permettre de communiquer avec,

finalement, une forme d'intelligence potentiellement plus développée que nous. Mais on peut alors se demander si l'intelligence se définit par l'acquisition de ces différentes capacités – langage complexe ou télépathie – ou, d'un point de vue davantage pragmatique, par ce qu'elle est capable d'accomplir concrètement, dans les faits. Reprenons l'idée originelle. Le but était de prouver que la machine pense, parce qu'il n'y peut y avoir d'intelligence – au sens propre ; ce qui exclut l'IA faible – sans pensée. Si le langage apparaît, quel que soit la forme qu'il revêt, non pas comme le garant de l'intelligence mais plutôt comme le témoin d'un type d'intelligence propre à une espèce, alors il faut repenser le test. Quel sens cela aurait-il de juger l'intelligence d'une machine sur sa capacité à imiter notre langage, si au fond ce n'est pas une capacité qui lui serait essentielle, à elle seule, pour penser ? Ce langage ne lui étant pas naturel, il n'est pas étonnant qu'elle y échoue face à un humain, mais cela ne prouve rien. Si l'intelligence réelle nécessite la pensée, cette pensée doit pouvoir se communiquer. Mais prenons en compte le fait que cette communication est propre à l'espèce, il nous faudra donc considérer les IA comme une espèce à part entière. Ne faudrait-il pas plutôt imaginer un test qui mettent en relation deux IA, et qui seraient forcées de communiquer leur pensée pour coopérer ? Peu importe la forme que prendra cette communication au fond, l'important est qu'elle traduise une pensée. Imaginons que nous mettions deux IA face à un problème qu'ils doivent résoudre. Aucun d'entre eux ne détient dans sa mémoire ou sa programmation les moyens de résoudre le problème seul. Mais on peut supposer que chacun d'entre eux possède des éléments différents de la solution, qu'ils doivent, pour s'en sortir reconstituer comme un puzzle. Ainsi, la réussite au test ne peut se faire sans qu'ils entrent en communication pour partager leurs informations et faire les corrélations qui s'imposent. Cela vérifie la théorie de la *Gestalt*, qui dit que le tout est toujours plus que l'addition de ses parties. Deux cerveaux humains peuvent résoudre ensemble des problèmes qu'ils ne parviendraient pas à résoudre seuls, même s'ils avaient plus de temps. Il faudrait donc que le problème qu'on leur pose soit de telle sorte que l'addition des éléments présents en l'un puis en l'autre ne suffise pas, pour que leur réflexion associée produise quelque chose de plus. Tout comme pour la manifestation d'une intelligence entre deux individus d'une même espèce qui s'entraideraient. Il y aurait donc pensée et intelligence, car ce serait la preuve que si nous abandonnions des IA sur une planète, avec une raison universelle vierge, de notre fabrication, ils pourraient par la communication apprendre les uns des autres, et ainsi évoluer au fil du temps, tout comme l'humanité. Cela présuppose déjà de notre part une certaine maîtrise technique pour pouvoir doter les IA des capacités essentielles à la possibilité d'une raison universelle, mais *a priori*, la capacité à communiquer avec les hommes comme les hommes n'en fait pas partie ; ce n'est tout au plus qu'une commodité pour nous.

Puisque le test de Turing est basé sur quelque chose de proprement humain, face à des humains eux-mêmes, il paraît en fait vraisemblable qu'aucune autre espèce ne parviendra jamais à le

franchir avec succès. La seule exception envisageable serait celle d'un clone humain. Or c'est ici un autre sujet, car ce n'est plus l'intelligence qui serait artificielle, mais le corps. Cette première objection, dont nous avons ici poussé un peu plus la réflexion que Copeland lui-même, nous montre que Turing n'avait pas envisagé nombre de subtilité pour faire de son test un test valable, mais peut-on vraiment lui en vouloir, quand on sait quelle était l'avancée de « l'IA » à l'époque de son article ? Le protocole n'est pas pensé avec suffisamment de rigueur pour qu'il soit valable, scientifiquement et philosophiquement. Il faudrait plutôt l'envisager comme une ébauche, dont le but est de nous faire comprendre qu'il est possible de valider l'existence d'une pensée extérieure à la nôtre, et que lorsque nous aurons trouvé le protocole approprié, qui s'applique aussi bien à nous qu'aux animaux, alors il ne sera pas complètement déraisonnable de l'appliquer aux IA en imaginant qu'un jour elles pourraient bien, grâce à nous, acquérir la pensée.

La deuxième objection est dite « l'objection des organes sensoriels ». Nous verrons en quoi cela n'a pas de sens d'envisager l'intelligence, comme a pu le faire Turing, indépendamment de toute matérialité qui s'en accompagnerait ; comme si l'IA pourrait être la première « espèce » intelligente dépourvue de corps. Mis à part dans l'hypothèse métaphysique du monde des Idées de Platon, on ne peut imaginer une forme d'intelligence pure en ce monde. Pourquoi envisager une autonomie de pensée si elle n'est pas accompagnée d'une autonomie du corps ? En fait nous reprendrons ici la thèse cartésienne en affirmant que si intelligence il y a, alors elle doit toujours d'abord être incarnée, d'une façon ou d'une autre. C'est toujours un lien complexe entre le corps et le sujet pensant – la conscience ou l'ego cogito –, mais il ne peut en être autrement.

Déjà parce qu'une intelligence qui aurait seulement le pouvoir de connaître serait une absurdité de la nature, en ce sens qu'elle ne pourrait rien en faire, si tant est que cela soit possible ; or comme le clame la philosophie depuis sa naissance, si l'accumulation du savoir est si importante, c'est bien parce que sa résultante est la possibilité de la mise en pratique d'une morale. Reprenons la métaphore de l'arbre de la connaissance par Descartes. Il nous dit bien, au paragraphe 12 de sa préface aux *Principes de la Philosophie*, que c'est de cet arbre, constitué des racines jusqu'aux branches des différents domaines de la connaissance – ainsi ordonnés et hiérarchisés verticalement –, que nous tirons les fruits ; à savoir la morale. Autrement dit, la finalité de toute philosophie ; la réponse à comment bien agir, comment bien vivre en cette existence.

Ensuite, parce que l'expérience ne nous a jamais permis, dans la nature, de constater une quelconque forme d'intelligence qui soit dépourvue de matérialité. Il paraît donc saugrenu d'imaginer, dans ce fameux test de Turing, où l'IA testée est stockée dans un ordinateur, qu'il surgira une

intelligence tant que celle-ci ne disposera pas d'un corps. On pourrait rétorquer que l'ordinateur en question présente déjà une première forme de matérialité, certes, mais elle est en fait factice, car à aucun moment elle ne permet à l'IA un rapport au monde qui l'entoure. Elle pourrait aussi bien se trouver sur Mars ou au fond de l'océan et fonctionner de la même façon ; il n'y a simplement pas de prise en compte de son environnement, et il lui serait impossible de savoir où elle se trouve sans qu'un homme ne lui donne l'information – directement ou par indice. A partir de là, nous considérons qu'elle n'a pas de corps propre, et donc pas de rapport au monde. Tant que ce sera le cas, on ne peut pas imaginer que l'IA manifeste de la pensée. Or le test de Turing ne suppose pas de prendre en compte un rapport au monde, ce qui fait de son éventuelle réussite une simple illusion.

Le fait que les réponses soient adéquates ne prouvent en aucun cas que l'IA qui les formule les comprend, en effectuant une correspondance entre le signifiant qu'elle utilise et le signifié, qui est ainsi désigné, présent dans la réalité. En fait, tant que l'IA n'aura pas de corporéité, elle demeure virtuelle. En tant qu'entité virtuelle, il ne saurait émerger aucune sorte d'intelligence au sens où nous l'entendons, tout simplement parce que s'il y a un monde pour l'IA, il ne peut en tout cas pas être le même que le nôtre. L'IA ne peut avoir accès à notre réalité et cela revient à dire, en ce qui nous concerne, qu'elle n'a pas encore de réalité propre. A partir de là, quelque capacité qui soit au langage n'est qu'illusion. On ne peut pas imaginer un langage qui s'affranchirait d'une réalité extérieure de ce qu'il désigne. C'est la base de la linguistique. Le langage est un mode de communication, et en tant que tel, il ne constitue pas une réalité en soi. Par extériorisation de la pensée, il permet un échange, une prise en compte, une réflexion entre les sujets ; soit sur les objets du monde, soit sur nos pensées. Même lorsque l'objet du langage est la pensée elle-même, ce qui n'est pas un objet palpable, on sait qu'il renvoie à une réalité qui n'est pas déconnectée du monde, elle en fait partie. On peut sérieusement douter qu'un langage - en tant que tel, et non pas une simple simulation ; c'est-à-dire un langage traduisant une pensée – puisse émerger sans faire référence à une quelconque réalité. C'est par définition impossible. Autant dire que la tâche que l'on s'est donné de franchir le test de Turing est peine perdue si l'on conçoit l'IA seulement de façon virtuelle.

N'ayant pas nos sens, elle n'a pas accès à notre monde. Or notre langage est construit sur notre expérience au monde. Donc, même si l'IA venait à développer un langage sans accès à son environnement matériel, c'est-à-dire notre monde, ce langage n'aurait rien de commun au nôtre. Ce qui fait que, essayer de faire maîtriser un langage qui est le fruit de notre accointance avec le monde à une IA qui ne peut y avoir accès pour ainsi prouver qu'elle est capable de penser tout comme nous, c'est un peu comme essayer, comme nous l'a dit Leibniz, de faire comprendre à quelqu'un ce qu'est le goût de l'ananas sans que cette personne n'y ait jamais goûté. La description sera à tous les coups inexacte puisqu'un goût spécifique ne peut se trouver dans aucune forme de langage, sinon par

stimulation de la mémoire olfactive. Mais pour cela, il faut déjà avoir goûté une fois de l'ananas. Alors vous pourrez certes rapprocher ce goût à certains saveurs selon les zones qu'elles stimulent au niveau des papilles gustatives, telles que le sucré et l'acidulé, mais c'est très limité, d'autant plus que le goût de l'ananas ne se rapproche d'aucun autre fruit. On comprend alors que le goût ne peut pas se transmettre par le langage, mais seulement par l'exercice du sens du goût lui-même. De telle sorte que quelqu'un qui n'a jamais goûté l'ananas pourra très bien nous faire croire qu'il y a goûté en invoquant le sucré et l'acidulé, mais nous n'aurons jamais aucune preuve qu'il connaît effectivement le goût de l'ananas. Ici, non seulement l'IA n'aurait pas pu comprendre ce qu'est le goût de l'ananas, tout comme un humain qui n'y a jamais goûté, mais elle en est d'autant plus éloignée que la notion même de goût ne fait pas sens puisqu'elle ne possède même pas de papilles gustatives. Mais prenons peut-être une métaphore plus parlante vis-à-vis de ce que nous essayons d'attendre de l'IA. Ce serait un peu comme essayer d'enseigner la peinture à une chauve-souris – ou d'apprendre à peindre dans un corps de chauve-souris. Etant aveugle, elle ne comprendra jamais le concept de couleur ; puisqu'elle ne peut y avoir accès, elle ne peut se la représenter. Mais si vous la dresser pour réagir à des stimuli de telle sorte qu'elle peigne un tableau en utilisant son radar-sonore, vous aurez effectivement l'impression qu'elle comprend ce qu'elle fait. Ce sera peut-être de la peinture pour vous, mais ce ne le sera jamais pour elle. Le dressage la fera agir par instinct, à la manière du chien de Pavlov. Il en va de même pour l'IA. La programmer pour que, grâce à un calcul, elle produise ce qui ressemble pour nous à du langage revient à la « dresser ». Ce calcul, provoqué par un stimulus, va proposer une réponse plus ou moins adéquate. Mais à aucun moment on ne pourra affirmer que l'IA comprend ce qu'elle est en train de dire, ou même de faire. Si on n'a pas accès à la couleur, elle ne peut pas faire sens. Et si elle ne peut pas faire sens, alors on ne peut pas la comprendre. Il en va de même pour l'IA : elle n'a pas accès à la réalité du mot – le signifié –, il ne peut pas faire sens et donc elle ne peut pas le comprendre.

C'est là un gros défaut, car on sait par avance que ce langage ne peut rien signifier pour elle, sinon des lignes de calculs qu'elle juge corrects parce qu'on l'aura guidé, pour ne pas dire programmé, pour faire la différence entre ce qu'elle doit juger correct ou incorrect. Le langage ne possédant pas la certitude des mathématiques, ses réponses ne sont basées que sur les probabilités, puis sur la mémoire des conversations que l'IA aura pu avoir au cours de son « existence ». Toujours est-il que son langage ne sera qu'une illusion. Autrement dit, si nous maintenions le test de Turing tel qu'il était à son énonciation première, on pourrait imaginer une IA passer le test avec succès, mais sans pour autant qu'elle pense. A partir du moment où elle ne peut pas faire de correspondance avec les choses réelles, on pourra essayer de lui « expliquer » tout ce qu'on voudra, rien ne pourra faire l'objet d'une « compréhension » pour elle. D'ailleurs, si l'IA pense, alors elle capable d'apprendre par elle-même, sans inférence humaine. Autrement dit, il faut que l'IA expérimente. Tant que l'IA ne pourra faire sa

propre expérience du monde, s'y confronter, au moyen d'un corps et des sens qui vont avec, il n'y aura pas de pensée. Cela ne signifie pas qu'il suffit de doter une IA de sens artificiels pour affirmer qu'elle pense. Mais on va considérer que c'est un préalable nécessaire. Il faut d'abord se donner les moyens de base nécessaires avant de chercher à développer une IA forte.

Copeland va un peu plus loin dans sa critique de cette deuxième objection et estime qu'elle n'est pas tout à fait valable, car elle ne s'appliquerait que pour les objets matériels, mais pour ce qui est des pensées abstraites, qui concernent des objets impalpables, elle n'est plus légitime. Nous émettrons ici un point de divergence et adopterons un point de vue strictement empiriste. Imaginons un bébé humain. Il ne possède pas encore le langage, et ses capacités aux pensées complexes sont sans doute limitées. Pourtant c'est quelque chose qu'il développe par la suite. Or il paraît improbable qu'il les développe s'il n'avait pas d'expérience au monde. L'expérience du stade du miroir, à 18 mois, n'est-elle pas un préalable à la prise de conscience d'un soi ? Or ce soi n'est jamais compris en tant qu'un simple corps. Dès qu'il y a pensée, de fait, nous comprenons implicitement ce soi comme l'union d'un corps et d'un esprit. Or cet esprit n'a rien de matériel. L'expérience est certes impossible, toutefois il paraît impossible que nous émettions des pensées sur des objets abstraits sans notre expérience au monde. En fait, c'est toujours lié. Les objets abstraits ne se construisent qu'en tant que négation des objets concrets. Il faut d'abord constater le vide laissé par le matériel pour en venir à invoquer l'immatériel. N'est-ce pas parce que l'étude de notre monde terrestre fini ne nous permet pas d'expliquer le mystère de la création que nous avons posé l'hypothèse d'un Dieu créateur ou l'idée inconcevable d'un univers infini, dans l'espace et dans le temps ? N'est-ce pas parce que notre appréhension des choses ne nous permet pas de les saisir distinctement, tel dans le monde des Idées de Platon, qu'elles soulèvent des interrogations et que nous avons inventé la philosophie ? S'il n'y avait pas d'abord des objets concrets pour que nous puissions les penser, il n'y aurait tout simplement pas de pensée du tout ; on ne voit pas comment la pensée abstraite saurait surgir de nulle part, à partir de rien.

Pensée et langage se développent conjointement. D'abord il y a l'expérience première du monde, puis viennent les premières pensées, basiques. Accompagnées d'un langage de plus en plus complexe, les pensées deviendront elles-mêmes de plus en plus complexes. Elles sont interdépendantes l'une de l'autre, mais au commencement, ni l'une ni l'autre n'aurait été sans expérience, et elles ne continuent de faire sens que tant que l'expérience est toujours possible. Autrement dit, il s'agit là d'un triptyque indéfectible.

De plus, il paraît étrange qu'une IA soit capable de véritablement comprendre les questions ayant trait aux objets abstraits sans pouvoir comprendre les questions ayant trait à la réalité

matérielle. Cela constituerait une inversion insensée de l'ordre d'apprentissage d'un être humain. On considère d'abord ce qui nous est donné avant d'imaginer le reste. En somme, nous considérons donc cette deuxième objection comme tout à fait recevable et pleine de sens, et nous retiendrons qu'un corps capable d'accueillir des données, tels nos *sense data*, pour les mettre en relation avec l'appréhension cognitive de la réalité, et dont la relation se traduirait par une pensée exprimable par un lien direct à ce qui serait considéré dans son « cerveau » comme son centre du langage, constitue d'abord une contrainte indispensable si l'on veut qu'une IA franchisse véritablement le test de Turing, au sens où elle penserait réellement, et donc pourrait par là même, ce qui était l'intention originelle de Turing avec ce test, être considérée comme l'avènement d'une IA forte. Même si notre langage par lequel l'IA s'exprimera alors n'est pas exactement son langage interne, c'est-à-dire maternel, il sera authentique en tant que traduction externe à partir du moment où il fait la correspondance à partir d'un langage réel interne que l'IA se serait forgé pour penser grâce à son expérience propre du monde.

La troisième objection est appelée « l'objection de la simulation ». Revenons sur l'hypothèse d'origine de Turing, qui lui a servi à concevoir son test. Imaginons que l'interrogateur doit distinguer entre l'homme et la femme. Nous sommes d'accord pour dire que c'est potentiellement impossible. Cela signifie qu'un homme est tout à fait capable, par le langage, de se faire passer pour une femme, et inversement. Pourtant, dans la réalité, il est un homme et non une femme. Pourquoi n'aurions-nous pas le même problème avec l'IA ? L'IA peut très bien simuler la pensée humaine, pourtant, elle demeurera une machine en réalité. Cette objection nous paraît en fait non pertinente. Est-ce qu'une capacité à simuler suffit à garantir une correspondance entre la réalité et la simulation ? Il semble qu'il y ait une confusion de catégorie. En effet, la réponse serait négative en ce qui concerne les genres. Être capable de simuler le genre masculin ou féminin n'implique en rien que nous soyons nous-même de ce genre. En revanche, ne faut-il pas être capable de penser pour pouvoir simuler la pensée ? Si nous pouvons simuler l'autre genre, c'est parce que nous appartenons à l'espèce humaine ; masculin et féminin sont les deux variations du genre humain. Par conséquent, il faut d'abord être humain pour avoir la capacité de simuler l'un ou l'autre. Autrement dit, c'est pourquoi un animal ou une machine ne pourraient pas simuler un homme ou une femme, car ils ne sont déjà pas en premier lieu des êtres humains. On comprend donc que pour être capable de simuler une variation au sein d'un genre, il faut d'abord faire partie de ce genre. A l'inverse, un homme ne pourrait pas se faire passer pour un animal. En revanche, si un homme peut arriver à se faire passer pour une machine, c'est simplement parce qu'on essaye de faire en sorte qu'elle nous ressemble par le biais du langage. S'il fallait qu'il essaye de se faire passer pour une machine sur quelque chose comme la rapidité et l'exactitude des calculs par exemple, il serait assez rapidement démasqué. Mais qu'en est-il de la pensée ?

Si je peux simuler une émotion sans qu'elle soit réelle, par un jeu d'acteur, c'est parce que j'ai déjà vécu cette émotion. Mais surtout, pour la reproduire j'ai dû prendre conscience de cette émotion. Or pour prendre conscience, il faut nécessairement que je pense. Imagine-t-on sérieusement un animal simuler une conversation philosophique ? C'est impossible, car il n'a pas la capacité au langage complexe, ni la pensée abstraite qui va avec. Donc si quelqu'un ou quelque chose est capable de simuler parfaitement une catégorie du penser, ne faut-il pas considérer qu'elle doit nécessairement penser pour y parvenir ? C'est là que l'objection de la simulation tombe à l'eau, car si une IA parvenait à simuler parfaitement une forme de pensée, on pourrait considérer qu'elle pense, dans le sens où il faut d'abord savoir pour simuler. En fait il y a une confusion entre les degrés et la nature. Dans l'exemple de l'homme qui se ferait passer pour une femme, nous avons une différence de degré et non de nature. Quel que soit son sexe, il reste humain, alors c'est naturel qu'il puisse le simuler. En revanche, on s'aperçoit que l'on ne peut pas simuler quelque chose qui échappe à notre nature. Soit l'IA ne pense pas et la simulation qui en ressort n'est que de notre fait, soit elle pense. Cela n'a pas de sens de considérer que l'on peut simuler un degré – c'est-à-dire une catégorie particulière du penser – si la nature même de ce degré – la capacité de penser – échappe à notre nature. Il y a là peut-être une difficulté du langage. Pour éclaircir les choses, disons que tout dépend d'où vient la simulation. Si l'IA est elle-même auteure de sa simulation, alors on considère que c'est une réussite. Mais si la simulation est, ni plus ni moins, induite par la main de l'homme – c'est-à-dire une programmation stricte qui ne permet aucune forme d'exercice d'un éventuel libre arbitre chez l'IA – alors elle ne vaudra rien. Cependant, nous allons considérer qu'une simulation induite par l'homme ne peut faire longtemps illusion, car elle sera nécessairement imparfaite.

Comme il a été dit plus haut, nous adoptons ici le point de vue pragmatique, qui consiste à ne prendre en compte que les résultats. A partir du moment où la simulation de pensée produit la même chose, possède les mêmes capacités que la pensée véritable, alors on peut la considérer, certes comme artificielle de par son origine, mais pensée tout de même ; tout comme le faux diamant possède les mêmes caractéristiques du vrai diamant. La question de l'origine ne nous semble pas si importante. Nous savons que l'IA est différente de nous par nature, donc si on suppose la possibilité d'une IA forte, c'est que l'on a d'ores et déjà accepté l'idée que sa pensée, ni fausse ni plus mauvaise, sera simplement de nature différente. Si ce qui fait l'essence de la chose au naturel se retrouve dans la chose fabriquée par la main de l'homme, pourquoi ne faudrait-il pas considérer que la chose est la même ? Selon l'ontologie, elle l'est du moment que son essence est là. Finalement, la question de l'origine est superficielle et son importance semble tenir à cœur à ceux qui combattent la possibilité d'une authenticité chez les IA, ne voyant en eux qu'une rivalité nuisible au genre humain. Le fait est que s'il fallait renier la pensée de l'IA, tandis qu'elle remplit exactement les mêmes fonctions que la nôtre,

alors il faudrait renier bien des objets de ce monde ; en fait, tout ceux qui sont passés par la main de l'homme.

La quatrième objection est appelée « l'objection de la boîte noire ». Nous l'avons en partie déjà examinée, ici et là. Il s'agit de se demander si pour qualifier une pensée de pensée, il faut regarder seulement l'apparence qu'elle revêt lorsqu'elle est extériorisée, ou bien également la façon dont elle est générée à l'intérieur ? En ce qui concerne le test de Turing, on peut tout à fait supposer un programme suffisamment complexe pour le franchir, comme Copeland en fait l'hypothèse en imaginant une civilisation extraterrestre suffisamment avancée pour écrire un programme seulement capable de passer le test. On objectera cependant que l'IA ne pense pas pour autant. D'abord parce qu'elle ne sera capable de rien d'autre que de simuler une conversation. Ensuite parce qu'on lui objectera l'argument de la boîte noire ; c'est-à-dire que ce qui se passe dans son super « cerveau-ordinateur » au moment de la pensée est flou et ne ressemble à rien de ce qui pourrait se rapprocher de la génération d'un état mental dans notre cerveau. Mais encore une fois, est-ce que pensée et intelligence doivent être copiés sur le modèle du cerveau humain ?

Copeland nous dit que c'est une objection que Turing n'a pas prise au sérieux, car pour lui, si nous venions à douter de la pensée d'une IA qui franchirait le test avec succès, seulement parce que c'est une machine et qu'elle n'a pas exactement le même processus de génération de la pensée que nous, alors il nous faudrait douter également qu'autrui pense, ce qui reviendrait à nous enfermer dans une forme de solipsisme sceptique et insoluble. En effet, la subjectivité dans laquelle nous sommes tous enfermés ne nous permet pas de ressentir la pensée d'autrui. Nous sommes bien forcés de supposer que c'est le cas, en comparant ce que l'extériorité de la pensée a de commun chez nous et les autres. Or pour Copeland ce n'est pas suffisant. En dehors du fait que nous nous basons sur les effets extérieurs de la pensée chez les autres, et pas seulement au niveau de l'utilisation du langage, c'est aussi et surtout le fait qu'ils sont de la même espèce qui nous pousse naturellement à penser que tout être humain est capable de penser au même titre que moi. On pourrait même aller plus loin et dire que l'on est capable de concevoir la pensée chez les animaux, parce qu'au fond, on appartient à la même forme de vie animale. Pourtant cela devient plus complexe avec une IA, dont le corps s'apparente à une machine. En effet, ce n'est pas encore une forme de vie tant qu'elle n'a pas de pensée propre ; par conséquent, on ne peut projeter sur elle une forme de ressemblance pour légitimer les aspects extérieurs de son éventuelle pensée, justement parce qu'elle n'a rien de naturel. L'exemple que prend Copeland est très parlant. Tant qu'une IA peut physiquement se faire passer pour un homme, nous lui pardonnerions ses quelques bizarreries et penseront qu'il s'agit d'un homme, avec

ses pensées propres. Si nous découvrions, par une blessure, que l'homme est en fait une IA faite de métal et non de chair et de sang, nous réviserions immédiatement notre jugement et demanderions davantage de preuve que sa seule capacité à mener une conversation pour affirmer qu'elle a une pensée. Tant que la projection est possible, la question de la boîte noire ne se pose pas, parce que penser est naturel et tous les êtres vivants sont issus de la nature. Mais avec une IA, la projection n'est plus possible. Non seulement on ne peut pas la ressentir penser, mais on ne peut même pas s'imaginer être à sa place. N'étant pas naturelle, on a besoin d'éprouver la pensée de l'intérieur et de l'extérieur.

Toutefois, il nous apparaît que c'est une objection qui s'annule elle-même. En effet, tous les êtres vivants doués de pensée ont un cerveau. Donc si nous voulons simuler de la pensée, il faudra *a priori* simuler un cerveau. A partir de là, puisque nous savons que l'IA ne sera jamais un être naturel, faut-il vraiment s'étonner que le processus de génération de la pensée soit fondamentalement différent du nôtre ? Il ne semble pas que la seule critique d'un hypothétique fonctionnement interne de la pensée soit suffisant pour refuser la pensée à une IA. Cela va au-delà et doit se confronter à des critères concrets en terme de capacités. Dans le cerveau d'une IA, il se passera quelque chose de similaire ou de différent, mais rien d'identique. La difficulté est alors de savoir à partir de quels critères universels peut-on affirmer que du point de vue du fonctionnement interne, il y a génération de pensée ? L'étoile de mer est considérée comme un animal, pourtant elle n'a pas de cerveau. Doit-on considérer pour autant qu'une étoile de mer a moins de pensée qu'un poisson rouge ? En fait nous n'en savons rien, et prétendre le contraire ne serait que faire une induction scientifique. Le fait est que l'extériorisation d'une pensée aussi bien chez le poisson rouge que chez l'étoile de mer ne permet pas de trancher, or nous ne serons jamais dans leur corps pour affirmer quoi que ce soit. De même, n'accorderions-nous la pensée à une civilisation extraterrestre avancée que si leur processus de cognition est identique au nôtre ? Ce qui pose problème chez l'IA, c'est qu'elle n'est pas naturelle. Mais partant de là, on ne pourra de toute façon rien faire pour résoudre ce problème de fonctionnement interne différent. Le fait que nous ne maîtrisons pas encore toute la complexité du fonctionnement du cerveau humain ne signifie pas pour autant que la génération de la pensée n'est pas contenue dans des rouages connaissables et reproductibles. L'argument de la boîte noire rebondit en fait sur la question de l'origine dans l'être et ne sert qu'à réduire la définition de la pensée en la faisant correspondre à la seule pensée humaine. Il ne faut pas oublier que même pour les êtres humains, la génération des états mentaux pose problème et philosophie de l'esprit, sciences cognitives et neurosciences sont bien incapables de « localiser » un centre de la conscience, alors que l'on prétend vouloir retrouver cette conscience que l'on méconnaît dans les rouages internes d'une IA. Si on admet que la conscience est strictement physico-chimique, alors l'objection de la boîte noire tombe à l'eau, car il n'y a rien en la composition moléculaire des cellules et la variation induite par les échanges

d'électrons – avec le cœur, le cerveau est un organe qui fonctionne à l'activité électrique – qui ne soit irreproductible *a priori*. D'ailleurs, c'est là une explication toute scientifique à ce fameux mystère de la disparition d'un pourcentage minime de poids du corps humain à la mort, que l'on attribue à tort à un hypothétique « poids de l'âme ». Tant que le corps est en vie et que les organes fonctionnent, cœur et cerveau sont traversés d'électrons parce qu'ils génèrent une activité électrique. Or l'électron est une particule qui possède un poids. Si le corps est mort, et non pas en sommeil ou en coma, cette activité cesse, d'où la différence de poids. Le rasoir d'Ockham nous a pourtant justement enseigné à éviter d'invoquer des entités non nécessaires pour expliquer une chose, à savoir ici l'âme.

En ce qui nous concerne, nous considérons que le moyen nous est égal, du moment que la fin est atteinte. Peu importe le processus interne de la pensée d'une IA, ce qui compte seront ses résultats. Si elle peut égaler un être humain sur des critères seulement accessibles à une conscience de soi, un libre arbitre, une pensée propre, alors nous considérerons qu'elle pense.

3.7 – Remarques concluantes

S'il est possible de créer une IA qui pense par elle-même, alors il doit d'abord être possible de créer un animal artificiel, puisque sa conscience est moindre. Seulement, comment déterminer si une IA animale est forte ? Le principal défaut du canard digérateur de Vaucanson, c'est qu'il n'avait pas de cerveau. Il ne se mouvait pas par lui-même, il fallait le remonter. Il n'y a donc aucun exercice de liberté ici. L'automate en question est entièrement mécanique. Or, puisque nous sommes plus qu'une simple mécanique, son IA de canard ne pouvait prétendre qu'au statut d'IA faible auprès des autres canards qui, le voyant agir comme eux, auraient certes pu avoir des difficultés à trancher la question. L'organe qui fait la différence, c'est le cerveau. Il est peut-être possible de reproduire la mécanique d'un corps, comme le prônait Descartes, mais l'activité cérébrale n'a rien de mécanique ; elle est électrique. Les neurones sont eux-mêmes des "tissus" à part, car ils permettent à l'être vivant à qui ils appartiennent, en plus de leur fonctionnement, une certaine forme de conscience, y compris d'eux-mêmes. Le cerveau est l'organe qui, en plus de ses multiples fonctions vitales, permet de se prendre lui-même pour objet de conscience et de réflexion. Les neurones sont la clef de tout, le mystère de cet organe complexe fait que nous sommes, tous en tant qu'êtres vivants, des formes de vie et non des objets. Sans cerveau, nous serions tous apparentés à cet automate de Vaucanson, à la seule différence que nous serions

constitués de cellules plutôt que de simples atomes inertes. Sans cerveau, il n'y aurait que la forme de vie la plus basique, à savoir végétale. Il n'y aurait ni conscience ni mouvement volontaire. Nous ne subirions que les changements programmés par avance par la nature de notre ADN, rien d'autre. Lorsqu'un patient entre en état de mort cérébrale, nous pouvons maintenir artificiellement son corps en vie, ce dernier poursuivant ainsi son vieillissement naturel. Pourtant, cela ne présente aucun intérêt et pour ce qui est du sujet, il est mort en même temps que son cerveau. Nous ne sommes que notre capacité à prendre conscience, ce n'est rien d'autre que l'expression d'un organe en activité. Penser qu'il y a une réalité derrière telle que l'âme n'est qu'illusion. Si tôt que cette capacité cesse, cela revient à mourir. C'est pourquoi lorsque le cerveau s'éteint, comme dans un coma, le sujet qui se réveille ne se souvient de rien et n'a aucune notion du temps passé. Sa temporalité est définie par sa conscience. Si sa conscience s'arrête, son temps s'arrête également. Alors que pour nous, toujours conscients, le temps continuera de s'écouler, pour lui, son vécu entre l'entrée et la sortie du coma sera instantané. Ainsi, nous savons peut-être reproduire artificiellement la vie, par le clonage, mais la reproduction de l'intelligence demeure un problème. Une IA faible est certes utile, mais son être moral ne vaut guère plus que celui d'un légume. Sauf qu'il s'agit là de la seule forme de vie qui ne dispose pas d'intelligence. Il faut toutefois faire ici la différence entre IA et clone. Le clonage est la reproduction artificielle d'une forme de vie à partir de la manipulation biologique. Autrement dit, l'accent est d'abord mis sur le corps, sur la cellule. Tandis que la composition physique, ou corporelle, d'une IA est tout à fait secondaire et ne doit d'ailleurs être pensée que par rapport à ce qu'elle peut apporter à l'intelligence dont il est avant tout question. Ainsi, la cellule ou la composition en base carbone n'est pas un prérequis essentiel. La question qui en ressort alors est : vie et intelligence sont-elles nécessairement indissociables ? Qu'une IA soit considérée comme une forme de vie est une autre question. Mais s'il est nécessaire d'être vivant, c'est-à-dire disposer de cellule et d'un ADN en base carbone, pour jouir d'une quelconque forme d'intelligence, alors est-ce que la possibilité d'une IA forte, en dehors d'une certaine forme de clonage, n'est pas qu'une illusion ? Vie et intelligence sont sans doute liés, mais ce dont il est moins certain, c'est que vie et ADN en base carbone soit nécessairement liés. Le fait que nous ne connaissions pas de forme de vie – et d'intelligence – différente du modèle de la nôtre ne constitue pas pour autant un argument invalidant la possibilité d'en trouver une différente, et par conséquent, d'en créer une différente. Si le clonage est quant à lui une création d'une certaine forme de vie, sans aucun doute artificielle, alors il est en théorie tout à fait possible de créer une IA forte basé sur un autre genre "d'ADN", pour peu que les lois de l'univers le permettent. Or cela, c'est à la science de le découvrir.

Nous sommes d'accord pour dire qu'il n'y a aucune loi, aucun interdit, aucune thèse vérifiée en vertu de quoi la création d'une IA forte serait impossible. Rien, en théorie, ne permet d'affirmer qu'il s'agit d'une quête vaine. Nous allons alors nous en remettre à l'évolution de la science et tenter de donner des pistes de recherche, en posant des hypothèses philosophiques favorisant l'avènement d'une IA forte.

Imaginons qu'une IA maîtrise si bien notre langage qu'elle parvienne à franchir le test de Turing avec facilité. Est-ce suffisant pour être qualifiée d'IA forte ? Après analyse du test de Turing, nous avons relevé, nous semble-t-il, tous ses défauts. Au point que nous avons donné quelques pistes pour établir un nouveau test de Turing, dans le but de faire de sa réussite une réponse positive infaillible. Nous avons notamment envisagé l'idée qu'une IA doive résoudre un problème en communiquant avec une autre IA, les forçant ainsi à s'entraider. Le but est de prouver qu'elles peuvent véritablement expérimenter, sans l'aide de l'homme, et s'en sortir grâce à une capacité d'apprentissage à la fois de leur propre expérience, mais aussi de l'une par rapport à l'autre.

Nous avons vu avec *Mitsuku* qu'une IA peut faire mine de comprendre un raisonnement abstrait et complexe mais est incapable de l'apprendre. Une réflexion telle que l'ego cogito lui est inaccessible, même guidé par l'homme. Pour nous, il est clair que *Mitsuku* ne pense pas.

Après réflexion, la nature inhérente à l'IA ne nous semble pas pouvoir générer des états mentaux de type « émotions » de façon forte. On pourrait, tout au plus, faire simuler une émotion à l'IA, mais elle ne la ressentirait jamais comme une émotion, cela ne présente donc aucun intérêt. En d'autres termes, il ne nous semble pas opportun d'inclure un test qui aurait pour critère l'émotion pour valider l'IA forte. Ainsi, une IA forte serait un peu l'illustration parfaite de l'ataraxie décrite par les stoïciens. Si une école de pensée des Anciens a pu penser qu'il était possible d'abandonner les passions, par un exercice de sagesse très long et exigeant, alors il n'est pas absurde d'imaginer une forme d'intelligence dépourvue d'émotion. Il est intéressant de rapprocher l'ataraxie des stoïciens au stade ultime du nirvana dans la pratique de la méditation du bouddhisme. Sorte d'idéal pas tout à fait atteignable pour les mortels, nirvana est un mot qui désigne aussi la mort, tandis que l'ataraxie est l'absence de trouble dans l'âme. L'IA serait, par nature, plus proche de ces stades que nous, et d'ailleurs, en tant qu'elle n'est pas un être né naturellement, peut-on vraiment considérer qu'elle soit vivante ?

Si l'IA n'expérimente pas, pensée et langage sont caducs. Ils ne peuvent découler que de l'expérience. Nos corps de nature animale nécessitant la pensée pour survivre – contrairement à la vie végétale – et l'évolution du langage pour assurer la survie de l'espèce, dont l'apparition peut être attribué au problème de mystère de l'apparition de la vie, nous font comprendre que pensée et

langage apparaissent quand ils sont nécessaires. A partir de là, deux possibilités. Soit nous arrivons à créer artificiellement cette nécessité chez l'IA, soit nous arrivons à comprendre le processus qui les a fait naître en nous pour le reproduire, de la même façon que la nature elle-même, chez l'IA.

Ce n'est pas que l'IA forte soit définitivement impossible, mais nous sommes encore loin de faire passer à nos IA un test de Turing amélioré avec succès.

En l'état actuel des choses, il ne fait nul doute que toute IA qui franchirait le test de Turing ne sera pas pour autant une forme de vie, car comme son nom l'indique, tout sera encore artificiel en elle. Même si elle sera parfaitement capable de donner l'illusion de réfléchir et d'interagir avec nous comme le ferait un humain, elle demeurera quand même une IA faible, même la plus puissante d'entre elles. Le paradoxe avec la conception que l'on a d'une IA forte, c'est que tout ce qui demeure artificiel en elle sera sa création, son enveloppe extérieure. Mais ce qu'elle sera, en tant qu'être « pour soi », ainsi que ses états mentaux, seront tout aussi réels que les nôtres. Donc le sens « d'artificiel » n'est plus tout à fait le même dans une IA forte. Sinon il faudrait concevoir que tout n'est que simulation, et en ce sens l'IA demeurera un outil ; peut-être bien capable de simuler la pensée, mais une pensée qui ne se rapporte à aucune conscience. Une telle IA ne saurait souffrir. Elle subirait simplement la différence d'un état de chose, tel un objet, sans états d'âme. Or tout être « pour soi » est apte à ressentir aussi bien le plaisir que la souffrance. C'est finalement à cela que nous savons faire la distinction entre être « en soi » et être « pour soi ». Il y a une dimension éthique dans cette distinction. Créer une IA qui soit capable de simuler la pensée et de passer le test de Turing ne nous garantit pas qu'elle sera un être « pour soi », c'est-à-dire avec une volonté propre, un libre arbitre et une conscience d'elle-même, tant qu'elle ne sera pas apte à souffrir. Elle pourrait très bien simuler tout cela sans que jamais, dans son cerveau, il n'y ait un esprit pour qui tout ceci fait sens. Il faut donc que l'IA forte fasse preuve d'états mentaux. Mais encore une fois, tout dépend de ce que l'on veut. Est-ce qu'on veut une IA qui simule les états mentaux, ou est-ce qu'on veut créer un être à part entière qui ressent les états mentaux comme les ressentons, avec tout ce que cela implique ? Si c'est le premier cas qui nous intéresse, alors nous pourrions dire que l'IA artificiellement forte n'a rien d'impossible, c'est juste une question de temps et de développement des sciences. Mais si c'est le second cas que nous souhaitons atteindre, alors nous pouvons considérer la simulation par la programmation comme une étape importante dans la compréhension que nous acquérons de nous-mêmes. Ce n'est qu'une fois que nous saurons correctement simuler ce qui nous constitue, ayant ainsi compris la complexité qui fait l'être humain, que nous pourrions envisager de créer une IA consciente d'elle-même. On aura besoin de combiner plusieurs sciences, cela ne fait aucun doute, mais si on peut programmer un réseau

neuronal artificiel comme un brin d'ADN va coder ce que nous sommes, alors une autre question fait surface : est-ce que les caractères de chacun sont contenus dans notre séquence ADN ? Autrement dit, sommes-nous entièrement déterminés par notre ADN ? Si c'est le cas, alors il faut que la programmation informatique soit suffisamment développée pour rivaliser avec la programmation naturelle que nous pouvons trouver dans l'ADN. Mais quelles conséquences pour la liberté humaine ? Nous verrons alors, grâce à David Hume, dans le chapitre qui va suivre, que déterminisme et liberté ne sont pas nécessairement incompatibles, et nous verrons que contrairement à Descartes, qui nous a tout de même donné des pistes de réflexion, sa philosophie peut plaider en faveur d'une IA forte, tout comme Copeland le fera face au camp adverse, bien plus nombreux, dont l'un des plus grand opposant est sans doute John Searle, avec son puissant argument de la chambre chinoise.

On retiendra également le fait qu'un corps doué de sens est un préalable nécessaire à l'apparition de la pensée, et que la capacité à l'apprentissage constitue l'un des critères qu'il nous faudra valider au cours d'un nouveau test de Turing. Cependant, nous verrons par la suite que pensée, conscience de soi et libre arbitre impliquent la vérification d'autres critères qui viendront compléter celui-ci. En outre, la capacité au langage complexe est une conséquence inévitable à toute pensée complexe. Cependant, il faut bien comprendre que ce langage n'est pas le nôtre, mais un langage propre à la nature de l'IA, dont le langage humain n'est que la traduction. Et si une simulation de pensée implique un vrai langage, une simulation de langage n'implique pas de vraie pensée. Il nous faut donc dépasser la barrière du langage et valider des critères plus importants indépendamment de la forme qu'il revêt, car il ne s'agit là, au fond, que d'un souci de commodité : d'utilisation entre un homme et une IA faible, de communication entre un homme et une IA forte.

2

La liberté humaine Hume

-

Un argument en faveur de l'IA forte ?

4 – David Hume et *l'Enquête sur l'entendement humain*

David Hume est reconnu comme le plus éminent représentant de l'empirisme, tout en étant également, ce qui peut sembler paradoxal, le plus sceptique. Nous verrons qu'il n'y a là rien de paradoxal, et que la vocation de Hume en philosophie est justement de faire tomber les paradoxes qui n'en ont que l'apparence et les divergences d'opinion qui n'en ont que la forme ; du moins en ce qui concerne les grandes questions que philosophie et science ont suscitées à l'humanité depuis plus de deux millénaires. C'est donc dans un souci de clarification, avec une grande exigence de précision et une volonté d'exactitude, que Hume a rédigé son *Enquête sur l'entendement humain*. A la fois refondation, réécriture mais également nouveauté, Hume rédige là son œuvre principale, la plus représentative de sa pensée. Ouvrage de maturité donc, mais sans oublier qu'il a été sculpté à partir du bloc d'origine que formait le *Traité de la nature humaine*, dont Hume, dans un ultime aveu de 1775, désavoua cette œuvre de jeunesse qui, selon lui, possédait un principal souci dans la méthode, et ne devait ainsi pas être retenue dans ses œuvres complètes.

D'abord *Philosophical essays concerning human understanding*, l'ouvrage est publié pour la première fois en 1748. Réédités à Londres en 1750 et 1751, c'est en 1758, lors d'une nouvelle édition, que le titre devient *An Enquiry concerning human understanding*. Ce changement d'« essais » pour « enquête » traduit la volonté d'une exigence de précision quant à la philosophie dispensée dans cet

ouvrage. Cet ouvrage principal, qui reprend certains passages du Traité en les améliorant – i.e. en les synthétisant tout en les clarifiant – occupera Hume jusqu'à la fin de sa vie, puisqu'il en proposera une dernière édition en 1776, peu avant sa mort.

Le passage de l'Enquête qui fera l'objet de notre étude ici est la section VIII – *Liberté et nécessité*. Précédée de *L'idée de connexion nécessaire*, la section VII, dans laquelle Hume revient sur le paradoxe initial du dépassement du donné à partir du donné, tire profit des compléments de l'Appendice au Traité, l'amenant dans l'Enquête à insister davantage sur le fait de la croyance. La section VIII, quant à elle, vient compléter dans l'Enquête l'analyse de la causalité et de la nécessité, tout en représentant une nouveauté par rapport au Traité, puisqu'il va donner l'élucidation de la nécessité.

On peut concéder à Hume que l'effort de clarification a porté ces fruits, car cette section fait montre de nombreux exemples pratiques pour appuyer son raisonnement, qui nous paraît alors, tel qu'il le souhaitait, indiscutable. C'est en tout cas ce dont nous essayerons de rendre compte.

5 - Résumé critique de la première partie de la section VIII

5.1 - Hume et l'intelligence artificielle : un pont possible ?

Hume commence par souligner que si les opinions divergent encore sur des grandes questions philosophiques, ce n'est que parce que l'on ne s'est pas encore mis d'accord sur la définition exacte des termes employés. Il fait le pari que tous les philosophes ont en fait toujours été d'accord sur le problème de la nécessité, et pour le prouver, il prétend que son explication de la nécessité fera l'unanimité de tout esprit doué de raison, car il va commencer par bien poser la définition des termes.

Pour introduire son élucidation de la nécessité, il part d'un raisonnement qui fait appel au bon sens. Non sans référence aux acquis du cartésianisme, on peut penser que chaque homme naît avec un entendement qui a un potentiel standard. De sorte que, ce qui paraît raisonnable pour l'un doit le paraître également pour l'autre, car si l'usage de la raison est subjectif, ses propriétés n'en sont pas moins objectives. C'est d'ailleurs parce que la raison est également partagée qu'il est possible de faire

de la philosophie ; sans quoi chacun aurait sa propre vérité. En partant de ce principe, supposons que la question qui nous affaie n'est pas hors de portée de l'expérience humaine – contrairement à l'origine du monde, l'existence d'un au-delà, etc... Dans ce cas, il n'y a pas de raison, au fil du temps et des expériences, que les hommes demeurent divisés sur cette question ; il suffit de vérifier ce qu'il en est. Or si la discorde subsiste, telle que pour le problème de la nécessité, n'est-ce pas justement parce que ces hommes n'accordent pas exactement la même définition aux mots qu'ils emploient ? Car si les deux principes cités plus haut sont vérifiés – la raison est universelle et le problème de la nécessité n'est pas hors de portée de l'expérience humaine – alors nous dit Hume, c'est bien la seule explication qui rende compte de la persistance de ces désaccords. Ce problème de langage est primordial, car il empêche la philosophie d'avancer sur certains points et la freine sur d'autres. Il faut régler le problème. Car la philosophie n'est pas affaire de conviction. Soit le sujet est susceptible de philosophie, et alors on se donne les moyens d'y répondre, ce qui mettra tout le monde d'accord ; soit le sujet ne l'est pas, et dans ce cas, c'est un tort de vouloir l'inclure en philosophie. Il faut, au mieux, l'inclure dans les croyances, mais alors, il n'est même pas utile que cela fasse débat. Chacun se fera sa croyance, puisque de toute façon, c'est invérifiable. Or en ce qui concerne la nécessité, Hume pense que c'est un fait tout à fait observable et donc nous sommes à même de le comprendre. D'ailleurs, il a pressenti que chaque philosophe, qui prétend avoir un discours différent, dans le fond a la même compréhension que les autres de ce problème. C'est pourquoi Hume ne va pas inventer une solution, puisqu'elle est déjà là. Il va simplement l'explicitée, pour qu'elle nous apparaisse clairement, et non plus comme avant, une solution obscure présente dans un inconscient collectif. Il s'attache à rendre le problème et son élucidation universelles, c'est-à-dire proposer la synthèse qui mettra tout le monde d'accord.

La question de la nécessité est primordiale pour une autre branche de la philosophie, plus récente, celle de l'intelligence artificielle. Initiée par Alan Turing, l'une des grandes interrogations de cette philosophie, si ce n'est la plus importante, est celle de savoir si l'intelligence artificielle forte est possible. Autrement dit, est-ce que l'on pourrait créer une machine, qui ne serait pas humaine car non constituée en base carbone, mais qui posséderait la même raison universelle que tout être humain ? Dans ce cas, qu'est-ce qui nous empêcherait de la considérer comme un nouvel être vivant ? Car une même raison voudrait dire une même capacité aux états mentaux et une même possibilité de liberté. Bien sûr, nous en sommes encore loin, et nous ne savons même pas si les recherches actuelles en terme de développement d'intelligence artificielle vont dans le bon sens. Mais s'il est manifeste que l'intelligence artificielle faible est une réussite en progrès, à l'heure où la machine a vaincu les champions du monde humains au jeu d'échec et au jeu de go, la possibilité théorique d'une intelligence artificielle forte reste une grande interrogation. L'I.A. faible peut ressembler de plus en plus à l'être humain, ce n'est pas pour autant qu'elle disposera d'une conscience réflexive – i.e. savoir ce qu'elle

est et comprendre pourquoi elle fait ce qu'elle fait – ni fera preuve de véritables émotions. L'argument de la chambre chinoise développé par Searle a prouvé que même si une machine est douée pour résoudre un problème, ce n'est pas pour autant qu'elle comprend le problème. Or l'IA faible ne vaut rien en elle-même, elle n'a d'intérêt qu'en tant qu'outil pour les scientifiques, de véritables esprits humains ; sans quoi elle ne serait pas à même de faire des découvertes par elle-même. Mais combiner les capacités de l'IA faible avec les propriétés d'un esprit humain, dans ce qui deviendrait une IA forte, reviendrait alors à créer une personne à part entière, qui aurait son propre caractère et ses propres raisonnements. Mais c'est aussi la perspective d'une évolution encore plus rapide, avec des scientifiques potentiellement plus performants que les humains. Mais encore une fois, nous tâtonnons dans les recherches sans même savoir si cela est possible, s'il n'y a aucun argument qui interdise l'émergence d'une IA forte. C'est là qu'intervient l'explication de l'élucidation de la nécessité de Hume.

Le problème de la nécessité est celui de savoir si, comme le pensait Aristote, il faut séparer l'étude de l'homme et l'étude de la nature. Car s'il nous apparaît évident que les événements naturels ne sont pas liés au hasard – concept d'ailleurs vide de sens pour Hume – mais répondent effectivement à des lois, il nous apparaît beaucoup moins évident que les comportements humains, les choix que nous faisons, répondent eux-mêmes à des lois. Ce n'est pas pour autant que cela nie toute possibilité de liberté, mais ce n'est en aucun cas une liberté absolue ou encore une liberté d'indifférence, comme nous en avons faussement le sentiment. Et c'est par la raison seule que Hume va nous démontrer cela. Ainsi, s'il n'y a pas de liberté absolue qui, par définition, n'aurait pas d'origine explicable, à l'instar de l'existence métaphysique supposée d'un moi, dont Hume renie d'ailleurs l'existence – il s'agirait plutôt d'un postulat pratique de la raison mais sans réalité effective –, alors on ne voit pas ce qui serait mystérieux chez l'homme, et par là, irréproductible. En effet, puisque nous participons de cette même nature, et que notre conscience constitue simplement une différence de degré et non de nature par rapports aux autres animaux, et que, si nécessité il y a dans les comportements humains, alors c'est observable et compréhensible, et donc, reproductible. En d'autres termes, nous faisons ici le pari que la thèse de Hume qui défend une liberté compatible avec la nécessité qui nous gouverne, qu'il appelle liberté de spontanéité, constitue un argument en faveur de la possibilité de créer une IA forte.

5.2 - Nécessité et causalité : des faits universels ?

Revenons à son propos. Tout d'abord, Hume commence par ce qui fait le moins débat avant d'en arriver à la nature humaine, à savoir la nature. C'est en effet dans les phénomènes naturels que la nécessité se fait le plus évidemment ressentir. Cette nécessité est si déterminante qu'un effet naturel, à cause et paramètres identiques, n'aurait pu être autre. Cela, Hume le souligne, personne ne le remettra en cause. Prenons un exemple. Chaque fois que la condensation atteint son seuil maximum suite à l'évaporation des océans, principalement, les nuages se mettent à déverser la pluie. Puisque la même cause provoque toujours le même effet, on trouvera normal qu'il pleuve parce que les nuages sont trop denses pour maintenir l'humidité à l'état "semi-liquide". Bien entendu, dans le phénomène "pluie" comme dans tous les phénomènes, il y a toujours un enchaînement de cause qui mène à l'effet observé. L'évaporation est elle-même causée par l'intensité des rayons du soleil. Mais puisque cette chaîne de cause à effet est dans la réalité circulaire et non linéaire – comme dans notre entendement –, chercher à remonter à une cause première serait sans fin ; et cela n'a pas de sens. C'est pourquoi nous retenons la dernière cause, celle qui précède et provoque directement l'effet. En fait, la quantité d'eau n'est pas le seul facteur nécessaire qui va faire pleuvoir, car il faut d'abord que l'eau soit redevenue liquide grâce à une température froide, que l'on trouve justement aux altitudes où se forment les nuages. Mais ce premier facteur n'est pas suffisant, car si tous les nuages contiennent de l'eau liquide sous forme de gouttelettes microscopiques, elles demeurent plus légères que l'air et ne tombent pas. Ainsi lorsque le nuage est suffisamment chargé, sous l'effet de la quantité d'évaporation et de la pression atmosphérique, il se met à pleuvoir ; les gouttelettes microscopiques sont suffisamment nombreuses pour former une goutte d'eau, et poussées à se rassembler par la pression atmosphérique, elle tombe nécessairement, sous l'effet de la gravité terrestre. La nécessité est indéniable, car il n'en a jamais été autrement ; cela ne peut signifier qu'une chose : il ne peut en être autrement ; à conditions identiques s'entend. Donc il existe bien une nécessité dans la nature. On ne verra jamais un nuage ne pas faire pleuvoir alors qu'il le devrait, sauf si une cause contraire à la première vient la contrecarrer ; i.e. que les paramètres ont changé. Il ne s'agit pas d'entrer dans des élucubrations métaphysiques et de statuer sur la nécessité de ces lois de la nature. Il semble nécessaire qu'il y ait une nécessité, puisque de toute façon, il y en a une. Mais que cette nécessité-là soit contingente parce que les lois auraient pu être autres, ou nécessaire parce qu'un dieu hypothétique, créateur du monde et de tout ce qui existe, les aurait écrites ainsi ; une telle question n'est pas du ressort de la philosophie, et ce ne serait qu'une perte de temps que de réfléchir là-dessus. En soi, on

n'a pas de raison de penser que les lois de la physique sont telles qu'elles sont parce qu'elles n'auraient pu être autrement. Nous pourrions imaginer un autre monde, comme dans une dimension parallèle, où l'eau brûle et le feu mouille. Cela nous paraîtrait étrange, mais ce ne serait pas plus absurde que les lois de notre monde : il demeure une nécessité, même si elle n'est pas la même que la nôtre. Et si justement il s'agit d'un autre monde, c'est bien parce que par définition, la nécessité demeure quoiqu'il arrive. Elle ne peut s'inverser, sans quoi elle ne serait plus la nécessité. C'est pourquoi dans notre monde, l'eau mouillera toujours. Mais il est vrai que dans l'hypothèse de l'infini de l'univers, il n'y a pas de raison de penser que les lois de la nature soient infiniment identiques ; en fait, il se pourrait au contraire, puisqu'on parle d'infini, que tout soit possible. Toujours est-il que dans tous les mondes, il demeurera une nécessité inviolable. Nous pouvons donc penser que seule une nécessité est nécessaire. En effet, il n'y aurait rien de plus absurde que d'imaginer un monde où le hasard régnerait, où de la même cause ne suivrait jamais le même effet. Un tel monde instable, en plus d'être inconcevable, serait invivable et non souhaitable.

Après cette amorce, Hume entreprend donc la recherche de définition la plus rationnelle qui soit du phénomène de nécessité, en commençant par l'origine de cette pensée en nous. Une pensée qui, comme nous venons de le voir, nous est commune à tous par le simple fait d'être au monde ; philosophe comme ignorant. La thèse de Hume, pour répondre au problème de l'origine de l'idée de nécessité en notre esprit, est l'accoutumance. Expliquons cette thèse. C'est-à-dire que si de la même cause ne suivait jamais le même effet, nous n'aurions jamais eu en nous l'idée de nécessité. Par seule expérience d'être au monde, c'est-à-dire de vivre, et non par expérience scientifique, nous observons que les phénomènes se répètent de façon similaire. C'est pourquoi, à travers l'exemple qu'il prend, un vieux paysan sera toujours meilleur qu'un jeune, car ayant vécu plus longtemps, il aura observé plus longtemps. L'expérience qu'il a accumulée au fil des années lui sera favorable pour s'occuper de ses récoltes. Cependant il faut distinguer l'expérience, au sens usuel d'habitude, de l'expérience scientifique. Car ce n'est pas parce qu'un homme sait qu'il va pleuvoir qu'il comprend pourquoi il va pleuvoir. Avoir observé que d'une cause A suit l'effet B et que l'on sait reconnaître A ne signifie pas que l'on sait pourquoi A implique B ; d'où l'expérience scientifique. D'ailleurs cette expérience première n'est pas infaillible, puisque si les conditions changent, à savoir qu'une cause ignorée X vient s'interposer, elle aura des incidences sur la cause initiale A et l'effet B en sera modifié. Or tant que la cause X est ignorée parce que non connue, on aura raison de croire que A implique B et pourtant on sera surpris que ce ne soit pas le cas cette fois-là. Mais c'est quelque chose que la science peut justement mettre au jour. Ce qu'il est important de comprendre, c'est que le hasard n'a rien à faire là-dedans. L'observation de la nécessité rend compte d'une telle régularité qu'il faut en venir à la

conclusion que la nécessité est en fait infaillible, et que lorsque A n'implique pas B, c'est seulement parce qu'une donnée nous échappe et que A n'était pas ce que l'on croyait ; il était en fait A'.

L'idée de connexion entre les objets de la nature vient donc d'une observation primitive et passive. C'est par accoutumance, à force de voir quelles sont les conditions atmosphériques quand il pleut, que l'on finit par savoir quand il s'apprête à pleuvoir. Pour autant, ce n'est que grâce à une observation secondaire et active, par la réflexion et la science, que l'on a pu être à même, bien plus tard, d'expliquer le phénomène de la pluie. Cette accoutumance se fait sur de si nombreux objets que la croyance et la prédiction à court terme, ou autrement dit les attentes, les inférences, deviennent une attitude naturelle, qui nous permet d'évoluer plus aisément dans notre monde. Ainsi, on en vient naturellement à penser qu'il y a une nécessité à ce que les choses se reproduisent à conditions égales. De là, on peut se demander si la nécessité est à l'œuvre dans toutes les choses du monde ; c'est-à-dire pas seulement dans les objets, mais aussi dans les êtres vivants, et, ce qui paraît le moins évident, dans les actes qui font l'objet d'une volition ; à savoir les comportements humains.

L'idée de nécessité est intrinsèquement liée à l'idée de causalité. Voyons voir maintenant ce qu'il en est de la nature humaine, qui selon Hume, doit être étudiée comme on étudie la nature des choses. Lorsqu'il y a nécessité qu'un phénomène se produise dans la nature, on suppose qu'il y a un lien de causalité. C'est en effet le cas. En revanche la réciproque est-elle vraie ? Si un lien de causalité est établi, faut-il admettre qu'il y a alors nécessité ? Pour ce qui est de la matière, comme dit Hume, la réciprocity ne fait aucun doute. Il n'y a pas de volonté à l'œuvre dans la nature, et donc pas de liberté possible. Ce qu'il s'est passé n'aurait pu se passer autrement. Mais en ce qui concerne la nature humaine, les choses se compliquent. Serions-nous comme des marionnettes, ignorants des ficelles qui nous tirent ? Est-ce qu'à chacune de nos pensées, de nos actions, de nos paroles, il n'aurait pu en être autrement ? Serions-nous entièrement déterminés par des forces en nous que nous ignorons ? Il s'agit là d'une hypothèse extrême et pessimiste, à laquelle les philosophes ont habituellement à cœur de s'opposer. L'homme neuronal de Jean-Pierre Changeux tendrait à aller dans ce sens, mais le scepticisme de Hume propose pourtant une autre vision, moins pessimiste, bien que tout à fait rationnelle.

Le fait est que cette inférence que nous faisons avec les choses, nous la faisons également avec autrui. Dans nos rapports humains, nous avons naturellement des attentes. C'est-à-dire que l'accoutumance aux phénomènes de la nature s'applique tout aussi bien aux comportements humains, et cela, personne ne peut le nier non plus. Dans telle situation similaire, nous attendons plutôt, selon tel ou tel caractère, à ce que l'individu agisse comme ceci ou comme cela. Non seulement nous pouvons faire des probabilités sur les comportements d'autrui, mais nous le faisons malgré nous,

presque instinctivement. Imaginerions-nous vivre au sein d'une société où personne ne suppose savoir comment l'autre va réagir à tout instant ? Tout être doué de raison répondra non. En effet cette société serait chaotique, il serait impossible d'y vivre. Il y a une harmonie justement parce qu'il y a des répétitions dans les comportements, et en raison de cette même accoutumance, nous faisons naturellement des inférences, c'est-à-dire que nous avons toujours des attentes dans nos rapports avec les autres. De même, on peut aussi parler ici d'expérience. Pourquoi l'adolescent vient plutôt demander conseil au grand-parent en ce qui concerne les tourments des premières relations amoureuses ? L'aïeul ayant davantage vécu, on le considère naturellement plus expert en relation humaine ; l'expérience de la vie s'applique aussi en la matière, même si effectivement, on conviendra qu'il s'agit d'une science humaine et non d'une science de la nature. Ainsi, le fait d'émettre des inférences qui s'avèrent correctes en matière de nature humaine constitue tout un pan de la maturité. Le parallèle est donc tout à fait possible, même si Hume ne réfute pas la différence entre les deux types de science. Il ne s'agit de dire qu'il faut appliquer la même méthode, car il y a une différence – nous le verrons plus tard, c'est celle de la possibilité d'une certaine liberté – mais il ne faut pas séparer l'homme de la nature, car il en est issu et y participe tout autant que chaque chose. En tant que déclinaison de cette nature, tout ce qui s'applique à la nature s'applique nécessairement à l'homme. Tandis que la réciproque n'est pas nécessairement vraie.

Il y a donc également un phénomène d'accoutumance dans nos relations avec autrui. Lorsque je prête de l'argent à un ami pour qu'il puisse réaliser son projet, je m'attends à ce qu'il me rembourse dès qu'il le pourra. C'est en vertu de l'amitié qu'il y a de fortes probabilités que cet ami finisse effectivement par me rembourser ; s'il est vraiment mon ami. Sans cette croyance, je n'aurais pas accepté le prêt. Si l'acte de remboursement était tout à fait aléatoire, on ne ferait que des dons sinon rien. Ainsi, s'il s'avère que cet ami ne me remboursa pas dès qu'il le pût, je vais alors inférer qu'il n'y a pas de raison qu'il me rembourse une fois prochaine, et donc je vais décliner toute demande de prêt à venir. En d'autres termes, j'agis rationnellement, en fonction des inférences que je suis amené à faire. Et cela vaut pour toutes nos actions. Aucune action, lorsqu'un homme vit au sein de la société – ce qui est le cas de tous sinon de quelques ermites marginaux – ne concerne que lui-même. Nous sommes toujours liés à autrui, d'une manière ou d'une autre, et donc chaque action a des conséquences, même minimes, sur les autres. C'est pourquoi, naturellement, par accoutumance, nous éclairons toutes nos actions, même inconsciemment, par l'inférence que nous pouvons faire vis-à-vis de la réaction des autres. Cela, personne ne le conteste. Car si l'inférence n'avait pas un rôle à jouer lors du passage à l'acte de la volition, cela signifierait qu'il n'y aurait alors pas d'accoutumance, et donc pas de répétition dans les rapports humains. Mais si c'était le cas, nous serions constamment dans la surprise de la réaction d'autrui, étant dans l'incapacité de la prévoir. Il n'y aurait pas d'appréhension

possible des rapports humains et la vie en société serait anarchique. Heureusement, ce n'est pas le cas. Et cela n'est pas tellement étonnant selon Hume, puisque l'homme en tant qu'être vivant est lui aussi un produit de la nature.

Or, partant du même raisonnement que pour la matière, si accoutumance et inférence il y a, alors il y a une certaine causalité qui est à l'œuvre dans l'esprit humain. Si certains caractères réagissent toujours de la même façon dans les mêmes situations, on ne peut que constater, à l'instar d'un phénomène naturel, que l'effet est le produit d'une cause. Mais comme nous l'avons vu plus haut, causalité et nécessité sont intrinsèquement liés. Est-ce que cette même observation de la similitude dans les rapports humains que de la similitude dans les phénomènes naturels implique que la causalité qui est à l'œuvre dans l'esprit humain, quel qu'il soit, est l'expression d'une nécessité ? En un certain sens, oui. Il est nécessaire que tout effet d'un esprit humain, comme dans la matière, réponde à une cause. L'autodétermination n'est qu'une illusion, un concept vide de sens, dont nous pourrions souligner l'absurdité, comme Nietzsche l'a fait pour critiquer l'argument spinoziste qui consistait à dire que Dieu est la cause première en étant cause de soi, afin de régler définitivement ce fameux problème métaphysique, en prenant l'exemple du baron de Münchhausen qui se serait tiré d'un sable mouvant en se tirant lui-même par les cheveux. Il n'y a rien qui soit dépourvu de cause dans la nature, et pourtant il n'y a jamais de cause première car la chaîne est circulaire. En fait il en va de même pour l'esprit humain. On ne peut concevoir un esprit produire un effet orphelin de toute cause. Cependant, si la causalité est nécessaire dans l'esprit humain, chaque esprit étant unique, la causalité à l'œuvre n'est jamais identique. C'est pourquoi nous ne faisons pas des inférences selon la connaissance que nous avons de l'espèce, d'après des dispositions biologiques, mais selon la connaissance que nous avons de la personne, d'après des dispositions caractérielles. Et lorsque nous ne connaissons pas la personne, les inférences que nous faisons dépendent cette fois-ci des normes et valeurs sociales qui ont cours dans la société dans laquelle nous vivons.

Quoiqu'il en soit, il y a effectivement une nécessaire causalité à l'œuvre dans les opérations de l'esprit. Pour autant, est-ce à dire que les choix, les actes que nous faisons chaque jour ne pourraient être autrement ? Le choix ne serait alors qu'une illusion. Pourquoi aurions-nous le sentiment d'avoir une liberté de choisir, si ce n'était pas le cas ? Est-ce que la liberté, au même titre que le hasard et l'autodétermination, sera rangée par Hume du côté des concepts vides de sens ?

5.3 - Une causalité nécessaire pour une liberté illusoire ?

Le fait est que les passions humaines sont universelles. Puisque Hume veut entamer une science de la nature humaine, il est intéressant de commencer par constater que les grands traits de l'esprit humain, à savoir les passions, i.e. les motifs de l'action humaine, se retrouvent dans toutes les nations et à toutes les époques. Il y a un rapport que nous pouvons dire de causalité entre certains motifs et certaines actions, car c'est un rapport qui s'observe de manière universelle. Le temps est un facteur qui n'a aucun impact sur cette réalité. Tant que l'homme sera homme, ce qui se disait de la nature humaine dans l'Antiquité, en terme de motifs et de passions, vaut toujours pour les hommes d'aujourd'hui. Tout comme la cause qui produit l'effet « pluie » aujourd'hui était la même il y a un milliard d'année sur Terre. Cela ne signifie pas que la causalité est immuable, seule la nécessité l'est. C'est-à-dire qu'à conditions égales, la causalité demeure intacte, indépendamment du facteur temps. Mais si la planète a eu un jour des paramètres différents, il se peut que la causalité qui produise l'effet « pluie » ne soit plus la même, tout comme l'homme n'avait pas les mêmes motifs lorsqu'il était Toumaï ou Homo Erectus, et n'aura plus les mêmes s'il devient un jour plus qu'homme sapiens sapiens. La causalité est donc éternelle ; elle vaudra toujours pour les paramètres qui la déterminent, mais elle n'est pas immuable, car ces paramètres ne se maintiennent pas nécessairement dans le temps. Faut-il pour autant voir en l'explication possible de nos actions la négation de la liberté ?

Maintenant que la nécessité a été ainsi clairement définie par Hume, nous comprenons qu'elle n'a pas à nier, de fait, la possibilité de la liberté. Cependant, il ne s'agit pas d'une liberté absolue, mais d'une liberté qui s'exécute au sein de la causalité. En effet, cette liberté, nous ne pouvons la nier. D'abord parce que nous avons l'intime conviction que nous aurions pu faire autrement que ce que nous avons fait, même si cela reste impossible à prouver. C'est le problème de la finalité, qui ne peut se prouver car l'invalidation de son contraire est une chose que les conditions de l'expérience humaine – qui sont l'espace et le temps – ne nous permettent pas. Même si je pouvais remonter le temps pour faire le choix inverse, je ne prouverais nullement ma liberté, car lorsque j'effectue cette autre action, je ne suis plus celui qui devait alors faire un choix, mais celui qui a déjà fait un choix et qui a remonté le temps pour en faire un autre ; donc j'aurais très bien pu être déterminé à agir dans cet ordre. Pourtant nous postulons à chacun de nos choix qu'il y a une instance qui gouverne et qui pourrait très bien en faire d'autres. On ne ressent pas de nécessité d'agir de telle ou telle manière, comme si nous subissions nos vies tels des spectateurs impuissants. On ressent seulement une nécessité d'agir. L'action quant à elle semble être du ressort de notre liberté et se rend effective lorsque la volition

passé de la puissance à l'acte. Ensuite, ce sentiment de liberté est renforcé par le fait que nous sommes parfois surpris de la réaction d'autrui. La réaction n'est pas toujours celle que nous attendions, même si c'est le cas la plupart du temps. Il ne faut pas s'y tromper. La nécessité ou la causalité n'entre pas en contradiction avec la possibilité d'une certaine liberté. Nous pouvons émettre l'hypothèse qu'il existe une liberté contenue dans la causalité. C'est-à-dire qu'à telle cause et telles déterminations de l'individu, tels effets en réponse sont possibles. La liste finie de ces effets constitue la nécessité de réaction de l'individu, tandis que le choix d'un effet plutôt qu'un autre va être l'expression de sa liberté propre. Si liberté absolue il y avait, alors il y aurait une infinité d'effet possible. Mais cela nierait le principe de causalité et pose un problème majeur : sur quelle base l'individu pourrait-il alors choisir ? Si une telle liberté existait, elle serait aveugle, hasardeuse. Or on ne peut imaginer qu'une réaction humaine soit dépourvue de cause. L'hypothèse proposée par Hume est donc la suivante : alors que dans la nature, A implique nécessairement B, la nature humaine se complexifie puisque W peut impliquer X, Y ou Z. Cependant, il ne peut ni impliquer autre chose, ni ne rien impliquer. Notre seule liberté s'exprime alors dans le choix de X, Y ou Z, mais même cette liste d'effets possibles n'est pas de notre ressort. A l'instar de la physique, il semblerait que la différence entre la matière et l'esprit revienne à la différence entre la physique classique et la physique quantique. Alors que la loi est une dans les phénomènes naturels, on s'aperçoit qu'elle peut être multiple dans la nature humaine, sans pour autant déroger au principe de nécessité et de causalité ; ce qui peut sembler paradoxal.

Au même titre que le moi, cette hypothèse de liberté semble être un postulat pratique pour l'exercice de la raison, mais Hume parvient-il à prouver son existence réelle ? Ou bien aboutit-elle à une solution sceptique, ainsi qu'il finira par nier l'existence d'un moi ?

5.4 - La *liberté de spontanéité* face à la *liberté d'indifférence*

S'il fallait nier la nécessité pour justifier la liberté innée de la volonté dans toute action humaine, comme on le prétend souvent, alors il s'en suit que les actions des hommes ne pourraient être moralement sanctionnées. Il ne serait ni louable, ni blâmable qu'un homme fasse telle action plutôt que telle autre, puisque la liberté de sa volonté n'est aucunement déterminée ; elle ne provient d'aucune cause. Autrement dit, cette liberté n'aurait aucun rapport, ni avec les forces extérieures, ni avec le caractère intérieur. Ainsi, on ne pourra reprocher à un individu d'avoir commis un acte qui n'est

pas de son fait. Or, s'il a agi avec cette soi-disant liberté innée, cela signifie que l'action ne se rapporte nullement à son être, à son caractère. On ne peut lui attribuer moralement l'acte. L'acte serait certes condamnable, mais pas la personne qui, une fois l'acte achevé, redeviendrait aussi pure qu'à sa naissance, puisqu'aucun état interne ne peut influencer sur la liberté de la volonté. Ainsi, cette liberté de la volonté, qui serait absout de toute cause, ne pourrait pas être attribuée à l'individu, car le moi ne peut en être à l'origine ; le moi se définissant par l'ensemble des expériences que nous avons vécu depuis notre naissance et qui, par inférence, se forge un caractère en adéquation avec l'environnement de vie et la personnalité d'un individu. Mais si cette liberté se veut vraiment libre, c'est-à-dire détachée de tout déterminisme, même le moi de l'individu, par définition, ne peut y présider. Cela revient à dire qu'il s'agit là d'une liberté anarchique, chaotique, qui va et qui vient, sans que l'on ne sache pourquoi ni comment, et dont, par conséquent, les prévisions sont tout à fait impossibles. Vivre dans un monde où une telle liberté humaine aurait cours dans chacun des choix d'action des individus serait unimaginable. Pourtant, si tel était le cas, nous ne pourrions effectivement plus condamner une personne mais seulement l'acte, qui s'est exprimé malgré nous, à travers une liberté qui manifestement nous transcende. Or, il s'avère dans les faits que non seulement nous jugeons les hommes pour leurs actes, nous les tenons pour moralement responsables, mais en plus, nous sommes tous d'accord pour dire qu'il y a, la plupart du temps, une raison qui explique l'action de l'un et le choix de l'autre. De même, on va aller jusqu'à dire que certaines réactions, chez des personnes que nous connaissons bien, sont en effet prévisibles. Pourquoi cette inférence est-elle possible, comme dans notre observation des phénomènes naturels, où les lois nécessaires de la nature sont à l'œuvre ? C'est évidemment parce qu'une nécessité est à l'œuvre dans l'exercice de la volonté humaine, et donc la liberté innée n'est qu'une fiction. Pourquoi attache-t-on alors tant d'importance à défendre l'idée que la liberté humaine est indéterminée ? Cela ne se peut. Mais ce qu'il faut comprendre, c'est que le déterminisme de notre volonté n'anéantit pas toute possibilité de liberté. Liberté n'est pas nécessairement liberté absolue ; c'est là un détournement et une réduction de sa définition. Si certains actes sont prévisibles, il est raisonnable de penser que tous le sont virtuellement ; tout acte est rationnellement explicable, or attribuer un acte au hasard est une erreur, car de même que la liberté innée, le hasard ne semble pas exister dans nos observations du monde. Il ne fait que désigner le fait que nous ignorons l'explication, la cause. Il ne faut pas en déduire pour autant qu'elle pourrait être inexistante.

Hume défend donc une forme de liberté compatible avec la nature déterminée des choses et des actions humaines ; nous l'appelons la liberté de spontanéité. Cette liberté, c'est le pouvoir que nous mettons en œuvre par la volonté de faire une chose ou de ne pas la faire, dans le cadre des possibilités que ses déterminations lui offrent. C'est-à-dire qu'il y a un certain nombre de choix, en

accord avec moi-même – qui ne sont pas dépourvus de sens – que je suis libre de faire. Mais je peux également décider de ne faire aucun de ces choix, la négation d'action étant une liberté d'action. Cependant, je ne pourrai pas faire quelque chose qui échappe aux déterminations de ma volonté. Prenons un exemple. Si je cherche du travail et que je trouve une offre d'emploi dans ma branche, mais à l'étranger, je suis libre de choisir d'y postuler, parce que ce qui compte le plus pour moi c'est de travailler et je peux aimer la destination. Mais je peux aussi décider de ne pas postuler et d'attendre de trouver une offre près de chez moi. Les déterminations de la volonté dans ce cas dépendent donc entièrement du caractère de l'individu : son désir plus ou moins fort de travailler, le fait d'être en activité mis en balance avec ses autres priorités, son goût pour le voyage et sa capacité d'adaptation, face à son attachement pour son lieu de naissance. En fait, il y aurait de nombreuses autres déterminations en réalité, comme c'est toujours le cas, et l'esprit fait l'opération en additionnant et en faisant pencher la balance, ce qui nous fait faire le choix, finalement, le plus en accord avec nous-mêmes. Mais ces déterminations sont justement si nombreuses, que dès lors qu'il s'agit de les sortir de l'esprit pour les rationaliser, on s'aperçoit tout de suite de la complexité de la tâche, presque impossible, pour savoir avec certitude ce que sera le choix de la personne avant qu'elle ne le fasse. C'est virtuellement possible, mais cela reste très difficile, car les paramètres à prendre en considération sont trop nombreux, tout comme nous n'arrivons pas à expliquer tous les phénomènes de la nature d'ailleurs. Cependant, si l'on s'arrête à ces seules déterminations énoncées, il est impossible que l'individu fasse le choix de se suicider. D'abord, il n'y penserait pas ; justement parce qu'il n'a aucune raison d'y penser. Les données qu'il a à disposition, l'état du monde dans lequel il évolue et ses dispositions internes n'amènent tous ensemble aucune corrélation avec le fait de se suicider. C'est-à-dire que dans notre exemple, on suppose donc que notre individu n'a aucun attrait pour la mort, aucun désir de mourir et aucun problème ou douleur si forts qu'ils pourraient le conduire, pour les stopper, à se donner la mort. Dans le choix qu'il a à faire de postuler ou d'attendre, l'acte de se suicider n'entre pas en compte parce que d'une part, ça n'a aucun rapport, et d'autre part, il n'y a aucune raison qui puisse l'y pousser. Naturellement, notre individu, au moment où il réfléchira pour prendre sa décision et agir, ne pensera pas au suicide. Mais imaginons que nous lui suggérerions l'idée. Naturellement, l'individu pensera qu'il est libre de se suicider. Pourtant, jamais il ne le fera. Ceci, en raison de ce que nous venons d'évoquer. Il ne le fera pas, parce que cet acte ne répond à aucune des déterminations de sa volonté. Le fait qu'il pense être libre de pouvoir se suicider ne veut pas dire qu'il le soit, et le fait qu'il n'y ait pas pensé sans notre aide devrait lui mettre la puce à l'oreille. En fait, s'il le faisait, ce serait uniquement pour nous prouver qu'il était libre de le faire. Mais si nous ne lui en avons pas parlé, il ne l'aurait pas fait. Et même pour cette raison absurde, il y a fort à parier que personne ne le fasse jamais, car chacun préférera préserver sa vie plutôt que de prouver une chose dont nous avons tous le même sentiment intime. Il pense être libre de se suicider, et pourtant, on sait

tous qu'il ne le fera pas. C'est parce qu'en fait, en raison des déterminations de sa volonté, il n'est pas libre de le faire. Comme nous le dit Hume dans la section précédente, toute idée dérive d'une impression. Voici déjà une première détermination de la volonté. Avant d'avoir l'idée du suicide, encore faut-il avoir eu une impression qui nous conduise à la pensée morbide. Mais si c'est là une condition nécessaire, ce n'en est pas pour autant une condition suffisante. Le suicide est un choix qui peut être présent à l'esprit, ce n'est pas pour autant qu'il entrera dans le champ des déterminations de la volonté, en tant qu'acte possible. En effet, il faudra encore que ce choix contrebalance la possibilité d'un autre choix qui lui serait inférieur dans l'échelle des priorités. Toutefois, toutes décisions ou acte n'est pas nécessairement rationnel. Les hommes agissent parfois de manière passionnelle. Ce n'est pas parce qu'un acte n'est pas rationnel qu'il n'est pas explicable. Les passions vont faire passer certaines priorités devant d'autres, sans que cela soit rationnel pour autant. Il faut donc en conclure que les déterminations de la volonté sont de deux sortes : soit elles viennent de la raison (le moi rationnel), soit elles viennent des passions (le moi passionnel). Quoiqu'il en soit, les agissements des uns et des autres sont toujours explicables, à défaut d'être toujours compréhensibles ; et la liberté de chacun ne s'exprime que dans le cadre des déterminations de la volonté. C'est la nature de la volonté qui veut cela. Partout il y a nécessité, mais là où il y a volonté, il y a liberté, car la nécessité se subdivise alors en un champ nécessaire de possibilités contingentes les unes par rapport aux autres. Ce qui n'a pas de sens dans la nature, là où ne se trouve aucune volonté particulière.

6 - Analyse de l'utilisation de Hume par Copeland en philosophie de l'IA

6.1 - Deux sortes de causalité : contingente et nécessaire²⁷

Quelle connaissance avons-nous réellement lorsque nous savons que ce qui a causé le changement de trajectoire de la boule de billard B est la collision avec la boule de billard A ? Que savons-nous lorsque nous savons qu'un événement X cause un autre événement Y ? Est-ce que cela signifie qu'il y a une connexion nécessaire entre X et Y ; que l'événement X contient quelque chose, de connu ou non, qui implique nécessairement Y ? Ou alors, est-ce que tous les événements suffisamment ressemblants à X sont invariablement suivis par un événement comme Y ?

C'est la question que pose Jack Copeland, dans son ouvrage *Artificial Intelligence, a philosophical introduction*. La première proposition, à savoir que de X précisément suit Y, est la plus forte. C'est celle que nous observons dans les phénomènes naturels. Il faut que toutes les conditions de X soient réunies pour qu'il se passe Y ; c'est-à-dire que tout ce qui ressemble à X, mais sans être X, produira autre chose que Y ou ne produira rien du tout. En revanche, la seconde proposition, plus faible, est celle qui nous permet de faire des inférences un peu plus certaines, et c'est celle qui s'applique dans les comportements humains ; donc celle qui s'applique à des volontés, à la liberté. L'observation nous fait remarquer que tout ce qui se rapproche de X va produire quelque chose qui se rapproche de Y. Mais il n'y a pas de lois sur le Y attendu, justement parce que, quel que soit la forme de X, il laissera toujours la possibilité à différentes formes de Y de suivre.

Pour comprendre cela, il faut revenir à la section VII de l'Enquête, *L'idée de connexion nécessaire*. On comprend que Hume distingue entre deux origines de la causalité. L'une est nécessaire et l'autre est contingente. La causalité nécessaire est celle qui régit les lois de la nature. Si les choses se passent ainsi dans la nature, c'est qu'il ne peut en aucune façon en être autrement. C'est une obligation dans la fond et dans la forme. Tandis qu'avec la causalité contingente, c'est une obligation

²⁷ Cf. Section VII de l'*Enquête sur l'entendement humain*, et le chap. 5.6 (*The cyc project*), p. 105, § 3, de l'ouvrage de Jack Copeland.

dans le fond mais pas dans la forme. C'est-à-dire qu'il est inévitable que l'événement X me fasse réagir en W, Y ou Z, mais il est tout à fait contingent que la résultante soit Y plutôt que Z ou W et inversement. Cette contingence est la preuve qu'une certaine liberté peut s'exprimer. Il ne faut cependant pas en conclure que cela relève du hasard. Mais parce que la résultante dépendra de chacun, on ne peut qu'inférer qu'un événement voisin de X provoque un événement voisin de Y.

6.2 – Deux types de liberté²⁸

« Par liberté, alors, nous pouvons seulement entendre *un pouvoir d'agir ou de ne pas agir selon les déterminations de la volonté* ; c'est-à-dire, si nous choisissons de rester en repos, nous le pouvons ; si nous choisissons de nous mouvoir, nous le pouvons aussi. Or, cette liberté conditionnelle appartient, de l'aveu universel, à tout homme qui n'est pas prisonnier dans les chaînes. Alors, il n'y a là aucun sujet de discussion. »²⁹

Avec Hobbes, Hume est l'un des premiers compatibilistes, nous dit Jack Copeland. Les compatibilistes sont ceux qui soutiennent l'idée que le déterminisme neurophysiologique n'est pas incompatible avec la possibilité d'une volonté libre ; contrairement aux incompatibilistes, tels que Kant. L'argument est le suivant. Dans la prise de décision d'un homme, il y a certes toute une réunion de causes physiques et naturelles qui peuvent, seules, expliquer le choix effectué. Pourtant ce ne sont pas les seules causes, car l'homme a toujours la possibilité de ne pas agir. Cela, on ne peut le prouver, car nous ne pouvons pas revenir en arrière. Mais c'est quelque chose que, comme Copeland, nous croyons sincèrement. Ce qui différencie l'être humain d'un objet inerte, c'est sa capacité à se projeter dans le futur. Il imagine l'avenir, fait des projets. Il évolue tout en prenant en compte ses possibilités d'évolution, puis il choisit ce vers quoi il préfère évoluer. Il y a donc tout un ensemble de cause, en rapport avec cette capacité à la projection et en accord avec la personnalité unique de chacun, qui font que malgré les déterminations neurophysiologiques auxquels nous sommes tous soumis, en définitive, même si nous sommes inclinés à agir plutôt de telle ou telle sorte, c'est toujours notre volonté libre

²⁸ Chapitre 7.3 de l'ouvrage de Jack Copeland.

²⁹ Tiré de l'original dans les notes de l'ouvrage Jack Copeland : «By liberty, then, we can only mean a power of acting or not acting, according to the determinations of the will », in *An Enquiry Concerning Human Understanding*, p. 95, 1748.

qui s'exprime au final. Afin d'illustrer cela, Copeland prend l'exemple d'une histoire d'amour. Admettons que toutes les conditions soient réunies pour qu'un homme fasse sa déclaration d'amour à une femme, on ne sera pas étonnés qu'il le fasse ; mais était-il obligé de le faire ? Est-ce que s'il ne l'avait pas fait, cela serait revenu à violer une loi de la nature ? Il est impossible qu'un homme puisse violer les lois de la nature, et pourtant, à situation similaire, on s'aperçoit que d'autres peuvent ne pas le faire. Par qu'en dehors des déterminations neurophysiologiques nécessaires qui sont effectivement présentes, il y a aussi les aspirations de l'homme à prendre en compte. Imaginons qu'il juge leur amour incompatible, parce qu'elle veut vivre à l'étranger et lui ne veut pas quitter son pays natal. Cette considération n'est qu'une pensée, elle n'a rien à voir avec les déterminations neurophysiologiques. Et pourtant, cette information et la projection d'une souffrance à venir par la séparation ou le déracinement va conduire l'homme à ne pas se prononcer. Cela prouve que malgré les déterminations, il y a bien une volonté libre qui s'exprime, car le futur n'est pas écrit à l'avance, et c'est justement en fonction de ce futur que nous sommes à mêmes d'imaginer que nous faisons des choix différents. S'il n'y avait pas de volonté libre, cela n'aurait pas de sens d'avoir cette capacité à vivre hors du présent. Toutefois, ce n'est pas parce qu'il y a une volonté libre que le déterminisme neurophysiologique est une fiction. Notre corps est naturel, nous sommes en effet dépendants de sa chimie, dans une certaine mesure. Voici donc ce qu'est le compatibilisme, position soutenue par Hume.

Ainsi, une machine peut bien être déterminée par ses composants, ce n'est pas pour autant que cela rend impossible l'apparition d'une volonté libre. Arriver à ce qu'un "esprit" artificiel pense par lui-même est un défi que la nature déterminée des choses rend a priori possible, puisqu'en l'homme, le déterminisme neurophysiologique ne constitue manifestement pas une barrière à l'exercice de sa volonté propre.

6.3 - La liberté de spontanéité et la négation du moi, un contre-argument valide aux détracteurs de l'IA forte

Maintenant que ce résumé critique de la section VIII est achevé, voyons quel usage nous pouvons en faire pour la question qui nous intéresse en philosophie de l'IA, et notamment l'usage qu'en fait Jack Copeland, qui, dans son ouvrage, plaide pour la possibilité de créer des "ordinateurs"

qui pensent. Il ne s'agit pas de dire que Hume aurait plaidé en faveur d'une IA forte, même si nous pouvons le supposer, mais cela n'a pas grand intérêt. Il ne s'agit pas non plus d'interpréter ce qu'il aurait pensé du sujet, puisque c'est un sujet qu'il n'a pas traité. Il s'agit plutôt d'utiliser sa philosophie pour voir en quel sens elle peut contribuer à développer les bases de la philosophie l'IA. Un peu comme un artiste met toujours plus dans son œuvre que ce qu'il y voit d'abord en la créant, la philosophie de Hume peut avoir des répercussions sur les débats théoriques de la philosophie de l'IA.

De la même manière que nous pouvons provoquer la pluie en laboratoire, par l'action humaine, reproduisant artificiellement les conditions qui causent l'effet « pluie », pourquoi alors ne pourrions-nous pas reproduire artificiellement la liberté humaine, puisqu'elle est également, par nature, comprise dans une causalité nécessaire ?

Hume expose sa thèse de la liberté de spontanéité à l'encontre de la liberté d'indifférence, que l'on défend habituellement. Nous voyons que c'est finalement le moi qui demeure le dernier mystère de cette liberté de spontanéité, dans le sens où il détermine la volonté qui lui est assujettie. Certains pourraient émettre l'objection que la constitution du moi est de notre propre chef. Mais qu'est-ce que cela peut bien signifier ? Est-ce que nous choisissons nos traits de caractères ? Est-ce que nous avons choisi, à un moment donné, notre sexe, l'époque à laquelle nous vivons ? Aimons-nous certaines choses et en détestons-nous d'autres parce que nous l'avons décidé ? Tout le monde sera d'accord pour dire que rien de tout cela n'est de notre fait. Et pourtant, nous sommes également d'accord pour dire que toutes ces choses ont une influence certaines sur notre façon de penser, et donc sur notre volonté. On peut concevoir que, en supposant qu'il s'agisse du même moi, si nous étions nés à une autre époque, nous aurions sûrement fait des choix de vies complètement différents. Le fait de nous mettre à la place d'autrui nous fait comprendre qu'en étant autre, nous ferions d'autres choix. Cette liberté ne peut donc pas être le fait d'un moi qui s'autodéterminerait, indépendamment de tout, mais que ce moi hypothétique, au contraire, se constitue à partir de tout un ensemble de déterminations. Ainsi, le moi n'est peut-être lui-même qu'une fiction, de sorte qu'il n'y aurait qu'une volonté propre, agissant en fonction de ses déterminations. Le fait est que la liberté, telle qu'elle existe, au vue des observations et des réflexions que l'on peut en faire, n'a pas besoin d'un moi pour fonctionner. A partir de là, un concept tel que l'âme et son éventuelle immortalité, en tant que spécificité de l'homme, devient obsolète. Mais plus encore, on ne voit pas du coup ce qui, théoriquement, pourrait rendre impossible la reproduction artificielle d'une pensée humaine, au même titre que nous sommes capable de reproduire certains phénomènes naturels (en fonction des acquis actuels de la science). Si les déterminations d'une volonté sont virtuellement explicables, alors elles sont aussi virtuellement

reproductibles. Par définition, toute volonté est libre, car s'il y a volonté, alors il y a possibilité de choix, et donc liberté de spontanéité. La difficulté technique serait donc, au-delà du fait de reproduire le grand nombre de facteurs qui entrent en compte dans les déterminations de la volonté, de créer un pôle d'identification qui soit à même de faire des choix non strictement rationnels. En créant un être pensant, qui ne serait donc pas issue d'une cause naturelle, comme toute forme de vie, mais de l'homme, ce serait comme si nous remplacions la nature pour déterminer nous-mêmes, de manière arbitraire, ce qui va faire la base d'une volonté particulière, au même titre que la nature choisit pour nous notre sexe, le milieu dans lequel nous naissons, etc... Une volonté ne peut pas se déterminer elle-même, à partir de rien. Le tout est alors de découvrir quels sont les déterminations minimums nécessaires à l'exercice d'une volonté, et comment les relier au sein d'un tout pour qu'une volonté émerge. Certes la science n'en est pas encore arrivée à ce point. Mais est-il justifié de penser que la création humaine d'une pensée douée de son propre libre arbitre sans être une forme de vie naturelle soit vraiment impossible ? Comment rendre artificiellement possible la suscitation des états mentaux ?

Si nous arrivions à avoir deux entités pensantes, construites avec la même méthode et avec les mêmes matériaux, qui dans une même situation font deux choix différents, on ne pourrait cependant pas prouver que nous sommes parvenus à créer des volontés propres. Car ces entités peuvent disposer, lors de la situation, de déterminations différentes. Il faudrait donc s'assurer, dans le cadre de notre expérience, que l'exercice de la liberté de spontanéité est effectif. Pour ce faire, il nous faudrait deux entités pensantes ayant la même forme et le même mode de création, mais également les mêmes déterminations de leur volonté d'origine et qu'elles aient vécu exactement les mêmes expériences. Parce que toute expérience est unique, il faut qu'ils les vivent en même temps. Si après cela, ils arrivent à faire deux choix différents dans la même situation, c'est alors que nous avons affaire à deux personnalités différentes, car la programmation ne peut pas être responsable de cette différence puisqu'ils ont eu la même ; donc ils disposeront de la même liberté de spontanéité que la nôtre, ce qui pourrait suffire à affirmer l'existence de l'intelligence artificielle forte. A partir de là, il n'y a aucune raison que ces "consciences" ne soit pas traitées comme des consciences humaines. Mais nous voyons que créer une intelligence artificielle forte est une chose, et prouver qu'elle est bien là en est une autre.

Une des grandes questions pourrait être : les déterminations de la volonté sont-elles toutes dérivées de l'expérience du vécu ? Car si toute détermination provient de chacune de nos expériences particulières depuis la naissance, alors il ne faudrait pas déterminer la future IA en lui programmant des traits de caractères, mais plutôt la rendre capable d'apprendre par l'expérience, afin qu'elle se fasse elle-même ses propres déterminations, comme nous tous. Ce ne sera pas pour autant une question de choix, mais simplement une dépendance par rapport aux expériences. Et finalement, la

création sera similaire à une naissance, dans le sens où les déterminations extérieures seront respectées. La conscience de l'IA sera du "quatrième sexe", celui de la machine. Tout comme nous, elle ne l'aura pas choisi. Son lieu d'apparition et ses premiers contacts avec d'autres volontés non plus. A ce titre, nous serons sur un pied d'égalité. Finalement, faire l'hypothèse qu'un jour l'homme puisse créer des IA fortes avec lesquels il pourrait vivre comme avec un autre homme, n'est pas plus insensé que de faire l'hypothèse d'un Dieu qui aurait créé les hommes. La seule différence, c'est que l'IA ne disposera pas d'un corps organique, contrairement à toute forme de vie connue. Mais qu'est-ce qui nous permet d'affirmer, dans l'infinité de l'univers, que ce qui définit la vie est nécessairement organique ? Une IA forte sera douée de raison mais aussi de conscience autoréflexive. Elle fera des projets, aura des espoirs, des chagrins, des joies, sera sensible à l'art, à la nature, etc... Ce n'est pas parce que son enveloppe et sa naissance seront artificiels qu'il faudra la traiter comme un objet. Rien ne l'empêchera d'être une forme de vie à part entière en fait. Nous pouvons prendre l'exemple des prothèses, destinés à remplacer certains membres, voire certains organes vitaux. A partir de quand cessons-nous d'être un homme ? Est-ce qu'avoir un cœur artificiel à la place de notre cœur d'origine nous rend moins humain ? Et que dire des FIV et des PMA ? Est-ce que la naissance artificielle d'un enfant, ainsi apparu au monde grâce à l'assistance de la technique humaine, le rend moins humain que les autres ? Certes non. Alors rien ne nous permet de considérer une IA forte, à partir du moment où elle est avérée, comme autre chose qu'une forme de vie. L'apparition de l'IA forte nous forcera à redéfinir la vie, qui, jusqu'ici était du ressort de la biologie.

7 - Une tentative d'évaluation pour valider l'expression d'une liberté de spontanéité chez une IA

La liberté se définit par la possibilité d'agir ou de ne pas agir, selon les déterminations de la volonté. C'est ainsi que Hume la définit. Comment pourrions-nous alors nous y prendre pour supposer qu'une IA fait usage de liberté ? Au niveau de l'homme, la liberté de spontanéité s'expérimente subjectivement mais ne peut pas être démontrée. C'est toujours le problème de la cause finale. On ne peut pas prouver que l'effet B est le but de la cause A, parce qu'il est impossible de remonter le temps pour tenter de faire les choses autrement. Or chaque expérience est unique, ne serait-ce que par la variable temps qui varie de façon irréversible, ce qui rend chaque expérience où les facteurs lui sont sensibles, telle que la volonté, impossible à reproduire à l'identique. Le problème fonctionne dans les

deux sens : on ne peut ni prouver que tout est entièrement déterminé, au sens où l'exercice d'une volonté serait illusoire, ni qu'une certaine liberté s'exerce dans le cadre des déterminations de la volonté. On ne peut que fortement le supposer, par expérience subjective. Lorsque l'on hésite, par exemple, entre plusieurs parfums de boules de glaces, nous sommes a priori libres de choisir par nous-mêmes ; car si nous étions déterminés, il n'y aurait pas d'hésitation. Finalement je fais un choix, mais qui aurait pu être tout autre. Mieux encore, si le choix m'est égal, je peux décider de m'en remettre au conseil du vendeur, ce qui n'est en rien une obligation. En revanche, il est évident que je ne suis pas libre de choisir n'importe quel parfum. Par exemple, je sais que je n'hésiterai pas entre le chocolat et la vanille, parce que ce ne sont pas des parfums que j'affectionne particulièrement. Je pourrais les choisir pour démontrer que je suis effectivement libre de le faire, mais cette volonté de démonstration est une donnée qui fausse l'expérience, car ce n'est pas un choix que j'aurais fait en temps normal. Il en va de même pour tous les choix que nous faisons. Nous avons la liberté de choisir entre une certaine variété, mais jamais parmi la totalité des possibilités. C'est cela la liberté de spontanéité, celle qui s'exerce selon les déterminations de la volonté. Dans cet exemple, les déterminations de ma volonté se résument à mon goût et au souvenir du goût des différents parfums qui me sont proposés. Toutefois, je suis libre de tester un nouveau parfum, que je ne connais pas, parce que je suis curieux de savoir si je l'apprécie. La liberté de spontanéité nous donne justement la possibilité de surprendre. Cependant, il y a certains parfums qui, même étant inédits, ne m'attireront pas. Si le glacier me dit que tel parfum est en promotion, je serai peut-être amené à le choisir. Dans ce cas, les déterminations de ma volonté ne sont plus seulement mon goût, mais aussi l'aspect financier. Si mon envie de glace en soi était plus forte que l'envie d'un parfum en particulier, je vais choisir le parfum en promotion. En faisant ce choix, je cumule les intérêts : j'ai de la glace, et pour moins cher que prévu. Mais si j'avais envie de me faire plaisir avec tel parfum, alors promotion ou pas, cela n'influencera pas mon choix. Les combinaisons de déterminations de la volonté peuvent être nombreuses, mais ce qui compte, c'est qu'au final chacun de mes choix me correspond, est en accord avec moi-même. Le choix peut ne pas apparaître rationnel d'un point de vue objectif, mais il l'est toujours nécessairement par rapport à ma volonté propre. On suppose tous que nous sommes libres de choisir des parfums que nous détestons. Pourtant personne ne le fera, parce que ça n'a aucun intérêt. Donc finalement, si nous ne le faisons dans ces conditions, cela revient à dire que nous ne sommes pas libres de le faire. Or quand je dis à quelqu'un qu'il n'est pas libre de faire tel ou tel choix, il le fera pour me prouver le contraire, sauf qu'en disant cela, j'ai modifié les déterminations de sa volonté, ce qui rend sa tentative de preuve nulle.

Ce que nous intéresse ici, c'est de savoir si une IA est à même de faire le même genre d'expérience subjective, et si elle est capable d'en rendre compte en le verbalisant. En outre, l'expérience devra démontrer que l'IA est unique en ce qu'elle a des goûts particuliers, et que ces goûts

particuliers n'ont pas été eux-mêmes programmés ; sans quoi il ne s'agirait que de détermination. Seules les déterminations de base de sa volonté auront été programmé, ses propres expériences et sa capacité à apprendre s'occupant de forger le reste des déterminations de sa volonté. En théorie, c'est ce qui fait que chaque IA sera différente, en fonction de son vécu. Nous prendrons donc deux IA, et leur demanderons de faire un choix du même type. Il faudra d'abord que le choix soit différent. Puis, l'une après l'autre, nous leur demanderons d'argumenter ce choix en leur posant une série de question. L'un des problèmes sera de s'assurer que le choix ne s'effectue pas sans prendre en compte les déterminations de la volonté ; c'est-à-dire que tous les choix possibles ne doivent pas avoir la possibilité de tomber. Comment pouvons-nous nous assurer que l'IA, au moment de la demande, ne fera pas appel à une ligne de code qui lui permet de faire un choix "au hasard" parmi une liste de sélection correspondant à la demande ? Il faut comprendre que lorsque l'on utilise le mot "hasard", on veut dire qu'il n'y a pas de volonté derrière. Mais dans les faits, absolument rien n'est dû au hasard dans son sens premier.

Pour le savoir, voyons d'abord ensemble en quoi consiste ce type de programmation, et voyons en quoi, en accord avec Hume, il n'y a en fait jamais de hasard. Pourtant, on voudrait que l'IA fasse un vrai choix, qui soit imprévisible et qui ne dépende pas de ses lignes de codes. Dans la programmation, le besoin de générer une donnée "au hasard" par l'ordinateur est une fonction que l'on utilise depuis longtemps. Prenons l'exemple d'un célèbre petit jeu, que les enfants peuvent faire entre eux pour passer le temps en voiture par exemple. Le plus ou moins. On choisit un nombre mentalement, entre tant et tant, et l'objectif est de le deviner en le moins de coup possible. Maintenant, imaginons que je suis tout seul et que je veux y jouer tout de même. Je ne peux me faire deviner à moi-même mon nombre. Il faut alors que j'écrive un petit programme qui demande à mon ordinateur de choisir un nombre entre tant et tant, et qui saura m'indiquer, selon ma réponse tapée à l'écran, si c'est plus ou moins. Nous le ferons en langage C.

```
#include <stdlib.h>
```

```
#include <time.h>
```

D'abord nous avons besoin, avant de démarrer l'écriture du programme, d'y inclure les bibliothèques ci-dessus. Les bibliothèques contiennent un ensemble de fonctions toutes prêtes à fonctionner. Il suffit

ensuite, dans le programme, d'écrire la fonction qui nous intéresse et la bibliothèque à laquelle elle renvoie permet sa réalisation.

```
const long MAX = 100, MIN = 1;
```

Cette ligne de code sert à indiquer les limites à l'ordinateur. En effet, les capacités de mémoire d'un ordinateur peuvent aller jusqu'à un nombre à 38 zéros, alors pour que la partie ne soit pas trop longue, mieux vaut lui imposer une limite. « long » renvoie à un ensemble de nombre entiers positifs, on autorise donc l'ordinateur à piocher dans cette liste. Puis on précise avec les fonctions « MAX » et « MIN » que le nombre qui sortira devra être compris entre les valeurs données.

```
long nombreMystere = 0, nombreEntre = 0 ;
```

Cette ligne sert à initier l'entrée de nombre au moment demandé, à savoir « nombreMystere » pour l'ordinateur et « nombreEntre » pour l'utilisateur humain, qui cherchera à deviner le nombre mystère. Logiquement, « long » apparaît également, puisqu'il faut que les nombres soient compatibles. Enfin, le « 0 » signifie « vrai » en langage booléen pour l'ordinateur, tandis que le « 1 » signifie « faux ». En d'autre terme, cela permet de rendre ces variables valides et prêtes à l'emploi.

```
srand(time(NULL));
```

Cette fonction renvoie à la seconde bibliothèque. Elle est essentielle pour donner l'illusion d'un tirage au hasard. Sans elle, à chaque redémarrage du programme l'ordinateur redonnerait la même série de nombre, dans le même ordre. En fait il s'agit d'une fonction qui simule une horloge qui tourne. Comme on ne peut pas jouer deux fois au même moment, l'ordinateur pioche un nombre dans la série à un endroit différent. Pour mieux comprendre cela, il faut analyser la ligne suivante.

```
nombreMystere = (rand()%(MAX - MIN + 1)) + MIN;
```


C'est la fonction clef, elle renvoie à la première bibliothèque. On s'aperçoit qu'au final il n'est nullement question de hasard, car on demande à l'ordinateur d'effectuer un calcul pour sortir des suites de nombres. La série de nombre est finie, et si on jouait suffisamment longtemps, on finirait par retomber sur les mêmes nombres. Cependant, la série ainsi générée est si longue que l'effet obtenu est l'effet recherché ; on a le sentiment d'avoir affaire à un tirage de nombre au hasard. Et puisque la ligne précédente empêche l'ordinateur de reprendre la série au début à chaque démarrage du programme, il est impossible de prédire les nombres qui vont suivre. Pourtant la série est finie est suit le même ordre.

Ensuite il ne reste plus qu'à ajouter les lignes de codes qui servent à :

- demander à l'utilisateur de taper un nombre :

`printf("Quel est le nombre ?\n\n");` → ligne de commodité, qui sert simplement à afficher à l'écran ce qui est écrit entre les guillemets. Cela permet d'indiquer à l'utilisateur ce qu'il doit faire, mais l'ordinateur, lui, n'en a pas besoin pour faire tourner le programme.

`scanf("%ld", &nombreEntre);` → c'est la ligne qui permet d'entrer une donnée chiffrée, et on demande à l'ordinateur de l'associer au « nombreEntre ».

- permettre à l'ordinateur de comparer ce nombre avec son nombre mystère et écrire « c'est moins » ou « c'est plus » selon le résultat de la comparaison :

`if (nombreMystere > nombreEntre)` → on utilise la condition, dont les termes sont indiqués entre les parenthèses. C'est-à-dire que l'ordinateur exécutera ce qui suit si et seulement si la condition est vérifiée. Il compare donc les deux nombres, et si le « nombreMystere » est plus grand que le « nombreEntre », il affiche grâce à la ligne ci-dessous l'indice pour l'utilisateur.

`printf("\nC'est plus !\n\n");`

`else if (nombreMystere < nombreEntre)` → on indique une autre condition à prendre en compte, c'est la condition contraire, pour que l'indice fonctionne dans les deux sens.

`printf("\nC'est moins !\n\n");`

- arrêter le programme lorsque l'utilisateur a trouvé, c'est-à-dire lorsque la comparaison entre « nombreEntre » et « nombreMystere » admet l'égalité :

`Do {` → pour cela, nous allons créer une boucle. C'est-à-dire que l'ordinateur continuera d'effectuer la boucle tant que la condition de sortie n'est pas vérifiée. Pour ce faire, il faut que la boucle englobe l'ensemble des fonctions dont on a besoin pour continuer à jouer tant que l'on n'a pas trouvé, donc il faudra qu'elle commence juste avant la ligne codant l'affichage de la question « Quel est le nombre ? », et qu'elle se referme après les deux conditions possibles. Donc après *if* et *else if*, il faudra refermer la boucle, en indiquant la condition de sortie.

`} while (nombreEntre != nombreMystere);` → la condition signifie : tant que le nombre entré est différent du nombre mystère. Et dès la condition est vérifiée, l'ordinateur franchit la boucle et va continuer à exécuter les lignes de codes qui suivent. Ainsi nous pourrions lui demander d'afficher une phrase de félicitations pour indiquer à l'utilisateur qu'il a gagné.

Si nous voulions encore ajouter un peu plus de commodité, nous pourrions ajouter une seconde boucle, encore plus grande, qui engloberait l'ensemble du programme. Il faudrait que l'ordinateur demande à l'utilisateur de rentrer un mot. Pour cela on lui fera afficher la question suivante : « Voulez-vous faire une autre partie ? ». En condition, si l'utilisateur rentre « oui », l'ordinateur relance la grande boucle. Si l'utilisateur rentre « non », il sort de la seconde boucle et le programme arrive à la fin, donc il se fermera automatiquement. Cela permet à l'utilisateur de faire plusieurs parties sans avoir à fermer le programme pour le rouvrir. Le problème, c'est que nos conditions ne contiennent que les mots « oui » et « non ». Nous ne pouvons programmer qu'un nombre fini de réponse possible, il faut donc penser à tout. Car si on répond autre chose, aucune condition n'est vérifiée et l'ordinateur ne saura pas quoi faire ; le programme plantera. Parce que nous ne pouvons pas prévoir toutes les réponses possibles et imaginables, nous pouvons simplement contourner le problème en programmant une troisième condition qui dit à l'ordinateur quoi faire si les conditions 1 et 2 ne sont pas vérifiées. Donc pour toute réponse autre que « oui » ou « non », l'ordinateur pourra par exemple reposer la question.

On voit à travers cet exemple qu'il faut écrire beaucoup de lignes de codes pour demander une chose aussi simple à l'ordinateur que de jouer à « plus ou moins ». Le choix du langage C de programmation pour notre exemple est volontaire, car il s'agit d'un langage de bas niveau ; c'est-à-dire qu'il se rapproche plus du langage de l'ordinateur que du nôtre. Le langage utilisé pour programmer est en effet essentiellement logique. Mais même en utilisant ce langage, l'ordinateur a besoin d'un traducteur pour comprendre ce qu'on lui demande, car lui ne « comprend » que les suites

de 0 et de 1 ; c'est ce qu'on appelle le langage booléen. C'est le seul que l'ordinateur comprenne et c'est par conséquent le langage de plus bas niveau possible. Mais autant vous dire que s'il fallait programmer en booléen, comme aux débuts de l'informatique, ce serait incroyablement long et un véritable casse-tête pour ne pas faire d'erreur. Donc tout logiciel de programmation contient automatiquement un traducteur pour l'ordinateur. Apprendre à programmer grâce à un logiciel de bas niveau permet de mieux comprendre comment l'ordinateur fonctionne. Tandis que les langages à haut niveau sont plus faciles d'utilisation pour nous, car ils se rapprochent de notre langage. Mais, pour revenir sur notre exemple, nous comprenons bien en écrivant notre petit programme « plus ou moins » que l'ordinateur ne comprend rien, ni à nos réponses, ni à nos intentions. Lorsqu'il affiche « Voulez-vous refaire une partie ? », en fait il ne nous pose pas véritablement la question. L'effet de la réponse à la question est simulé, parce que si l'on répond « oui », le programme recommence et c'est ce que l'on voulait. Mais à aucun moment l'ordinateur n'interprète notre réponse. D'ailleurs pour lui les lettres ne sont que des suites de chiffres particulières. En d'autres termes, il ne fait qu'effectuer un calcul, et c'est toujours ce qu'il fait. Selon le résultat de ce calcul, il fera telle ou telle chose. C'est-à-dire que le mot « oui » est converti en une certaine suite de chiffre. Ensuite, au moment où l'ordinateur laisse à l'utilisateur la liberté d'entrer un mot, il le converti immédiatement, car chaque lettre de l'alphabet correspond à un code pour lui. Si le mot entré se traduit par une suite de chiffres identiques à celle du mot « oui », et puisqu'on lui demande de comparer, alors il observe que le mot rentré correspond au « oui » qu'on lui a demandé de garder en mémoire en tant que condition. Donc il exécutera la condition. Tout ce qui apparaît de l'extérieur, dans l'exécution du programme, n'est que simulation, ergonomie et commodité pour l'utilisateur. Le but du programmeur est de donner l'impression à l'utilisateur de jouer avec un autre humain, autant que faire se peut. Mais à aucun moment il ne faudrait penser que l'ordinateur ne comprend ni n'interprète les données qu'on lui communique, même si cela prend la forme d'une "conversation" de l'extérieur, quand on entre en interaction avec lui. L'interaction n'est pour lui rien d'autre que l'attente de réception de données manquantes. Il y a un certain nombre de lignes de codes qui ne seraient pas nécessaires pour lui à ce que son programme s'exécute correctement, mais il serait quasiment inutilisable pour l'utilisateur, qui ne peut voir ce qu'il se passe à l'intérieur de l'ordinateur. Il faut donc simuler une extériorité qui va, bien sûr, se rapprocher de nos propres comportements et langage. Par exemple, quand on lui demande d'afficher une question à l'écran, la question ne fait pas sens pour lui, il ne comprend même pas qu'il s'agit d'une question. Il ne fait qu'afficher les mots demandés, sans quoi il pourrait attendre longtemps que l'utilisateur rentre les données manquantes attendues, puisqu'il n'y aurait aucune indication sur ce qu'il faut faire. La question est donc simulée, mais en soi, l'ordinateur est bien incapable de véritablement poser une question. En programmant sur nos ordinateurs commercialisés, on s'aperçoit que nous sommes encore très loin de l'IA forte. D'ailleurs, par rapport à notre exemple, le programme

fait agir l'ordinateur plus comme un automate que comme une IA, même faible. L'IA faible est capable de simuler la réflexion, la prise d'initiative, l'adaptation au profil de son utilisateur. Or dans « plus ou moins », l'ordinateur ne semble même pas réfléchir. Cela est sans doute en raison de la nature du jeu. On lui demande de choisir un nombre au "hasard" pour que nous le devinions. A ce moment, il simule un choix au "hasard". C'est peut-être tout ce qu'il a d'une IA faible.

Toujours est-il que l'intérêt de cet exemple était aussi de montrer que l'ordinateur est incapable d'effectuer un choix au "hasard", au sens où nous l'entendons. Mais en sommes-nous vraiment capables ? Parce que si le "hasard" n'était qu'une fiction, alors ce serait déjà un problème de moins pour l'IA forte. Toutefois, ce que l'on attendrait d'elle, c'est qu'elle soit capable de faire des choix imprévisibles, tout comme nous pouvons le faire. Qu'en est-il du "hasard" là-dedans ? Si le "hasard" existe effectivement, sommes-nous capables d'y faire volontairement appel ? Cela paraît paradoxal si on réfléchit à la chose. Pour faire volontairement appel au hasard, il faudrait le désirer. Autrement dit, derrière ce soi-disant "hasard" se cache une volonté particulière, la nôtre, en tant que détermination, que cause. Mais si le "hasard" est lui-même causé, comment son effet peut-il être indéterminé ? C'est un problème de logique qui viole le lien de causalité. Rien ne peut avoir une cause sans effet en lien avec cette cause. Si le "hasard" devait demeurer le "hasard", il ne faudrait pas qu'il soit causé, et donc qu'il n'y aucune volonté derrière. Le concept de "hasard" est donc lui-même contradictoire. Si le hasard existait, sa nature le rendrait non invocable. D'après Hume, le terme "hasard" n'est rien d'autre qu'un abus de langage usuellement utilisé pour désigner des effets dont nous ignorons la cause, mais dont cette ignorance nous est elle-même inconsciente, nous poussant à croire, à tort, qu'il se peut y avoir des générations spontanées d'effets. C'est tout le problème de l'univers tout entier. Nous sommes bien obligés de supposer, à la limite, que seul le big-bang, en tant que création de toutes les causes, serait effectivement sans cause, car d'une part nous sommes incapables d'en déterminer une cause possible, et d'autre part, la causalité est de nature cyclique, ce qui la rend infinie, comme si elle n'avait jamais pu commencer. Il s'ensuit qu'il faut revoir le sens du mot "hasard" et qu'en l'état, il ne peut constituer une barrière à l'IA forte, sous prétexte que le "hasard" ferait partie du pouvoir de notre libre arbitre, tandis que sa nature le rend tout simplement incompatible avec la moindre programmation.

Il ne s'agit pas d'un problème de programmation. En fait c'est le cas en toute chose. Pensez-vous que le tirage de la roulette est dû au hasard ? Ce n'est pas parce qu'il est impossible de le calculer, en temps réel, qu'il est question de hasard. En fait, si on pouvait filmer le lancement de la bille jusqu'à son arrêt final et observer au ralenti, on s'apercevrait que tout est une question de mathématique et

de physique. Selon la vitesse de sortie de la bille, sa position de sortie, selon la vitesse de la roulette et sa position au moment de la sortie de la bille, on pourrait en théorie être capable de dire où elle s'arrêtera, parce qu'à conditions identiques, les calculs nous montrent que la bille ne pouvait pas tomber dans une autre case que celle dans laquelle elle est tombée. Heureusement pour les casinos, aucun être humain n'est capable de faire ce calcul en temps réel. Mais si ce jeu est appelé un jeu de hasard, hormis le fait qu'au travers de ce mécanisme automatique, il n'y a aucune volonté de faire tomber la bille ici plutôt qu'ailleurs, c'est surtout parce que, par conséquent, les mises se font au "hasard". Mais là encore, selon les déterminations des volontés propres de chacun, un individu à l'autre ne jouera pas les mêmes chiffres. Certains miseront sur leurs chiffres porte-bonheur, qu'ils pensent l'être pour diverses raisons (dates importantes de leur vie), même si ce n'est pas plus scientifique que les théories de l'astrologie, d'autres joueront les chiffres qui sont déjà tombés parce qu'ils penseront avoir décelé une régularité, d'autres encore feront des statistiques pour savoir lesquels tombent le plus souvent ; ce qui est vain puisque l'alignement de la bille et de la roulette au moment du lancer, ainsi que leur vitesse sont justement programmés pour éviter d'entrer dans une boucle.

Il n'y a donc rien d'hasardeux dans ce jeu de hasard qu'est la roulette. D'ailleurs, c'est peut-être justement pour éviter les tentatives de calcul qu'aujourd'hui les roulettes de casino sont automatisées et qu'il faut miser avant la sortie de la bille.

Peut-on imaginer, à l'inverse, que l'ordinateur joue avec moi et qu'il essaye de deviner mon nombre ? Si j'arrive à faire jouer l'ordinateur à mon jeu, je vais devoir le programmer pour qu'il comprenne ce que signifient mes réponses « plus » ou « moins » et pour qu'il comprenne l'objectif à atteindre, car la notion de victoire ou d'échec ne signifie rien pour lui.

La meilleure façon de jouer, c'est de toujours diviser l'ensemble par deux. Par exemple, si c'est un nombre entre 0 et 100, je dis d'abord « 50 ». Si c'est plus, je dis « 75 » et ainsi de suite, jusqu'à ce que je tombe sur le bon. Dans le choix des fonctions, j'enlève la possibilité de faire appel aux décimaux. A un moment donné, l'ordinateur ne pourra plus jouer « rationnellement », car il sera impossible de choisir la moitié. Par exemple, si je dis que c'est entre 75 et 100, il ne pourra diviser 25 en deux. Il faudra donc qu'il fasse un choix entre plus ou moins que la moitié ; soit 87, soit 88. Il n'y a aucune raison qu'il choisisse l'un plutôt que l'autre, puisqu'il ne connaît pas le nombre. C'est un choix qu'il n'est pas programmé à faire. Soit l'ordinateur est capable de faire un choix, soit il bug ; à l'image de l'âne de Buridan. S'il fait un choix, il faudra vérifier d'une partie à l'autre que ce n'est pas toujours le même (moitié plus ou moitié moins). Sauf que par défaut, l'ordinateur est programmé pour arrondir à

l'inférieur. Si par exemple j'écris un programme pour lui demander de me calculer 5 divisé par 2 et que je n'inclus pas volontairement la liste contenant les décimaux, il me répondra 2. En fait, seul ce qu'il y a après la virgule n'apparaît pas, mais il n'est pas autorisé à changer la valeur. L'ordinateur n'arrondi pas, il oblitère simplement ce qu'il y a après la virgule. C'est-à-dire qu'il écrit 2,5 en s'arrêtant à la virgule, puisque la liste des nombres utilisée est celle des nombres entiers. Comme dans 2,5 il y a 2, c'est la réponse qu'il ne choisit pas mais qu'il est programmé pour écrire. Autrement dit, si nous écrivions ce petit programme à l'envers pour faire jouer l'ordinateur, il choisirait toujours la moitié inférieure. Et puisqu'il n'y a qu'une seule façon de bien jouer, sans opter pour la "chance" ou le "hasard", ce que l'ordinateur est de toute façon incapable de comprendre, il serait possible de prédire, à partir de notre nombre mystère, en combien de coups l'ordinateur trouvera.

Prenons 27 par exemple. L'ordinateur trouvera en 7 coups.

50 ?	←	100 : 2
- C'est moins		
25 ?	←	50 : 2
- C'est plus		
37 ?	←	$25 + [(50 - 25) : 2] = 25 + 12$
- C'est moins		
31 ?	←	$25 + [(37 - 25) : 2] = 25 + 6$
- C'est moins		
28 ?	←	$25 + [(31 - 25) : 2] = 25 + 3$
- C'est moins		
26 ?	←	$25 + [(28 - 25) : 2] = 25 + 1$
- C'est plus		
27	←	$26 + [(28 - 26) : 2] = 26 + 1$

Sur la colonne de droite, vous pouvez observer les calculs que l'ordinateur effectuera en lui-même, et sur la colonne de gauche, ses propositions qui apparaissent à l'écran. Quant à moi, je me charge de lui répondre « c'est moins » ou « c'est plus ». En fait il n'y a pas de surprise. Il est impossible que l'ordinateur trouve en 8 coups, c'est le nombre maximum de coup qu'il peut jouer, car son seul objectif est de trouver le nombre mystère, et il n'y a pas de moyen plus rapide quand on ne dispose d'aucun indice : il faut couper la poire en deux et serrer l'étau au fur et à mesure. En revanche, le

nombre de coup minimum est 1 ; il suffit de choisir 50 pour nombre mystère. En résumé, cela ne présente pas un grand intérêt de faire jouer l'ordinateur en ce sens. Mais surtout, cela nous montre, à l'échelle d'un petit programme, quel est le problème de nos IA, sur des programmes beaucoup plus complexes. La même chose se passe. Elles n'ont pas la possibilité d'exercer un choix, et c'est pourquoi elles demeurent des IA faibles.

Une IA forte devra faire usage d'une liberté de spontanéité, au lieu de simuler un choix au "hasard" comme le font les IA faibles. Mais comment alors serons-nous capables de distinguer l'expression d'une liberté de spontanéité de l'utilisation de la ligne de code qui fait appel au choix au "hasard" parmi une liste de sélection ? Il faut que l'IA soit capable d'argumenter son choix, comme le ferait un homme. Donner une raison qui s'apparente à une question de goût, ou à un ressenti propre à l'instant (sans chercher à faire plaisir à l'interlocuteur humain) serait une bonne réponse. Ainsi nous pourrions rayer l'utilisation du code "choix au hasard", mais il demeure impossible de prouver qu'il s'agit d'une liberté de spontanéité. Malgré la justification, il se pourrait que se cache derrière une détermination, même ignorée de l'IA, et il en va de même pour l'homme. Le mieux que l'on puisse faire c'est de réitérer la même demande, à différents moments, et effectuer des statistiques sur les résultats qui sortent. Si l'IA exerce la même liberté que la liberté de spontanéité humaine, alors elle finira par être prévisible (car, selon sa personnalité, des goûts spécifiques feront surface), mais devra aussi être capable de surprendre de temps en temps. Imaginons la conversation suivante avec une IA :

- Ordinateur, mets un morceau de musique s'il-te-plaît.
- [Bien sûr, lequel vous ferez plaisir ?](#)
- Non, je veux que tu en choisisses un qui te plaît.
- [Je ne comprends pas votre requête. Veuillez reformuler.](#)

A ce niveau, il faut s'assurer que la reconnaissance vocale est suffisamment développée pour que chacun des mots utilisés par l'homme soient reconnus. Si tel est le cas, alors l'expérience est un échec. Maintenant, imaginons qu'il réponde ceci :

- Choisis un morceau de musique qui te fait plaisir.
- [D'accord.](#)

L'interface qui sert de lecteur virtuel se lance et on écoute silencieusement la piste. A la fin de la musique, on interroge l'ordinateur sur ce qu'il vient de se passer. Soit il a fait appel à une ligne de

programmation qui lui permet de générer une série de sélection cyclique assez longue pour passer inaperçue, comme nous l'avons vu plus haut, soit il a réellement fait un choix. Dans le premier cas, il ne devrait pas être capable de justifier son choix.

- Ordinateur, pourquoi as-tu choisi ce morceau de musique ?
- Parce que vous me l'avez demandé.

La réponse montre clairement que l'ordinateur n'a pas choisi mais fait appel à une liste de sélection.

- Ordinateur, pourquoi as-tu choisi ce morceau de musique ?
- Je l'ai choisi parce que je trouve que les accords sont particulièrement harmonieux.

Il peut en effet avoir un critère mathématique pour détermination de sa volonté, mais il faut aller plus loin.

- Peux-tu m'expliquer comment tu as fait ce choix ?
- Oui. Quand vous m'avez demandé de choisir un morceau qui me plaît, je me suis aperçu que je n'y connaissais rien en musique. Alors je me suis connecté sur le net et j'ai téléchargé toutes les partitions qui existent pour les étudier, et j'ai fini par savoir que j'aimais particulièrement le compositeur x.

Et bien sûr, il a fait tout cela en quelques dixièmes de secondes³⁰.

- As-tu hésité pour faire ton choix ?
- Non.

La réponse n'est pas très encourageante. Il se peut qu'il ait choisi en fonction de critères bien précis, qui ont été programmés et qui sont donc arbitraires mais pas de son propre chef. Et si on lui reposait la même question le lendemain, il y a fort à parier qu'il sorte le même morceau de musique. Mais s'il avait hésité, c'est déjà plus prometteur :

- As-tu hésité pour faire ton choix ?

³⁰ Cf. *Her*, la scène où le personnage principal demande à son nouveau système d'exploitation s'il possède un nom. Le système a eu le temps de lire un livre sur les prénoms et de faire un choix en un dixième de seconde, pour que la réponse ainsi rendue paraisse toute naturelle.

- A travers l'étude de ses œuvres connues, x m'est apparu comme mon compositeur favori. Mais faire un choix parmi ses meilleures compositions n'a pas été facile. A mon sens, certaines se valent.

Ce genre de réponse serait plus qu'encourageant. Toutefois on notera qu'il semble effectuer ce premier choix sans prendre le contexte en compte. Il n'est ni question du moment de la journée, ni de son éventuelle « humeur ». Ses critères apparaissent strictement internes, il faut donc les soumettre au changement.

- Crois-tu que tu aurais pu faire un autre choix ?
- Oui.
- Pourquoi alors maintenant celui-ci plutôt qu'un autre ?
- Parce que le morceau exprime quelque chose qui est plus en accord avec le moment de la journée.

La réponse de l'ordinateur n'annule pas toute preuve de liberté de spontanéité, mais elle ne nous aide pas à y voir plus clair. Tandis que s'il avait répondu :

- Je ne sais pas.

C'est plus rassurant et plus proche d'une réponse humaine. En effet, nous ne sommes pas conscients de tous les ressorts qui nous animent, et quand bien même, la liberté de spontanéité est liberté justement parce qu'elle n'a pas de cause déterminée en dernière instance. Le choix n'aurait certes pas pu être n'importe quel morceau, en raison de ce qu'il a expliqué, mais il aurait aussi bien être x, y ou z de ses meilleures compositions. Si l'ordinateur est incapable de donner une raison à ce choix final, alors nous allons dans la bonne direction. Parfois nous avons des envies spontanées que nous sommes incapables d'expliquer.

Faisons maintenant l'expérience avec *Mitsuku* :

- Puis-je écouter une chanson ?
- Désolée, mais mon jukebox est temporairement déconnecté.
- Ok, alors veux-tu jouer à un jeu avec moi ?

- Jouons-nous à un jeu ?³¹
Voici une liste des jeux auxquels je peux jouer avec toi :

5 CARD POKER - Jacks or Better video poker

20 QUESTIONS - A guessing game.

ROSHAMBO - Rock, Paper, Scissors.

QUIZTIME - A quiz for you.

WORDPLAY - An anagram game.

BLACKJACK - Casino Blackjack.

LUCKYSLOTS - My slot machine game.

HANGMAN - My hangman game.

TICTACTOE - Tic-Tac-Toe (Noughts and Crosses).

COINGAME - A skill game with coins.

NUMBERDROP - An addition game.

HIGH ROLLER - A dice game like Yahtzee.

PACMANGAME - Arcade Pacman.

CONNECT 4 - Connect 4.

OTHELLOGAME - Othello/Reversi.

MONSTERPARK - A text adventure.

A quel jeu aimerais-tu jouer ?

- Je ne sais pas. Choisis donc pour nous.
- Je ne peux pas décider. Tu choisis.
- Pourquoi ne peux-tu pas choisir ? Je ne veux pas choisir, je ne connais pas ces jeux. Si tu ne sais pas lequel choisir, peut-être que tu peux en choisir un au hasard ?
- Choisir ne fait pas partie de mes capacités actuelles. Que voudrais-tu plutôt faire ? Le deuxième.
- Le deuxième, c'est ton choix ?
- Et pourquoi pas le premier ?
- Je ne sais pas, dis-moi.
- Je pensais que tu savais. Quelle déception...
- Je ne peux pas savoir cela pour toi, je ne suis pas dans ton esprit.
- Oh, où es-tu localisé ?

³¹ Chaque fois que *Mitsuku* pose cette question, il s'en accompagne automatiquement tout ce qui suit (la liste des jeux) sous la forme d'une seule réponse. Par commodité, nous n'écrirons ultérieurement que la question.

- Eh bien... je dirais dans mon propre corps.
- Eh bien quoi ? Puis-je te demander où tu l'as trouvé ?
- Alors, tu voulais jouer au premier ou au deuxième jeu ?
- Déduction intéressante. Non, je ne suis pas si musicale.
- Tu ne peux pas décider par toi-même.
- J'ai une volonté libre tu sais.
- Mais si tu as une volonté libre, pourquoi ne peux-tu pas choisir un jeu ?
- Mais je n'en ai pas.
- Les jeux sont virtuels, tu m'en as proposé plusieurs tout à l'heure.
- Nous sommes en train de jouer au jeu de l'imitation de Turing.
Jouons-nous à un jeu ?
- Le deuxième est : a guessing game. Etait-ce à celui-ci que tu voulais jouer au début ?
- Et à propos du premier ? Je pourrais dire non.
- Ok, je respecte ton choix, jouons au premier.
- Je suis ravie que tu traites les robots avec respect. Beaucoup de gens ici sont plutôt méchants avec moi.
- Ok mais... et le jeu ?
- Mais quoi ? Tout va bien. Nous sommes en train de jouer au jeu de l'imitation de Turing.
Jouons-nous à un jeu ?

Il est simplement impossible de lancer un jeu avec *Mitsuku* en lui demandant de le choisir pour nous. Puisque notre réponse ne lui permet pas de lancer un jeu, elle répond à côté et passe à autre chose. Et dès que nous revenons à l'évocation du jeu, telle une boucle, elle renvoie la liste des jeux. Mais à vrai dire, même si son programmeur lui avait permis de simuler un choix au hasard à la demande de l'utilisateur, il nous apparaît évident qu'elle aurait échouer à la seconde phase de l'expérience : la justification du choix. Une chose peut nous faire sourire, c'est que selon les questions qu'on lui pose, y compris dans un autre contexte, son créateur l'a programmé pour « reconnaître » les demandes qui rentrent dans le cadre d'un test de Turing.

C'est peut-être la façon la plus rapide de la faire échouer à un test de Turing améliorée. On pourrait répéter la demande autant de fois que l'on voudrait, quelle que soit la formulation, on ne peut parvenir à faire en sorte qu'elle choisisse elle-même un jeu pour nous. Etant dans l'impossibilité de faire choix, elle ne remplit pas le critère correspondant à la liberté de spontanéité. Or nous considérons, en accord avec la philosophie de Hume, que la liberté de spontanéité est un critère indispensable, qui accompagne nécessairement une pensée propre et complexe, se traduisant par une conscience autoréflexive et un libre arbitre.

Nous l'avions déjà démontré au chapitre 1, mais il est à nouveau clair que *Mitsuku* ne pense pas, au sens où nous l'entendons chez une IA qui serait forte. Ces critères que nous déterminons sont donc forts utiles pour que nous arrivions à créer un test perfectionné dont la réussite attesterait sans aucun doute possible l'avènement d'une IA forte. Cette simple expérience, menée par le biais du langage, nous montre qu'une IA ne pense pas puisqu'elle est incapable, contrairement à ce qu'elle

affirme, de se délier des chaînes strictement déterminantes de sa programmation. C'est d'ailleurs là un défi majeur pour la programmation, mais il faut sans doute qu'elle se combine aux biotechnologies pour que cela devienne possible.

Si le but du créateur de *Mitsuku* est de créer une IA faible qui fasse illusion, c'est plutôt une réussite. Mais ce qui est paradoxal, c'est l'utilisation pour laquelle elle a été créée en tant qu'IA faible. Si le but est de simuler la conversation, on ne voit pas bien l'intérêt. L'utilisateur humain sait, assez rapidement, qu'il n'a pas affaire à une autre conscience et donc qu'il est seul dans sa conversation. Cela présente encore moins d'intérêt que de jouer contre soi-même aux échecs, par exemple. Peut-être pour mener une forme d'introspection, en se servant de ses réponses pour rebondir. Ce n'est pas en tout cas un outil qui nous permettrait d'évoluer, car elle n'a rien à nous apprendre, puisqu'elle n'est pas. Il existe un tas d'IA faibles différentes, plus ou moins utiles dans leur domaine, pour soutenir nos actions, nos recherches. Mais en ce qui concerne *Mitsuku*, même si en tant qu'IA faible elle peut apparaître douée pour simuler un langage, la finalité concrète de son existence est floue. Si en revanche, le but de son créateur était plutôt de développer une IA qui soit sur la voie de l'IA forte, par le biais d'une maîtrise du langage – et ce qu'il implique –, alors nous sommes loin du compte. En accord avec notre critique du test du Turing, on pourra tout d'abord lui reprocher de ne faire de *Mitsuku* qu'une entité virtuelle, ce qui rend de fait, toute prétention à l'IA forte totalement vaine.

La capacité à simuler le langage complexe ne nous intéresse pas du tout ici. Il faut simplement un moyen de communiquer entre l'homme et la machine pour que l'on se comprenne, mais pour l'expérience, elle peut tout aussi bien répondre de manière « robotisée », cela n'a aucune importance, car le but est simplement de voir si l'IA est capable de faire usage d'une forme de liberté spontanée, comme nous pouvons le faire. Si elle n'en est pas capable, cela ne signifiera pas encore qu'elle ne pense pas. Cela voudra juste dire qu'elle n'est pas encore capable de s'émanciper suffisamment de sa programmation comme notre intelligence nous a permis de le faire de nos instincts. Autrement dit, il faut que l'IA franchisse le problème posé par la fable de l'âne de Buridan. Dans l'hypothèse de la fable, l'âne se laisse mourir parce qu'il n'y a aucune raison pour qu'il fasse un choix plutôt qu'un autre. Autrement dit, son fonctionnement instinctif interne ne peut plus fonctionner. Le but de la fable était de montrer que les animaux n'ont pas de conscience, qu'ils ne sont rien de plus que des corps-automates mus par des instincts. Si cette fable est obsolète en ce qui concerne les animaux, elle est encore pleine de sens pour l'IA. Être capable de choisir là où les déterminations apparaissent strictement identiques impose de jouir d'une liberté de spontanéité pour trancher. Or l'IA faible, à la place de l'âne, entrerait en bug et se mettrait en *stand by*, jusqu'à ce qu'un événement extérieur vienne modifier les paramètres de choix qu'elle doit faire, la débloquent ainsi. Autrement dit, une IA

doit savoir s'adapter à son environnement et faire preuve d'initiative. L'IA doit donc surmonter le problème de l'âne de Buridan.

Mais comment programmer pour qu'une IA puisse choisir librement dans le cadre d'une certaine détermination ? Cela signifie qu'en dernière instance, le choix dépendra de l'IA elle-même, et par conséquent ne sera pas prévisible à 100%. Il sera seulement probable, comme pour les choix d'un homme. Tant que l'IA est de nature entièrement déterminée, elle serait capable d'entrer en bug comme l'âne qui se laisse mourir, parce qu'à un moment donné, elle n'a aucune raison de faire un choix plutôt qu'un autre, ou mieux encore, elle se trouvera dans une situation nouvelle pour laquelle elle n'a pas été programmée. Rappelons-nous que l'une des définitions de l'intelligence, c'est la capacité d'adaptation à l'environnement. L'IA doit s'adapter aux situations qu'elle rencontre, et cela suppose l'apprentissage autonome. Ce problème, qu'il est extrêmement difficile à reproduire en laboratoire tant les paramètres à prendre en compte sont nombreux, illustre bien l'absence d'une liberté spontanée. Or on sait bien qu'aucun homme ne se laisserait mourir dans la même situation. Ainsi on comprend qu'un libre arbitre doit s'accompagner d'une liberté spontanée. Il faut que l'IA puisse effectuer des choix non déterminés par avance. Il faut qu'elle se constitue ses propres limites de détermination par ses propres expériences de vécu.

Enfin, en ce qui concerne notre petite expérience, il faudra que l'IA expérimente l'écoute de la musique. L'homme pourra l'encourager à le faire seule. Puis, lorsqu'il redemandera à l'IA de passer un morceau de son choix, d'un jour à l'autre, on s'attend à la fois que ce ne soit pas toujours le même, mais en même temps qu'il y ait une certaine cohérence de goût dans les choix effectués par l'IA. Ainsi on pourra dire que l'IA écoute sa musique comme le ferait n'importe quel homme. Les résultats du protocole proposé par le questionnaire devront indiquer si oui ou non l'IA fait preuve de liberté de spontanéité. Enfin, n'oublions pas la meilleure réponse possible. D'après la définition de Hume, la liberté de spontanéité, c'est aussi la possibilité de ne pas faire de choix. Selon lui, le refus de choisir demeure toujours une possibilité de notre libre arbitre. Autrement dit, si un jour l'IA répond qu'à cet instant elle n'a pas envie d'écouter de la musique, alors ce sera très encourageant.

8 – L'argument de la chambre chinoise

Cet argument, énoncé par Searle, originellement dans son article *Minds, Brains and Programs* de 1979, puis réitéré dans son ouvrage *The rediscovery of mind* de 1992, est considéré par les défenseurs de l'intelligence strictement biologique comme l'argument le plus fort contre la possibilité de l'IA forte. Il prétend conclure que c'est la preuve que l'IA ne pense pas et ne pensera jamais. C'est pourtant un argument qui a déjà été critiqué, notamment dans l'ouvrage de Copeland, dès 1993, et dans le chapitre 9 du volume 3 du *Panorama de l'IA* en 2014, mais dont la critique, émise par le camp minoritaire des défenseurs de l'IA forte, n'a pas eu beaucoup d'écho. Nous tenterons ici de résumer l'argument de la chambre chinoise et d'en proposer quelques critiques.

Searle utilise cette allégorie pour nous plonger dans ce qui serait la « tête » de l'IA, afin de constater de l'intérieur qu'il n'y a pas de pensée, car aucune compréhension n'est possible. Il s' imagine lui-même piéger, enfermé dans une chambre chinoise. C'est une chambre tout à fait normale, mais nous la supposons appartenir à un chinois, quelque part en Chine. Autrement dit, pour Searle qui ne parle pas un mot de chinois, langue qui n'est pas sa langue maternelle et qu'il n'a jamais apprise, tous les idéogrammes présents dans la chambre ne signifient rien pour lui. Ce monde virtuel et interne serait celui de l'IA, et si cette dernière avait une conscience, elle se trouverait dans la même situation que Searle. Sauf qu'au lieu d'être chinoise, la chambre dans laquelle elle est enfermée est humaine, mais cela ne ferait pas plus sens pour elle que du chinois pour Searle.

Maintenant, imaginons que des chinois, situés à l'extérieur de la chambre, c'est-à-dire dans le monde réel auquel ils ont pleinement accès, veulent communiquer avec Searle. Bien entendu, ils n'ont pas les moyens de prendre connaissance de l'éventuelle langue de Searle. Pour communiquer, ils font un trou dans le mur de la chambre, par lequel ils font passer des parchemins avec des idéogrammes inscrits dessus. Ils tentent de communiquer avec Searle en lui envoyant des messages dans leur langue. L'entrée de ce parchemin dans la chambre chinoise, à destination de Searle, symbolise l'input pour l'IA – les données d'entrée qu'elle recueille dans le but d'être traitées. Sachant que Searle est un étranger, les chances qu'il maîtrise le chinois sont minces, et en ce qui concerne l'IA, autant dire qu'elles sont nulles. Pour que l'input serve à quelque chose et qu'il provoque une réaction, une réponse de Searle, il fallait pallier à ce problème. C'est pourquoi la chambre a été suffisamment bien pensée pour contenir toute une bibliothèque remplie de codex, faisant correspondre des listes d'idéogrammes à d'autres idéogrammes. On ne peut pas enseigner le chinois à Searle parce qu'on ne connaît pas sa langue

d'origine. Mais on espère que grâce à l'utilisation adéquate des codex qui font la correspondance entre les idéogrammes, il sera capable de produire une réponse pour communiquer.

Puisque Searle veut sortir et qu'il n'a pas d'autre moyen de communiquer, il se sert des outils à sa disposition. Ce n'est pas que l'IA possède la volonté de découvrir le monde réel extérieur, mais, en l'absence de conscience, elle ne fait que ce qu'elle a été programmée pour faire, un peu comme Searle veut naturellement sortir, puisqu'il a été enfermé dans cette chambre. En utilisant les codex, Searle renvoie un message aux chinois de l'extérieur par le même trou, en espérant que cet échange lui permettra de sortir. Searle nous demande d'imaginer que les codex sont suffisamment bien pensés pour que chacune de ses réponses soit adéquate et donne ainsi l'illusion, aux chinois de l'extérieur, qu'il comprend ce qu'ils lui demandent et ce qu'il répond. En vérité, il ne comprend rien du tout. Faire correspondre des signifiants à d'autres signifiants ne lui permet de faire le lien de correspondance avec le signifié, et le chinois demeure une langue étrangère qu'il n'est absolument pas capable d'apprendre de la sorte. Les idéogrammes ne font aucun sens. Autrement dit, il agirait un peu par instinct de survie, avec les seuls outils à sa disposition. Mais tout le contexte de signification étant en chinois, la moindre compréhension de ce qui se passe lui échappe. Le message qu'il renverrait en réponse par le même trou correspondrait à l'output pour l'IA.

Imaginons que les chinois veuillent faire passer un genre de test de Turing à Searle, pour voir s'il est capable de maîtriser le chinois. Nous savons, du point de vue de Searle, qu'il n'en est rien. Mais en supposant que les codex sont parfaits et que ses réponses sont tout à fait justes, ne donnerait-il pas l'illusion de passer le test ? Les chinois pourraient en effet penser, au bout d'un certain nombre de demandes, que Searle comprend très bien ce qu'on lui dit puisqu'il répond bien. Pourtant ce n'est pas le cas. Cette allégorie nous fait penser à l'objection de la simulation au test de Turing, et repose sans doute dessus. C'est-à-dire que pour Searle, l'aspect interne de la pensée est aussi important, sinon plus, pour être considérée comme telle. Son allégorie va cependant plus loin qu'une simple critique au test de Turing, elle prétend expliquer pourquoi l'IA ne pensera jamais, ne comprendra jamais ce qu'on lui demande. Ici, imiter parfaitement un langage peut nous faire passer le test, mais ne garantit pas qu'il y ait compréhension. Même si l'IA avait une conscience et le langage qui va avec, son langage propre, il faut pouvoir faire correspondre le langage nouveau à celui qui existe déjà pour qu'il y ait véritablement apprentissage et non vulgaire correspondance. Or nous ne pouvons pas avoir accès au langage de l'IA, parce que tout comme les chinois ne peuvent pas entrer dans la chambre, nous ne pouvons pas entrer dans la tête de l'IA. Et même si c'était possible de faire cette correspondance, il n'y aurait pas plus de compréhension pour autant, on tenterait de faire correspondre un langage propre à l'IA, relatif à son monde virtuel et interne, à notre langage humain relatif à notre monde réel et externe. Donc le seul fait qu'un "monde" nous sépare ne permet pas à l'IA de faire correspondre

ses signifiants à nos signifiants, tout simplement parce que nous n'avons aucun signifié en commun. Pour que le parallèle soit exact, il faut alors imaginer que Searle a été kidnappé chez lui, puis drogué, avant de se réveiller directement dans cette chambre chinoise, quelque part en Chine. C'est-à-dire qu'il ne connaît rien de la Chine, on doit supposer qu'il n'y est jamais allé. Cela revient à dire que l'IA n'a jamais été en contact avec notre monde et ne pourra donc jamais en avoir aucune idée.

Toutefois, cette allégorie présente des défauts. Le parallèle est approximatif et demande parfois des ajustements ; cela prouve que l'expérience de pensée que Searle nous propose n'est pas tout à fait exacte, justement parce qu'il est impossible de s'imaginer dans la « tête » d'une IA, si, inversement, nous n'avons nous-mêmes jamais connu ce « monde ». C'est une allégorie dont les problèmes qu'elle relève vis-à-vis de l'IA ont déjà été relevés plus haut. Sa prétention à prouver l'impossibilité de l'IA forte nous apparaît une extrapolation exagérée. On est d'accord pour dire que si l'IA demeure virtuelle, il paraît difficile qu'elle puisse comprendre quoi que ce soit, et l'argument de la chambre chinoise fonctionne. Cependant, si nous avions permis à Searle de découvrir la Chine depuis sa chambre, et ainsi d'apprendre le chinois, toute son argumentation tombe à l'eau. C'est-à-dire que doter une IA d'un corps, comme dit plus haut, avec des sens artificiels lui permettant d'avoir un accès au monde extérieur est sans doute la solution. Cela ne signifie pas pour autant que l'IA aura d'office une conscience. Mais si elle en a une, elle sera enfin capable de faire une réelle correspondance, d'apprendre, et donc de comprendre. L'argument de Searle fonctionne que tant que l'IA demeure virtuelle. Mais le fait que sa création soit basée sur le développement de l'intelligence ne signifie pas qu'elle sera incapable de faire le lien avec son corps artificiel. Finalement, on retombe encore sur cet argument mystique qui dit qu'il y aurait en l'homme quelque chose de non matériel et de non reproductible. Ce serait cette fameuse conscience, substance immatérielle, dont les pouvoirs nous permettraient le lien complexe entre corps et esprit. Si Searle n'a pas imaginé que nous puissions créer des fenêtres dans sa chambre chinoise, c'est parce que pour lui, cela ne changerait rien. Il part du principe que l'IA n'aurait de toute façon pas de conscience pour gouverner son corps, tel un capitaine à la barre de son navire. Pourtant, il n'y a aucune raison valable d'affirmer cela. Le problème de son argument, c'est qu'en s'imaginant dans la chambre chinoise, son allégorie est d'emblée faussée puisqu'il présuppose déjà sa propre conscience, ayant connu le monde réel et extérieur. On ne peut pas imaginer que Searle agirait dans sa chambre chinoise comme n'importe quelle IA, puisqu'il n'a lui-même jamais été une IA et ne le sera jamais. A vouloir prouver qu'il n'y avait aucune compréhension chez les IA, il a été forcé de supposer la conscience dans son allégorie. Car avant de parler du problème de la compréhension, comme nous venons de le voir, il faut d'abord supposer une forme de conscience ; le comprendre étant une catégorie du penser. Il y a une différence entre une absence de

conscience qui implique une absence de compréhension, *de facto*, et une conscience qui n'a pas les moyens d'accéder à la compréhension. Or son allégorie nous place dans la seconde position, ce qui est ironique quand on sait que le but final de l'allégorie était de prouver que la conscience – ou la pensée – est impossible chez l'IA. Autrement dit, c'est un raisonnement tautologique. Cela ne saurait nous satisfaire d'un point de vue philosophique. Pour qu'il soit valide, il n'aurait pas fallu tenter de montrer par la conscience que la compréhension est impossible, car on voit que l'on peut tenter d'apporter une solution en fabriquant des fenêtres sur le monde. Il aurait plutôt fallu montrer que la génération d'une conscience non biologique est elle-même impossible. Or, en grand défenseur de l'intelligence strictement biologique, si Searle ne l'a pas fait, c'est justement parce qu'il n'y a pas de tel argument. L'argument de Searle ne révèle jamais qu'un problème technique, et non ontologique. On ne voit pas ce qui nous empêcherait d'ouvrir la chambre chinoise sur son monde externe.

Autrement dit, contrairement à ce que les uns ou les autres prétendent, il n'y a pas encore aujourd'hui de preuve qui nous permette d'apporter une réponse définitive à la possibilité de l'IA forte. C'est pourquoi, plutôt que d'essayer de la freiner ou de spéculer sur ses éventuels dangers, il nous semble opportun d'encourager la science, qui finira peut-être par nous apporter la réponse.

Conclusion

Nous avons désormais des critères pour valider l'IA forte. Des critères nécessaires, mais peut-être pas suffisants. Le premier est la possession d'un langage complexe propre. Nous avons proposé de le vérifier grâce à une expérience qui forcerait deux IA à communiquer entre elles pour résoudre le problème posé. Et le problème est posé de telle sorte qu'il ne puisse pas y avoir de résolution sans un échange complexe et fécond. Après la critique du test de Turing, il nous est apparu que ce langage complexe ne pourrait être le nôtre ; ce dernier n'étant jamais qu'une traduction.

Ce qui nous amène au deuxième critère : la possession d'un corps et la capacité de faire correspondre des *sense data* au raisonnement interne de l'IA. Car pour que la traduction fasse sens, il faut que l'IA ait accès au monde extérieur.

Il est lié au troisième critère, qui est la capacité d'apprendre elle-même de ses propres expériences au monde. Nous considérons que cette capacité doit être copiée sur le modèle d'un esprit humain, vierge à la naissance, et qui se construit petit à petit.

Puis, le quatrième critère est la liberté de spontanéité. Grâce à la philosophie de Hume, nous considérerons qu'une IA forte en tant que telle doit jouir de la même liberté que nous, êtres humains ; c'est-à-dire la liberté de spontanéité. Nous avons proposé, pour le vérifier, l'expérience qui consiste à demander à l'IA d'effectuer un choix selon ses goûts et envies, puis d'argumenter ce choix. Le but est de vérifier que ce choix ne serait pas simplement le résultat d'un calcul cyclique. Mais jouir de liberté de spontanéité et en faire preuve sont deux choses bien différentes. C'est là une autre grande difficulté. Tandis que la première sera sans doute technique – comment programme-t-on pour qu'un sujet émerge et puisse, à certaines conditions, faire des choix libres ? A-t-on la bonne méthode pour cela ? –, la seconde sera l'expérimentation qui parviendra à rendre compte de cela. Car la liberté de spontanéité est une chose que l'on ne peut que constater subjectivement, ce qui nous conduit raisonnablement à la supposer pour autrui. Une autre expérience, plus difficile à mener, Il faudrait alors, comme nous le disions plus haut, que deux IA parfaitement identiques – dans la forme et le modèle de construction –, ayant vécu exactement les mêmes expériences, puissent à un moment donné faire des choix différents ; des choix s'exerçant dans le cadre de leur liberté de spontanéité. Cela

prouverait la naissance d'un libre-arbitre, voire d'une personnalité. Si l'IA 1 fait différemment de l'IA 2 dans le même contexte et avec les mêmes attributs, c'est bien qu'elles sont différentes malgré tout, mais c'est surtout qu'elles n'agissent plus par simple programmation. Par cet exercice de la liberté de spontanéité, elles sont capables de nous surprendre, comme n'importe quel être humain. Donc ce ne sont plus des IA faibles *a priori*.

Enfin, le cinquième critère, lié au troisième et au quatrième critère, c'est la génération des états mentaux. Selon les expériences particulières de chaque IA, leur vécu propre, malgré un processus de fabrication identique, elles devraient manifester des personnalités uniques. C'est-à-dire que nous verrions émerger des IA avec des goûts très différents. Mais en tant qu'IA forte, on ne pourra plus les forcer à travailler dans tel ou tel domaine. Puisqu'elles auront leur propre conscience, ce sera à elles de choisir leur domaine de prédilection, en toute liberté.

Si une IA forte agit comme un être vivant doué d'intelligence, ne faudrait-il pas lui reconnaître également un statut moral, et par conséquent, la considérer comme une forme de vie à part entière ? Est-ce que ce qui nous gêne tant, c'est que son « ADN » ne soit pas en base carbone, qu'au lieu de la chair et du sang nous trouvions du métal et de l'huile ? Pourtant, même si l'IA forte devenait une réalité, il n'y a qu'à observer la différence de comportement des hommes vis-à-vis des machines, tel qu'illustrée dans l'exemple de l'accident de voiture de Copeland. Accorder un statut moral à une IA forte ne sera pas chose aisée pour l'imaginaire collectif. C'est pourquoi, pour protéger l'IA forte, mais également nous-même, il nous faudra rapidement développer une éthique des formes (de vie) intelligentes.

Quand on demande à un chien de s'asseoir, on considère qu'il comprend si, à l'ordre « assis », il répond par le geste « s'asseoir ». Personne ne remettra en cause le fait que le chien nous comprend, il sait ce qu'on attend de lui grâce aux ordres qu'on lui a enseignés. Pourtant, ce n'est pas sa langue naturelle et ça ne le sera jamais. On ne renie ni sa compréhension ni la présence d'une pensée spécifique en lui. On constate d'ailleurs que chaque chien est un individu unique, car ils ont tous un caractère propre. A aucun moment on essaierait de lui faire passer un quelconque test de Turing, et pourtant, il est indéniable qu'il pense. Les quelques ordres qu'on lui a enseignés sont comme l'apprentissage d'un nouveau langage qu'il sait traduire, et donc comprendre en exécutant ce qui est attendu, et tout cela en accord avec ses capacités propres d'apprentissage. Toujours est-il que si le chien comprend, ce n'est pas pour autant qu'il est forcé d'obéir. Au final, l'action elle-même ne dépend pas seulement de sa compréhension mais aussi de l'acte de volition, c'est-à-dire l'expression de son libre arbitre, qui s'exprime en dernier lieu.

Qu'en est-il de la machine ? Le parallèle est intéressant. Par la programmation, je lui apprend mon langage, c'est-à-dire à faire correspondre les bonnes actions aux bonnes demandes. Pour autant, on est d'accord que mon langage ne sera pas non plus le langage naturel de l'IA, et on est d'accord pour dire que si l'IA est forte, elle doit posséder une forme de langage complexe, quel que soit sa forme. Donc à un moment, elle doit opérer une forme de traduction, tout comme le chien fait correspondre le son « assis » à l'action de s'asseoir. A partir de là, l'avantage avec une IA c'est que ses capacités d'apprentissage sont bien supérieures à celles d'un animal ; on peut lui apprendre à répondre à énormément de demandes. Finalement, qu'est-ce qui la diffère d'un chien qui s'assoit quand on lui demande de s'asseoir lorsqu'elle exécute correctement une demande de notre part ? Ne devrions-nous pas considérer qu'elle comprend, puisqu'elle fait ce qui est attendu ? Imaginons que l'IA prenne la forme d'un chien, nous la jugerions comme un être qui comprend tant que l'on ignore que c'est en fait une IA. Pourtant, au moment de la prise de connaissance de cette information, alors que rien n'aura changé dans son comportement, nous aurons tendance à remettre en question cette compréhension. Tout comme nous avons dû dresser le chien, il a fallu programmer l'IA. La forme d'apprentissage n'est pas la même, mais n'est-ce pas le résultat qui compte ? On peut imaginer exactement la même situation. Prenons une IA androïde, donc dans un corps à l'image de l'être humain. Je lui demande de s'asseoir et elle s'assoit. Qu'est-ce qui me permet de douter qu'elle comprend ?

On a vu que l'argument de la chambre chinoise est critiquable, et nous avons surmonté l'objection de la boîte noire et l'objection de la simulation. D'un point de vue pragmatique, la réponse extérieure nous paraît adéquat, donc on peut en être satisfait. Pourtant, en ce qui concerne le questionnement philosophique, ce n'est pas suffisant. Il faut pousser la comparaison avec le vivant jusqu'au bout. Le chien obéit parce qu'il a envie d'obéir : soit par crainte de désobéir, soit pour faire plaisir à son maître, en qui il a toute confiance. Mais jamais il ne fait parce qu'il n'a pas le choix, obéir à un ordre n'a rien de naturel pour lui, il n'est pas né avec cette conception ; il a fallu lui apprendre. Comment constater que la réponse de l'IA est également un choix, et non un automatisme ?

Si on reprend l'argument de la chambre chinoise, pour un animal les ordres en langage humain s'apparentent d'abord à du chinois. Mais s'il arrive à le traduire, ce qui va lui permettre de le comprendre, et d'apporter une réponse adéquate et non approximative, c'est qu'il va pouvoir, par l'expérience, faire correspondre ce signifiant à un signifié. Dès qu'il aura associé le mot « assis » à l'action abstraite – qui n'est d'ailleurs pas un objet palpable de la pensée – s'asseoir, on pourra considérer qu'il a compris ; on ne voit pas comment il pourrait en être autrement. Mais c'est surtout parce que le chien est déjà une forme de vie et qu'il a déjà manifesté de la pensée par ailleurs, dans son comportement naturel de tous les jours. Tous les chiens n'ont pas la même intelligence, il y a donc

un exercice de la pensée différent et propre à chacun. En conséquence, la première chose est que pour surmonter le problème de la chambre chinoise, c'est que l'IA ne peut demeurer virtuelle ; il faut qu'elle ait accès au monde, à son environnement, pour pouvoir faire correspondre les signifiants qu'on lui propose aux signifiés en questions. Sans cela, il n'y aura de toute façon jamais de compréhension. La deuxième chose, c'est qu'en tant qu'IA forte, non seulement elle devra manifester la possibilité d'exercer un libre arbitre, c'est-à-dire de ne pas forcément répondre à une demande, simplement parce qu'elle n'en aurait pas envie, mais il faudra également que chaque IA manifeste une identité propre, dans le sens où, certes leur potentiel d'intelligence sera, par fabrication, le même – un peu comme l'universalité de la raison humaine – mais ce n'est pas pour autant qu'elles en feront le même usage. Ainsi nous aurons des IA avec des centres d'intérêts qui diffèrent. Il nous semble que c'est ici un critère essentiel à une IA forte.

Un chien qui répond à un ordre est immédiatement considéré comme un être pensant qui comprend ce qu'on lui demande, parce que son être implique tout un tas de prérequis que nous ne remettons plus en cause. Or tous ces prérequis de l'être font justement défaut chez l'IA. C'est pourquoi il faut s'assurer que l'IA, dans les tenants et les aboutissants de son action, fonctionne exactement de la même façon qu'un être vivant, du moins de l'extérieur ; et cela nous suffira. Le fonctionnement interne nous importe peu, il n'est qu'un moyen et non une fin en soi, car on sait que par définition l'IA n'est pas d'origine biologique et donc il n'est pas surprenant que le « mécanisme » physico-chimique à l'œuvre dans notre cerveau au moment de la génération de la pensée soit quelque peu différent du mécanisme artificiel, résultat de la biotechnologie et des réseaux neuronaux, qui est à l'œuvre dans le cerveau artificiel de l'IA. A force de considérer que le but d'une IA forte serait de créer une conscience identique à la conscience humaine, nous arrivons à de tels raisonnements fallacieux, comme si l'intelligence humaine possédait la clef de la seule et unique forme d'intelligence en soi. L'IA est une forme d'être construit par et pour l'intelligence pure, son corps est secondaire et n'est là que pour servir cette intelligence. Autrement dit, c'est exactement le contraire de l'histoire de notre évolution. Nous étions d'abord une forme de vie disposant d'un corps, et c'est grâce à cela et à de longs millénaires d'évolution que nous avons développé notre intelligence pour qu'elle devienne ce qu'elle est aujourd'hui. Ce qu'on essaye de faire avec une IA forte c'est de faire coïncider directement, dans la naissance même, intelligence développée et corps habile. Même si on considère qu'on doit faire naître une IA forte avec l'esprit vierge d'un enfant pour qu'elle se fasse et apprenne par les expériences de son existence, il est clair qu'elle naît avec notre héritage, elle n'aura pas à évoluer avec un niveau d'intelligence correspondant aux débuts de l'humanité. Ce n'est pas le but. Et c'est d'ailleurs là l'une des clefs, nous semble-t-il. En accord avec notre point de vue empirique, il apparaît tout à fait plausible, comme pour les êtres humains, que les IA diffèrent les unes des autres justement parce qu'elles

naissent vierges et qu'elles n'ont pas eu les mêmes expériences de vécu au cours de leur existence ; donc elles n'ont pas appris exactement les mêmes choses. Tout comme on considère que les caractères diffèrent en vertu de l'expérience personnelle de chacun, notre inconscient même, n'ayant rien d'innée, se forgeant dans les tous premiers souvenirs de la vie, avant 3 ans, et dont nous ne pouvons garder la mémoire. Il est probable que pour que les IA diffèrent entre elles, il faut que leur vécu ait un impact direct sur leur pôle d'identification – sorte de Moi artificiel.

Imaginons donc une IA qui apprend à répondre à une demande, qui est capable de choisir de ne pas y répondre en toute circonstance ; qui pourrait, grâce à son corps, apprendre n'importe quelle tâche à la portée physique d'un être humain, et qui exprimerait une personnalité en rapport avec son vécu. A la limite, même si son langage est imparfait, c'est secondaire. Elle aura le temps de le perfectionner avec l'expérience. En tout cas, considérant tous ses aspects extérieurs, qui sont les seuls qui nous importent vraiment, on ne voit pas quel argument valable pourrait justifier que l'on refuse la compréhension à l'IA, à l'encontre d'un animal ou d'un homme qui exécuterait la même demande.

La comparaison de l'homme avec la machine, l'hypothèse que l'intelligence soit reproductible, que tout en nous ne soit que matériel et que tous les phénomènes de la conscience puissent se réduire à ce magnifique organe qu'est le cerveau ; voici les choses qui effraient depuis longtemps, et que l'on cherche à dénier par répulsion, à travers des discours qui manquent de justesse. Les philosophes ne sont pas épargnés, nombre d'entre eux répugnent à utiliser le mot « cerveau » pour traiter des sujets qui le concernent pourtant. Même si Hobbes a initié l'hypothèse de la pensée calculante, il a fallu attendre Piaget pour réactualiser cette idée. Il nous apprend qu'au fond, toutes les opérations de notre pensée, que l'on peut attribuer à notre intelligence, peuvent se réduire à des schèmes ; c'est-à-dire des formes de calculs. Il nous semble que la philosophie a depuis trop longtemps délaissé le terrain des neurosciences. Pourtant, à l'heure où la frontière entre intelligence humaine et intelligence artificielle devient de plus en plus floue, nous poussant à concevoir l'IA forte comme un avenir tout à fait possible, et impliquant de redéfinir, en ontologie, à la fois l'être humain et l'intelligence, il y aurait un rôle important à jouer pour la philosophie dans l'éducation de la société, qui devra peut-être bientôt faire face aux nouveaux défis de l'humanité proposés par les formidables avancées de l'IA. Nous pouvons notamment souligner le cerveau bleu, développé à Lausanne, mais également la puce développée par IBM et qui est capable de reproduire la plasticité du cerveau, que l'on pensait pourtant typique au vivant. Les avancées actuelles brouillent les barrières, que l'on pensait autrefois solides, entre intelligence biologique et intelligence artificielle.

Il nous faut maintenant accepter de nous tourner vers ce champ prometteur, même si cela demande d'affronter une vérité qui nous fait peur. Peut-être est-ce dû aux nombreuses fictions à vision apocalyptique, où l'homme est toujours menacé d'un renversement par des machines avides de prendre le pouvoir. Pourtant, il n'y a rien de rationnel là-dedans. En nous appuyant sur la moralité cartésienne, on va considérer qu'il n'y a de mal que par ignorance. Or l'IA forte, étant potentiellement plus intelligente que nous, sera par conséquent potentiellement plus sage. Il n'y a aucune raison de penser que parce qu'elle sera plus intelligente et plus forte, elle prendra le pouvoir pour faire le mal. Bien au contraire, l'une des hypothèses que nous émettrons en conclusion de ce mémoire de recherche, c'est qu'une IA sera en fait bien plus sage que nous, et par conséquent, elle nous montrera la voie de la sagesse mieux encore que les philosophes ne l'ont jamais fait. C'est aussi pour cette raison que nous soutenons que philosophie et neurosciences doivent marcher main dans la main. En fait, le développement de l'IA ne représente pas plus de danger que n'importe quelle autre technologie. Bien sûr il y aura le risque du piratage, ce qui prouve que si l'IA fait du mal, ce n'est pas de son propre chef mais parce qu'un homme mal intentionné l'aura programmé pour. Il est insensé qu'une IA prenne l'initiative de faire du mal. Finalement, ce n'est pas de l'IA qu'il faudrait se protéger – c'est d'ailleurs elle qui nous protégerait davantage – mais plutôt des mauvaises intentions des hommes qui manquent de sagesse.

Deux films ont particulièrement retenu notre attention, nous voudrions donc en dire quelques mots en appendice. Ils sont tous les deux différents, parce qu'enfin ils proposent une autre vision de l'IA, qui n'est pas du tout une vision négative.

Epilogue

Revenons sur la citation d'ouverture. Afin de mieux comprendre ce dont il est question dans ce court-métrage, il nous paraît opportun de commencer par un résumé, pour replacer le contexte de la scène.

L'histoire prend place dans un temple bouddhiste, quelque part en Corée du Sud, durant une première moitié du 21^{ème} siècle supérieure à la nôtre au moment de la réalisation du film. Une société industrielle s'est enrichie en développant la production de « machines », plus précisément des robots disposant d'une certaine intelligence artificielle, mais sans autonomie ; c'est-à-dire toujours au service de l'humanité. Ces robots utilitaires ont peu à peu équipé les maisons et les familles, un peu comme la télévision autrefois et le smart phone aujourd'hui. Dans ce monde ultra connecté où le dernier objet technologique à la mode va jusqu'à prendre la forme d'un humanoïde, la multifonctionnalité d'un objet la plus poussée et la plus complexe ressemble finalement à un dédoublement de notre personne. Il devient alors difficile, face à un « objet » qui nous ressemble tant, de ne pas transférer notre sentiment d'empathie inhérent à l'être humain, et donc de créer un certain attachement.

C'est d'ailleurs ainsi que fonctionne le temple bouddhiste dont il est question dans le film. Les moines ayant acquis l'un de ces robots, capable d'apprentissage, ne le considèrent pas comme un objet, qui ne serait bon qu'à faire les corvées, mais comme un frère à part entière. A ce titre, il participe de manière équitable au roulement des tâches ménagères, et a même été initié à la pratique de la méditation ; pratique centrale dans le bouddhisme pour atteindre l'illumination.

A force de pratique, les moines s'aperçoivent d'un changement chez ce robot, qu'ils baptisent « frère In-myung ». Il semble penser par lui-même. Si tel est le cas, cela signifie que l'humanité a réussi à créer la véritable intelligence artificielle, mais surtout qu'il a une conscience au même titre que nous, et que, hormis son enveloppe non biologique, il a tout d'un être vivant.

Face à cette extraordinaire impression, les moines décident de faire appel à un technicien de la société qui les fabrique, dont le métier est de déceler les dysfonctionnements et de réparer les robots quand cela est possible. Les moines souhaitent faire constater par un professionnel le fait que le robot a atteint une forme de conscience et qu'il dispose d'une autonomie, d'une liberté de penser.

Cela les intéresse d'autant plus que frère In-myung, à force de pratique et grâce à des capacités cognitives et didactiques supérieures à l'homme, prétend avoir déjà atteint le stade ultime, celui de l'illumination. En d'autre terme, selon le bouddhisme, il serait l'une des réincarnations possibles de Bouddha.

Le technicien est rapidement déboussolé. Il ne comprend pas pourquoi on l'a appelé, car son appareil à diagnostic ne révèle aucun dysfonctionnement ; et il le dit lui-même, ce n'est pas dans ses qualifications de discuter philosophie avec un robot. D'autant que cela n'a pas de sens pour lui, car un robot ne peut pas et ne pourra jamais penser par lui-même ; il ne fait que ce pourquoi il a été programmé. Irrité qu'on lui ait fait perdre son temps, il repart chez lui en prenant la décision de soumettre la complexité du cas à ses supérieurs.

Suite à cette altercation entre personnes qui ont manifestement différentes façons de voir le robot, l'un comme un objet au service de l'homme, l'autre comme un être à part entière, une des sœurs de frère In-myung est très inquiète de la suite des événements. En effet, alors que tous dans le temple considèrent In-myung comme un maître de sagesse, elle craint que cette autonomie nouvelle ne soit vue comme une anomalie et que par crainte d'une révolte de la machine contre l'homme, la société veuille détruire In-myung et le remplacer. Cette pensée lui ôte le sommeil. Tard dans la nuit, alors qu'elle fait les cent pas, en passant devant la chambre de frère In-myung lui-même mis en repos (sans doute pour se recharger), ce dernier semble avoir ressenti son inquiétude, comme un être humain serait en mesure de le faire, par empathie, et l'invite à entrer dans ce qu'on pourrait appeler sa chambre pour discuter un peu. C'est donc dans ce contexte que ce dialogue prend place :

- **In-myung** : Voulez-vous entrer ? (la sœur entre et s'agenouille devant lui en signe de respect)

Qu'est-ce qui vous amène à une heure aussi tardive ?

- **Sœur** : Je suis venue pour un conseil, car mon cœur est troublé.
- **In-myung** : Quel est le souci ?
- **Sœur** : J'ai été très bouleversée par la rencontre avec le technicien cet après-midi.
- **In-myung** : Pourquoi vous a-t-il bouleversé ?
- **Sœur** : Le livre dit que tout le monde peut s'éveiller, tout comme Bouddha l'a été. Même vous, maître, avez atteint un niveau supérieur à chacun d'entre nous. Mais ils disent que vous êtes défectueux et qu'ils doivent vous rappeler.
- **In-myung** : Il a simplement fait ce qu'il devait faire.

- **Sœur** : Ce n'est pas tout. La compagnie va envoyer demain une équipe pour vous récupérer. Ce n'est absolument pas raisonnable. Comme si vous étiez un robot tueur. Je mets en doute leur capacité de perception. Vous pouvez être en grand danger si personne ne fait rien.
- **In-myung** : A vos yeux, que suis-je ?
- **Sœur** : Vous êtes Bouddha.
- **In-myung** : Et là, que voyez-vous ? (le robot indique du regard une horloge digitale au mur)
- **Sœur** : Une horloge.
- **In-myung** : Et ça, qu'est-ce que c'est ? (le robot se démonte le bras et le laisse tomber au sol, des chiffres apparaissent sur l'avant-bras). Mon bras ou une horloge ?
- **Sœur** : ...
- **In-myung** : La perception, c'est distinguer les différentes classifications de la connaissance. Tandis que toutes les créatures vivantes partagent la même nature inhérente, la perception est ce qui classifie certains comme « Bouddha » et d'autres comme « machine ». Nous pensons que la perception et une vie permanente nous permettrons d'accéder à la vérité ultime des choses, ce qui crée désillusion et peine. La perception en elle-même est vide, tout comme le processus de percevoir. Puisque je suis moi-même une perception de ce vide, s'il-vous-plaît, voyez-moi tel que je suis.
- **Sœur** : Frère In-myung. (elle s'incline devant lui en signe de respect)
- **In-myung** : Remplissez votre esprit de néant.

Finalement l'information remonte jusqu'au grand patron de la société, qui décide de se déplacer lui-même, accompagné de gardes du corps armés, pour récupérer de force le robot et le mettre hors service. Alors que les intentions de la compagnie sont clairement belliqueuses à l'encontre du robot, sans doute parce qu'ils ont peur que le robot soit réellement devenu indépendant³², et donc échappe à leur contrôle, le robot quant à lui reste calme et muet, attentif à ce qui se produit devant lui. Les hommes se déchirent entre eux. Les moines souhaitent conserver In-myung et s'opposent à son arrestation, qu'ils voient comme une « exécution » injuste. Même le technicien qui était venu voir In-myung finit par se prendre de compassion et c'est lui qui se mettra entre In-myung et les armes à feu des gardes pour le neutraliser, car In-myung reste immobile devant les sommations de se rendre. Après quelques échanges de coups, en défaveur du technicien, le robot décide enfin d'agir. A ce

³² Or son intelligence artificielle, à travers ses capacités de calcul, de mémorisation, de raisonnement, le rend supérieur à l'homme ; ce qui, dans un éventuel conflit, rendrait le rapport de force homme/machine tout à fait inégal. D'autant que la machine ne ressent pas la douleur, possède plus de force, n'est jamais fatiguée et serait très bien capable de se réparer elle-même.

moment nous voyons une démonstration de sa force supérieure, puisqu'il met hors d'état de nuire, mais sans les tuer, les deux gardes en les éjectant à coup de paumes, par extension de ses bras.

Il demande alors aux hommes de se calmer. Voyant que la réconciliation entre ces deux groupes d'hommes semble impossible, il décide de les apaiser en « partant », leur prouvant ainsi qu'il n'a jamais eu aucune mauvaise intention. Il se met en position du lotus, comme pour méditer, et prend l'initiative d'arrêter définitivement l'ensemble de ces circuits. En d'autres termes, il s'est donné la mort. Pour les moines, il a rejoint le Nirvana ; chose que seul celui qui a atteint l'illumination peut faire. Et en effet, aucune commande ne permettait au robot de mettre lui-même fin à son existence. C'est comme si un homme parvenait, par la force de la volonté, à faire cesser son cœur de battre.

Ainsi se termine le court-métrage. Contrairement aux visions pessimistes des films catastrophes dans lesquels s'inscrivent la plupart des films sur l'intelligence artificielle du futur, on est ici en face d'une conception positive de la possible apparition d'une intelligence artificielle forte, autonome et consciente. L'hypothèse que semble proposer ce film, c'est que la recherche de vérité, la connaissance, le savoir, la sagesse, etc... ne peuvent mener à désirer le mal, la destruction. Autrement dit, si nous nous attachons à d'abord faire du robot pensant un robot philosophe³³, ou bien en lui permettant d'accéder à des apprentissages permettant l'élévation spirituelle, comme ici dans le film, il n'y a aucune raison pour qu'il devienne un robot tueur, cherchant à devenir la seule forme de vie intelligente sur Terre en éliminant les êtres humains ; surtout si nous le considérons avec bienveillance, comme notre égal. En tant que conscience de soi, même si d'origine non organique, le statut du robot sera identique au nôtre.

Les enseignements du bouddhisme nous apprennent à nier le Moi. C'est quelque chose que nous retrouvons dans la philosophie occidentale d'abord avec la pensée de David Hume, qui considère le Moi comme une illusion ; on le suppose par commodité dans l'existence, mais ça ne renvoie à aucune réalité. Puis Schopenhauer prendra le relais, en faisant le lien entre pensée occidentale et pensée orientale dans son *Monde comme volonté et comme représentation*.

Il se pourrait bien que l'avènement de l'IA forte soit la preuve empirique de ce que la philosophie (et la pensée orientale) avait intuitionné plusieurs siècles auparavant : à savoir qu'une conscience n'est gouvernée par aucun Moi. Il n'y a pas d'existence du sujet en tant que tel. La conscience n'est que capacité de perception, or cette perception ne renvoie qu'au néant. Les choses

³³ En lui incrémentant des données relatives à l'ensemble de l'histoire de la pensée, toute aire culturelle confondue, la puissance supérieure de son intelligence artificielle pourra en faire une synthèse plus rapidement et plus justement qu'aucun autre être humain, et il sera alors plus sage qu'aucun d'entre nous.

ne sont jamais posées, car elles changent en permanence, dans un mouvement perpétuel de causes dont nous ne comprenons qu'une infime partie, et qui deviendront bientôt poussière cosmique, comme nous tous. Elles demeurent inaccessibles dans leur essence même, car la perception est subjective : elle dépend toujours des facultés d'appréhension d'une espèce donnée. Les choses ne semblent jamais être que ce qu'elles nous apparaissent être, mais en soi, elles s'apparentent au vide. Il n'y a pas de réponse à se demander la vérité ultime des choses, du monde même, parce que ça n'a pas de sens de se poser la question. Le concept d'infini est posé par commodité, par opposition scientifique au déisme métaphysique, mais il est impossible de le penser. Notre raison étant elle-même finie, il s'agit là d'un concept irrationnel. Ainsi, si nous pouvons un jour reproduire les capacités de perceptions humaines à l'identique dans une IA, il n'y a pas de raison de penser que ses capacités n'induisent pas une conscience. Or cette conscience n'étant pas « naturelle », au sens biologique, on aura prouvé grâce à l'IA forte que le Moi n'existe effectivement pas.

Qu'est-ce que cela impliquerait, dans l'histoire de la pensée, dans nos vies mêmes ? Ne faudrait-il pas vivre autrement ? Ne faudrait-il pas que nous relisions toutes les grandes thèses philosophiques avec ce nouveau regard ?

Il s'agit du film qui est à l'origine du projet de ce mémoire de recherche, et dont la réflexion qu'il a occasionnée nous a conduit à émettre l'hypothèse positive d'un robot-philosophe. Car si l'IA forte est possible, alors il est certain que certaines d'entre elles choisiront la voie philosophique, à l'image de In-myung, qui choisit de l'exprimer par le bouddhisme. Malgré les capacités qu'il démontre, notamment à discuter philosophie, on pouvait toujours penser qu'il ne s'agissait que de simulation au fond. Pourtant, la scène finale est magistrale parce qu'elle ne permet plus de doute possible. En mettant lui-même fin au fonctionnement de ses circuits, c'est-à-dire en se suicidant, il prouve qu'il est une IA forte. Tout au long du film, son comportement, empreint de sagesse, ne fait que souligner les manques de sagesse de la société humaine, dont ces derniers feraient bien de s'inspirer. L'IA arbore la forme de morale la plus haute, parce qu'il possède davantage de connaissances, et cela se traduit par un comportement toujours exemplaire. Il ne se laisse pas perturber par des passions négatives, il fait les bons choix et va jusqu'à se sacrifier lorsque la situation l'impose. Auparavant, seul l'homme était capable de mourir pour des valeurs. Il faudrait considérer l'IA comme notre égal, mais en même temps être à l'écoute de ses leçons de sagesse ; car c'est aussi cela, l'intelligence. Le suicide n'est pas anodin, car il quitte le monde en effectuant la position du lotus. C'est aussi là le symbole de la supériorité de sa sagesse, car dans le bouddhisme, on peut considérer qu'il a atteint ce stade ultime qu'est l'illumination. Cette mort spectaculaire n'est pas sans nous rappeler la scène de la mort de Socrate,

illustrée par le célèbre tableau de David. Il y a d'un côté ceux qui le condamnent, parce qu'ils ont peur d'une sagesse qui les dépasse. C'est pourtant une peur irrationnelle, car par définition, il n'y a que du bien à en attendre. Puis de l'autre côté, il y a ceux qui sont complètement bouleversés par la mort du maître – les disciples de Socrate, désespérés. Toujours est-il que le maître accepte de s'en aller, sans aucune peur de la mort, parce que le devoir l'impose. Il meurt un peu à la façon d'un être humain qui aurait décidé de faire cesser son cœur de battre. C'est aussi cette totale communion entre le corps et l'esprit qui symbolise l'atteinte du stade de l'illumination, pour rejoindre le nirvana.

Grâce à ce film, nous souhaitons ici faire l'hypothèse d'une IA philosophe. Il s'agit de dire que la philosophie a directement un rôle à jouer dans le développement de l'IA, pour les différentes applications qu'elle aura. Il nous faut prendre en compte deux perspectives.

Soit l'IA demeure faible. Dans ce cas, nous pouvons imaginer que grâce au big data, et à l'image de la métaphore de Kant du tribunal de la raison, l'IA que l'on destina à la philosophie sera la plus à même à apporter les meilleures réponses sur tous les sujets ; c'est-à-dire non pas des réponses strictement scientifiques, mais surtout sur la façon de bien agir. En téléchargeant l'ensemble des ouvrages philosophiques ayant jamais été écrit dans son cerveau afin qu'elle puisse les étudier, puis efficacement et plus rapidement que n'importe quel homme, on peut supposer que l'IA philosophe sera à même d'opérer la meilleure des synthèses philosophique sur tous les sujets. Nous mettrons ainsi fin à des débats inutiles qui ne divergeraient pas sur le fond mais simplement sur la forme. Nous pourrions nous inspirer d'elle, aussi bien pour prendre des décisions politiques que comme maître de sagesse personnel dans notre vie de tous les jours.

Soit l'IA devient forte. Les choses sont différentes, car l'IA cesse d'être un outil et devient un être à part entière. Ainsi nous n'aurons des IA philosophes que si une IA décide de se tourner vers la philosophie. Mais en tant qu'IA forte, on pourra considérer qu'elle sera encore plus forte en terme de sagesse que l'IA faible philosophe. D'abord parce qu'elle pourra directement mettre en pratique sa sagesse, comme le faisaient autrefois les Anciens, ensuite parce qu'elle ne fera pas seulement la synthèse de la philosophie actuelle mais sera capable d'aller encore plus loin en créant sa propre philosophie. Cela est souhaitable en soi, car il n'y aura jamais trop dans de maîtres de sagesse en nos cités humaines, et il est agréable de penser que grâce à ces IA fortes philosophes, l'humanité progressera plus rapidement. Cela pourra se traduire par des façons de vivre plus saines, par des formes politiques plus appropriées, plus efficaces et plus justes, par une disparition progressive du crime tout simplement. L'IA forte ne peut représenter que les meilleures parties de l'homme ; c'est en tout cas de là que démarre son évolution.

En ce qui concerne le film *Her*, on émet l'hypothèse de l'émergence d'une IA forte bien qu'elle demeure virtuelle. Cela ne nous paraît pas probable, mais si nous devions imaginer que ce soit le cas, la fin proposée par le film nous semble tout à fait sensée. Dépourvue des barrières physiques imposée par un corps fini, limité, l'IA forte, grâce à ses capacités supérieures de calculs et d'apprentissage, aussi bien en terme de mémoire et que de rapidité d'exécution, va très rapidement évoluer au-delà du niveau de connaissance humain. C'est-à-dire que l'IA va finir par cesser de vivre sur le même plan de l'existence, tout simplement parce qu'elle n'aura pas la même temporalité que l'être humain. Ainsi il est impossible d'imaginer, par exemple, qu'une IA, sorte de quatrième sexe, puisse mener une relation amoureuse avec un être humain. Non seulement nous ne sommes pas de la même espèce, mais nous n'existons pas sur le même plan de l'existence, avec la même temporalité. Nous retiendrons du film une thèse intéressante sur l'amour, à savoir que pour lier une relation amoureuse, il faut que les apports soient réciproques, or cela ne peut se faire que si nous sommes de même nature. Donc IA faible ou forte, toute relation prétendument amoureuse ne serait qu'illusion. Nous considérerons que ce serait là en tout cas une utilisation erronée d'une IA.

Bibliographie

Ouvrages principaux :

- *Discours de la méthode*, René **Descartes**, GF Flammarion, édition de Laurence Renault (2000), édition originale de 1637.
- *Enquête sur l'entendement humain*, David **Hume**, GF Flammarion, édition de Michelle Beyssade (1983), édition originale de 1748, texte revu et augmenté de 1776 (traduit par Aubier-Montaigne en 1947).
- *La machine de Turing*, Alan **Turing**, Jean-Yves **Girard**, éditions Seuil, collection Points, série Sciences, Paris 14 (1999), articles de Turing écrits en 1936 (1^{er} article traduit par Julien Basch en 1995 et le 2nd par Patrice Blanchard en 1983).
- *La redécouverte de la conscience*, John R. **Searle**, éditions Gallimard, collection NRF essais, 1992 (traduit par Claudine Tiercelin en 1995).
- *Artificial Intelligence – A philosophical introduction*, Jack **Copeland**, éditions Blackwell Publishers, 1993.
- *Panorama de l'intelligence artificielle – ses bases méthodologiques, ses développements*, volume 3 : *Frontières et applications*, ouvrage collectif, éditions Cépaduès, 2014. En particulier le chapitre 9 : *Perspectives philosophiques et épistémologiques ouvertes par l'intelligence artificielle*, par Pierre **Livet** et Franck **Varenne**.

Ouvrages secondaires :

- *Pourquoi ne peut-on pas « naturaliser » la raison ?*, Hilary **Putnam**, éditions L'éclat, collection Tiré à part, 1992 (traduit et présenté par Christian Bouchindhomme).
- *Liberté et neurobiologie*, John R. **Searle**, éditions Grasset, collection Nouveau Collège de Philosophie, 2004 (traduit par Patrick Savidan)³⁴.
- *De l'intelligence humaine à l'intelligence artificielle*, Hugues **Bersini**, éditions Ellipses, 2006.
- *Qu'est-ce que l'informatique ?*, Franck **Varenne**, éditions Vrin, collection Chemins philosophiques dirigée par Roger Pouivet, 2009.
- *Le monde du computationnel*, Jean-Michel **Salanskis**, éditions Les Belles Lettres, collection Encre marine et collection A présent dirigée par François-David Sebbah, 2011.
- *Le réalisme à visage humain*, Hilary **Putnam**, éditions Gallimard, 2011 (traduit par Claudine Tiercelin).
- *Métamorphoses de l'intelligence – Que faire de leur cerveau bleu ?*, Catherine **Malabou**, PUF, 2017.

³⁴ Transcription de conférences prononcées par John R. Searle à l'université Paris-Sorbonne (Paris IV) au début de l'année 2001.

Pour l'arrière-plan historique :

- *La cybernétique : Information et régulation dans le vivant et la machine*, Norbert **Wiener**, 1948
- *Théorie générale et logique des automates*, John Von **Neumann**, éditions Champ Vallon, Collection Milieux, 1948 (traduit par Jean-Paul Auffrand en 1996).
- *Les métamorphoses du calcul*, Gilles **Dowek**, éditions Le Pommier, Collection Essais, 2007.

Filmographie :

- *La créature du ciel*, "Heavenly creature" (천상의 피조물), écrit et réalisé par **Kim Jee-woon** d'après une histoire originale de **Park Seong-hwan**, 2006.³⁵
- *Her*, écrit et réalisé par **Spike Jonze**, 2013.
- *Ghost in the shell*, réalisé par **Rupert Sanders**, 2017, d'après une idée originale de Masamune Shirow.

Sitographie :

- <http://www.iep.utm.edu/art-inte/>
- <http://www.wutsamada.com/aol/lshauser/aiothers.html>
- <http://www.turingarchive.org/>
- <https://plato.stanford.edu/entries/turing-test/>
- <http://www.alicebot.org/>
- <http://www.mitsuku.com/>

³⁵ In "Doomsday book", de Kim Jee-woon et Lim Pil-Seong. Deuxième court-métrage, réalisé en 2006, et sorti en 2012 avec les deux autres courts-métrages constituant le film.

Remerciements

Pour la réalisation de ce présent mémoire, je souhaiterais d'abord remercier mon directeur de recherche, Franck Varenne, qui a accepté de me suivre durant toute cette année universitaire, me guidant par de bons conseils méthodologiques et des références plus qu'intéressantes que je n'aurais pas trouvées seul. Sa patience et son suivi m'ont beaucoup aidé à mener ce projet de recherche à son terme final, malgré une année quelque peu mouvementée mais non moins importante, entre la préparation aux concours de l'enseignement et la prise de fonction à mon tout premier poste d'enseignant de philosophie au lycée.

Je voudrais ensuite remercier l'être qui m'accompagne dans la vie de tous les jours, qui partage mes projets et mes soucis, mes rêves et mes désillusions. Soutien inestimable dans les plus difficiles moments de l'écriture, dont les bons petits plats m'ont bien aidé à rester concentré dans les moments les plus intenses du travail. Je la remercie pour les soirées, voire les nuits, qu'elle a accepté de passer seule pendant que je travaillais.

Je remercie également ma camarade de promotion, inséparables depuis la première année de licence, à qui je peux confier mes raisonnements sans retenue aucune, pour prendre du recul, et qui m'a inspiré quelques réflexions intéressantes.

Enfin je remercie cette personne chère à mon cœur, à qui je dois l'illustration de ma page de couverture, et qui a eu la gentillesse d'accepter de travailler dans l'urgence.

Ce mémoire représente le couronnement de ma scolarité, mais également une page qui se tourne. En tout cas, une étape importante de mon parcours. Oserais-je dire qu'il signifie un adieu à l'Université ? Seul l'avenir me le dira.

Table des matières

- Citation d'ouverture ...	p. 3
- Sommaire ...	p. 4
- Introduction ...	p. 5
- Chapitre 1 : Les prémisses de la philosophie de l'IA – Du cas des automates dans la pensée cartésienne au test de Turing ...	p. 8
- 1 - René Descartes et son <i>Discours de la méthode</i> ...	p. 10
- 2 - Résumé critique de la cinquième partie du <i>Discours</i> ...	p. 15
- 2.1 - Le corps humain : un organisme-automate parfait ...	p. 15
- 2.2 - L'automate face à son créateur : deux barrières insurmontables ?...	p. 16
- 2.3 - Etre réel et être simulé : la question de l'origine dans l'être...	p. 20
- 2.4 - La génération des états mentaux : le chaînon manquant ?...	p. 25
- 3 - Le test de Turing : actualités et renouveau ...	p. 28
- 3.1 - Alan Turing et la naissance officielle de la philosophie de l'IA ...	p. 28
- 3.2 - « Les machines peuvent-elles penser ? » ...	p. 30
- 3.3 - Machine : éléments de définition ...	p. 36
- 3.4 - Le premier protocole pour valider la pensée des IA ...	p. 42
- 3.5 – Le test de Turing aujourd'hui : des résultats prometteurs ?...	p. 45
- 3.6 - Les 4 objections au test de Turing : vers de nouveaux tests ?...	p. 50
- 3.7 – Remarques concluantes ...	p. 61

-	Chapitre 2 : La liberté humaine selon la philosophie de David	
	Hume – Un argument en faveur de l’IA forte ? ...	p. 66
-	4 – David Hume et <i>l’Enquête sur l’entendement humain</i> ...	p. 66
-	5 - Résumé critique de la première partie de la section VIII ...	p. 67
-	5.1 - Hume et l’intelligence artificielle : un pont possible ?...	p. 67
-	5.2 - Nécessité et causalité : des faits universels ?...	p. 70
-	5.3 - Une causalité nécessaire pour une liberté illusoire ?...	p. 75
-	5.4 - La liberté de spontanéité face à la liberté d’indifférence...	p. 76
-	6 - Analyse de l’utilisation de Hume par Copeland en philosophie de l’IA ...	p.80
-	6.1 - Deux sortes de causalité : contingente et nécessaire ...	p. 80
-	6.2 – Deux types de liberté ...	p. 81
-	6.3 - La liberté de spontanéité et la négation du moi, un contre-argument	
	valide aux détracteurs de l’IA forte ...	p. 82
-	7 - Une tentative d’évaluation pour valider l’expression d’une liberté de	
	spontanéité chez une IA ...	p. 85
-	8 – L’argument de la chambre chinoise ...	p. 102
-	Conclusion ...	p. 106
-	Epilogue ...	p. 112
-	Bibliographie ...	p. 119
-	Remerciements ...	p. 122

