UCLouvain

FIAL
Un éventail de formations, une multitude d'horizons

Faculté de philosophie, arts et lettres

Agence morale et intelligence artificielle

Un agent moral artificiel est-il possible ?

Auteur : Thomas Cardon
Promoteur : Pr. Charles Pence
Année académique 2021 - 2022
Master en philosophie à finalité approfondie

Table des Matières

Introduction

Propos introductif

S'il y a bien une réalité qui est aussi vieille que l'humanité elle-même, c'est l'anthropocentrisme. A l'échelle de l'histoire de l'humanité, ce n'est que récemment que certains penseurs remettent en question le primat de l'humain, son caractère fondamentalement différent. Il a fallu attendre la théorie évolutionniste de Charles Darwin pour comprendre à quel point l'humain est un animal, et pourtant même Darwin de son vivant n'a pas véritablement adopté cette position, malgré les preuves à sa disposition. Le contexte socio-culturel de cette époque, empreint d'un christianisme omniprésent, rendait presque impossible de remettre en question l'origine divine de la nature humaine. D'après la Bible, l'homme a été façonné par Dieu à son image, et l'espèce humaine est en ceci très distincte des autres animaux, dont elle est le gardien. C'est à Noé, un homme, qu'incombe la tâche de rassembler toutes les espèces animales avant le déluge. Cette conception de la distinction entre l'être humain et l'animal, et par extension, tout autre être que l'humain, nous a longtemps empêché de réaliser la complexité des animaux avec lesquels pourtant nous sommes en relation depuis des millénaires. Ce n'est que relativement récemment que l'on a pu comprendre et prouver que certains animaux les plus sophistiqués avaient une conscience d'eux-mêmes, via notamment le test du miroir[1]. Et pourtant, l'anthropocentrisme a continué de nous empêcher de comprendre l'évidence qui est parfois juste sous nos yeux. Si de nombreux mammifères passent avec succès le test du miroir, un animal pourtant connu pour son intelligence sociale notamment comme le chien ne passe pas aussi régulièrement un tel test. C'est une fois de plus l'anthropocentrisme qui nous empêche de réaliser que le test en lui-même est biaisé et a été constitué en fonction du point de vue particulier de l'humain sur la question de la conscience de soi. Précisément, c'est ici le biais selon lequel la vue est l'organe de reconnaissance primaire qui biaise l'expérience pour le chien. Chez le chien, le fait de reconnaitre autrui passe notamment par l'odeur, et une image d'un individu quelconque sans odeur ne ressemble pas, pour lui, à cet individu. Cette importance de l'odeur dans l'identification veut elle

[1] Ce test consiste à placer un animal devant un miroir après avoir au préalable disposé une tâche de couleur sur le visage de l'animal. Si en se voyant dans le miroir il comprend que la tâche est sur lui et tente de la toucher, c'est, a priori, qu'il a une conscience de lui-même et qu'il comprend que ce qu'il voit dans le miroir c'est son propre corps.

dire que le chien n'a pas conscience de lui-même ? Une telle question est évidemment absurde.

Un des éléments que l'être humain a longtemps tenu pour lui être propre, est évidemment l'agence morale. Plus que cela même, l'idée selon laquelle l'humain agit de manière délibérée, pose des choix et est véritablement libre alors que l'animal n'agit que par instinct est encore très répandue aujourd'hui. Pourtant, nous sommes bien capables de comprendre que nous-mêmes sommes sujet à certaines réactions instinctives, et pouvons observer que des animaux sont capables eux-mêmes de certains raisonnements complexes. Au plus l'on se penche sur la question, au plus on réalise que la distinction qualitative entre l'être humain et l'animal est plus floue qu'elle n'y parait. Cela étant dit, doit-on pour autant revenir aux procès d'animaux[2] du Moyen-Age ? Doit-on reprocher au lion de chasser la gazelle, à l'état naturel ? C'est évidemment absurde. Nous ne rentrons pas ici dans la différentiation entre agence morale et patience[3] morale, parce qu'elle dépasse la portée de notre questionnement. Néanmoins, cet exemple de l'animal nous permet de souligner la façon dont certaines notions, comme l'agence morale, sont traditionnellement réservées par l'être humain à l'être humain, comment elles ont été pensées sur base de notre modèle unique, et comment pour autant ce n'est sans doute pas la seule définition qui a du sens, pour l'ensemble des agents moraux potentiels.

Si nous tentons de déconstruire cette approche anthropocentrée, et que nous essayons de penser une place pour des agents moraux non-humains, nous serons peut-être à même d'en apprendre plus sur l'agence morale de l'humain lui-même. Dans notre monde actuel, les agents artificiels prennent de plus en plus de place. On a longtemps distingué l'humain du reste du vivant de par sa capacité à créer des outils, mais il semble bien que ces mêmes outils se soient complexifiés, automatisés au-delà de ce que l'on avait sans doute imaginé pendant une bonne partie de l'histoire. Ces outils sont si sophistiqués qu'on en vient à se demander dans quelle mesure ils ne sont pas capables de penser par eux-mêmes. Evidemment, la question de la liberté dont peut disposer une intelligence artificielle (IA) reste une question très prospective, mais il n'empêche que nous devons reconnaître que certaines capacités de l'IA sont fascinantes. Une des raisons pour

[2] Notons brièvement que ces mêmes procès sont à comprendre dans un contexte propre à l'époque qui est souvent oublié quand ils sont pris pour exemple.
[3] L'agent étant dans cette définition celui qui a des devoirs moraux envers un ou plusieurs patients moraux qui sont ceux qui ont des droits moraux.

lesquelles beaucoup d'entre nous ont l'intuition d'être des agents libres, c'est sans doute parce que nous ne sommes pas en mesure de pointer du doigt une force qui nous contraint à agir comme nous agissons. De la même manière, au plus l'IA se développe, au moins le grand public est capable de comprendre ce qui détermine le résultat de ses apparentes décisions. Et cela va même maintenant jusqu'à ce que les concepteurs de l'IA eux-mêmes ne puissent plus comprendre les délibérations qui ont menés à certaines de ses décisions. De cette inconnue, émerge l'idée que l'IA est peut-être capable de choisir, et donc d'agir en exprimant un certain degré de liberté. Ce n'est évidemment pas simplement parce qu'on ne comprend plus quelque chose que l'on doit lui accorder une spiritualité, mais c'est suffisant pour au moins attiser notre curiosité. Cette curiosité, elle nous amène à considérer l'agence morale artificielle, et cette réflexion se présente à nous, dans une certaine mesure, comme un miroir sur notre propre agence morale. En cherchant à mieux comprendre l'agence artificielle, on est naturellement amenés à questionner l'agence que l'on connait prétendument le mieux, la nôtre.

Par ailleurs, si les questions relatives à l'IA ont longtemps fait partie de la science-fiction, elles sont devenues de plus en plus réelles au fur et à mesure des dernières décennies. Certes les androïdes les plus poussés tirés de l'imagination des auteurs de roman ne parcourent pas encore nos rues, mais il n'empêche que l'IA s'est introduite discrètement dans nos vies jusqu'à devenir indispensable. Elle est dans la poche de nos jeans, sous la forme d'un smartphone, elle est au bout de mes doigts lorsque j'écris ce texte, elle aspire les maisons de certains, nettoie les piscines d'autres. Et si ces exemples paraissent peut-être triviaux et sans conséquences, on ne peut certainement pas en dire autant des voitures autonomes qui sont de plus en plus fréquemment déployées, et encore moins des drones militaires qui font déjà partie du quotidien de certains soldats. Dès lors, une réflexion sur les conditions de possibilité de l'agence morale artificielle, sur la façon dont on peut interagir avec de tels agents moraux et sur les conséquences de leur existence n'est plus un luxe, elle est devenue une urgente nécessité. Il est du devoir des penseurs et des philosophes de réfléchir le cadre moral et le cadre légal dans lequel ces nouveaux agents vont évoluer avant que l'absence de cadre n'entraîne des conséquences malheureuses que l'on pourrait éviter. Cela étant dit, il ne semble pas que les mises en garde ou les prophéties, selon qu'elles soient pessimistes ou optimistes, de ceux qui

prêchent l'arrivée de la singularité[4] doivent nous éloigner de réflexions plus réalistes. Au contraire, l'objectif de ce mémoire sera de tenter d'éviter autant que faire se peut de partir dans des spéculations. Il s'agira de tenter de définir un cadre aussi objectif que possible, de défricher autant que possible les différents concepts qui entourent la question de l'agence morale afin de pouvoir montrer comment agence morale et IA peuvent non seulement coexister, mais aussi nous en apprendre plus sur nous-même.

Plan du mémoire

Ce mémoire est construit en trois parties. Sa structure logique correspond dans une certaine mesure à la méthode hégélienne de l'« aufhebung[5] », en ce que nous allons prendre deux concepts relativement distants, l'agence morale et l'IA afin de les réunir dans une troisième partie qui aura bon espoir de les sublimer. En effet, la première partie concernera tout d'abord l'agence morale et certaines considérations éthiques, la seconde partie portera directement sur l'Intelligence Artificielle (IA), son histoire, son fonctionnement et les différents paradigmes qui la définissent, et la troisième partie sera enfin la réunion des deux précédentes, et aura pour but de montrer comment les différents concepts interagissent ensemble.

Dans la première partie de ce travail, nous introduirons d'abord le concept d'agence morale en discutant notamment la circularité entre agence morale, action morale et responsabilité morale. Nous présenterons ensuite diverses approches de la responsabilité morale, en mobilisant pour se faire l'article de Matthew Talbert publié dans la Stanford Encyclopedia of Philosophy dans l'édition d'hiver 2019 et intitulé « Moral Responsibility »[6]. Nous aborderons ensuite la distinction entre compatibilisme et incompatibilisme en ce qui concerne les concepts de liberté et de déterminisme. Nous ferons ensuite un aparté pour discuter le problème de l'intentionnalité avant de présenter

[4] La singularité est le terme qui désigne le moment à partir duquel l'intelligence humaine serait dépassée par l'intelligence artificielle, si bien que cela entraînerait des conséquences fondamentales sur nos conditions de vie. Certains y voient une apocalypse, d'autres y voient l'avènement d'une nouvelle ère d'abondance. Nous n'allons pas discuter ces points de vue extensivement, parce qu'ils rencontrent un certain nombre de problèmes logiques, et que la pure prospection dépasse le cadre de ce mémoire.

[5] Parfois traduit par le terme « sublimation », ce terme désigne, chez Hegel, le dépassement d'une contradiction dialectique qui permet de dépasser ses prémisses en les élevant. Agence morale et IA ne représentent pas dans l'absolu une contradiction dialectique, même si ce sont des concepts relativement paradoxaux, et notre ambition est moindre que celle d'une véritable sublimation, mais il y a néanmoins un parallèle dans la méthode appliquée.

[6] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019).

notamment l'approche dite « forward-looking », l'approche par « quality of will » ainsi que le problème de la chance morale. A ceci nous ajouterons une proposition de réflexion autour d'une approche différente du concept d'agence morale et de responsabilité à partir du concept de prise de risque, pour tenter de nous débarrasser des inconvénients du conséquentialisme et de l'intentionnalité pour nous concentrer sur ce qui est effectivement disponible à notre analyse, l'action elle-même. Pour conclure cette première partie, nous introduirons quatre théories éthiques très distinctes, à savoir l'utilitarisme, le déontologisme, l'éthique de la Vertu et l'éthique du Care. Ces quatre théories seront présentées très brièvement, dans le but d'être mobilisées dans la troisième partie de ce travail.

Dans la seconde partie de ce travail, nous nous concentrerons donc sur l'IA. Nous l'introduirons par une traditionnelle reprise des grandes lignes de son histoire, en ne manquant pas de soulever l'intéressante distinction entre le terme français « Intelligence artificielle » (IA) et le terme anglais « Artificial Intelligence » (AI), qui, s'ils partagent un sens commun, mobilisent des termes qui ont des sens légèrement différents dans leur langues respectives. Nous parlerons ensuite du fonctionnement de l'IA, en mobilisant pour se faire l'article de Selmer Bringsjord et Naveen Sundar Govindarajulu publié dans la Stanford Encyclopedia of Philosophy dans l'édition d'automne 2018 intitulé « Artificial Intelligence »[7]. Nous évoquerons alors les différents types d'IA, les différentes formes d'apprentissage qui peuvent être mobilisées, en présentant notamment les grandes lignes du fonctionnement des réseaux de neurones artificiels. Après ceci, nous exposerons les distinctions entre les IA de type descendante (top-down) et les IA ascendante de type (bottom-up). Cette distinction sera importante pour la suite de notre réflexion. Nous conclurons cette partie par un questionnement plus philosophique, à savoir la distinction entre une IA au sens fort (Strong-AI) et une IA au sens faible (Weak-AI).

Dans la dernière partie de ce travail, nous allons tenter de faire communiquer ces deux mondes distincts que sont l'agence morale et l'intelligence artificielle. Pour se faire, nous allons commencer par discuter les conditions de possibilité d'un agent moral artificiel, en partant de l'article de John P. Sullins publié en 2006 dans la International Review of Information Ethics et intitulé « When a robot is a moral agent ? ». Nous

[7] BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in Stanford Encyclopedia of Philosphy, (2018).

évoquerons à ce titre les trois conditions qu'il pose, à savoir autonomie, intentionnalité et responsabilité, ainsi que l'approche minimaliste qui est la sienne. Dans un second temps, nous mobiliserons les quatre courants éthiques discutés dans la première partie et essayerons de montrer leurs intérêts et leur problématique dans le cadre de l'agence morale de l'IA. Ensuite, nous évoquerons différentes difficultés supplémentaires que rencontre l'agence morale artificielle, au travers notamment de liens avec les approches de l'agence morale que nous avons discutées précédemment. Enfin, nous conclurons cette partie par une discussion plus prospective sur différentes problématiques de l'IA en société, en apportant des pistes de solutions, et en soulignant la présence actuelle de l'IA dans notre société mais aussi ses limites intrinsèques.

Première partie

Introduction à l'agence morale

L'agence morale désigne, très simplement, la capacité à agir moralement. Cette capacité, la tradition considère qu'elle est réservée aux humains dits « normaux ». Cette tradition considère donc que les animaux non-humains ne la possèdent pas, et qu'elle manque aussi aux très jeunes enfants, mais encore à ceux qui souffrent de handicap mental sévère ou de démence (pour en citer quelques-uns).

> Whatever the correct account of the powers and capacities at issue (…), their possession qualifies an agent as morally responsible in a general sense: that is, as one who may be morally responsible for particular exercises of agency. Normal adult human beings may possess the powers and capacities in question, and non-human animals, very young children, and those suffering from severe developmental disabilities or dementia (…) are generally taken to lack them.[8]

L'expression que nous avons utilisé « agir moralement » est en soit relativement peu claire elle-même. Prenons d'abord la première partie, « agir ». « Agir » consiste à poser un acte, ce qui est à distinguer du simple geste ou du simple mouvement. Là où les branches de l'arbre s'agitent au passage du vent, là où le caillou dégringole sur le flanc de la montagne, il n'y a pas d'acte, il n'y a que de simples mouvements. Pour qu'il y ait acte, il faudrait qu'un « agent » (traditionnellement un humain) jette le caillou, secoue la branche. Certaines traditions philosophiques distinguent encore selon si l'agent avait l'intention de poser le geste ou s'il l'a fait accidentellement, mais nous évoquerons plus tard les raisons qui nous font éviter le sujet de l'intention[9]. L'acte est donc posé par un agent, qui de par son action agit de telle sorte à ce qu'il se rende responsable de l'action qu'il a posé. Il se trouve donc contraint d'accepter la responsabilité morale des conséquences de son action. Celle-ci peut dès lors se trouver être félicitée (praise), blâmée (blame), ou bien encore être reçue avec indifférence. La question se pose dès lors de déterminer ce qui est primordial entre toutes ces notions. Il semble y avoir une circularité entre les concepts d'action et d'agent, l'un requérant inévitablement l'autre. L'action, c'est donc ce mouvement (ou

[8] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p1.
[9] Voir page 20.

cette absence[10] de mouvement) posé par un agent. Ces éléments sont fondamentaux, et ce n'est que lorsqu'une forme de relation interpersonnelle[11] intervient que la dimension de moralité et donc de responsabilité peut surgir. Par ailleurs, la responsabilité est secondaire à l'action parce qu'elle ne peut être considérée qu'à la lumière des conséquences de l'action. Nous verrons ultérieurement que cette approche conséquentialiste n'est pas la seule approche possible, mais il n'empêche que la dimension de responsabilité est postérieure, logiquement, à l'action elle-même. Néanmoins, la dimension morale de notre rapport interpersonnel à autrui ne peut se penser qu'avec l'idée même de responsabilité en tête. L'un requiert l'autre. Enfin, la dernière dimension que nous avons évoquée, à savoir celle de la félicitation ou du blâme, n'arrive que dans un dernier temps, parce qu'elle représente la façon dont une action non seulement est perçue par autrui, mais en plus elle constitue la manière dont autrui réagit à l'action initiale. Logiquement, la réaction est secondaire à l'action. Il y a donc plusieurs circularités à l'œuvre entre ces différents concepts, l'action morale n'est possible qu'en la présence d'un agent moral, et l'existence même de ces concepts n'est possible qu'avec le prérequis d'un concept de responsabilité, qui n'a lui-même de sens qu'à la lumière d'une action. Finalement, il convient de se demander si une action peut être différenciée d'un simple geste sans que le cadre moral ne soit directement mobilisé. Si ce n'est pas le cas, l'action morale, l'agent moral et la responsabilité morale sont chacun des prérequis mutuels et doivent donc être posés de façon axiomatique. Nous prenons en particulier le spectre de l'agence morale[12], parce qu'il s'agit de la qualité de l'IA que nous étudions, mais nous mobiliserons aussi régulièrement

[10] On peut argumenter que le fait même de ne pas faire quelque chose peut constituer un acte en lui-même. La différence entre un choix actif ou un choix passif ouvre une discussion relativement complexe que nous n'allons pas aborder ici. L'exemple classique, bien que limité, est celui du dilemme du tramway qui montre que ce n'est pas la même chose que de laisser mourir quelqu'un parce que le sauver aurait couté la vie d'un certain nombre d'autres personnes ou bien d'activement choisir de sacrifier une personne sous le prétexte d'en sauver d'autres.

[11] Il est possible d'envisager que cette relation interpersonnelle se vivre avec soi-même. La question se pose, est-ce une exigence morale que de se respecter soi-même ? Que ce soit ou non le cas ne change pas les conclusions de notre propos.

[12] Notons tout de même qu'un primat de l'agence pourrait être argumenté sur base du fait d'une part que l'agence est une capacité nécessaire à l'action et qu'elle précède donc l'action, et d'autre part que la responsabilité ne peut qu'être secondaire logiquement à l'action. Si elle est chronologiquement simultanée à l'action, il n'empêche que l'on peut argumenter que la responsabilité n'est possible que logiquement après que l'action ai été effectuée. Dès lors, l'agence morale, qui ne se manifeste que grâce à l'action et à la responsabilité dans la relation interpersonnelle pourrait bien être logiquement première.

la responsabilité morale, parce qu'un certain nombre de concepts sont plus propices à cet axe donné.

Ces manières de caractériser l'action d'autrui, d'apparence symétrique, selon qu'elle soit jugée positive ou négative sur le plan moral, se sont insérées dans les pratiques sociales et dans les traditions d'une façon qui montre que les blâmes sont bien plus prépondérants que les félicitations. C'est en tout cas ce que Matthew Talbert défend dans l'extrait suivant.

> Blame is a response that may follow on the judgment that a person is morally responsible for behavior that is wrong or bad, and praise is a response that may follow on the judgment that a person is morally responsible for behavior that is right or good. It should be noted at the outset that the above schema, while useful, may be misleading in certain respects. For one thing, it suggests a correspondence and symmetry between praise and blame that may not exist. The two are certainly asymmetrical insofar as the attention given to blame far exceeds that given to praise. One reason for this is that blameworthiness, unlike praiseworthiness, is often taken to involve liability to a sanction. Thus, articulating the conditions of blameworthiness may seem to theorists the more pressing matter. [13]

Il montre en effet que l'attention donnée aux blâmes est d'une nature plus significative que celle accordée aux félicitations. Personne n'est félicité pour un comportement décent, qui est en fait simplement attendu de chacun sans contrepartie, alors qu'un comportement problématique, même léger, sera la cible de reproches. Dans le cas d'évènement plus significatifs, il est légitime de s'attendre à des remerciements si l'on venait à sauver la vie de quelqu'un, peut-être même à une légère récompense, mais c'est sans commune mesure à la sanction qui est appliquée à celui qui commet un meurtre. Deux actions d'apparence symétriques, sauver une vie et prendre une vie, sont reçues de manières très asymétriques. Et c'est encore plus vrai si celui qui a sauvé une vie l'a fait sans effort et sans risques, auquel cas, ce serait tout simplement attendu de la part de n'importe quel agent. En fait, l'asymétrie entre ces deux concepts semble venir du fait que le point de départ ne se situe pas au centre, dans un terrain neutre. Le point de départ est l'action considérée comme décente, celle qui est attendue de tout agent moral, sans pour autant représenter du zèle. Cette action type standard est celle qui permet à l'agent d'éviter d'être blâmé, sans pour

[13] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p2.

autant être félicité pour l'acte qu'il a posé. Si le contexte du blâme est plus présent que le contexte de la félicitation, c'est assez simplement parce que l'action standard est moins souvent atteinte qu'elle n'est dépassée. Il arrive plus régulièrement qu'un agent ne parvienne pas à élever son action au statut moral de l'indifférence, et mérite donc d'être blâmé, qu'il n'arrive qu'un agent ne dépasse ce statut par son zèle, et mérite donc des félicitations. Que ceci soit dû à la nature humaine ou à une exigence trop élevée, nous ne l'aborderons pas ici. Néanmoins, il était important de comprendre pourquoi le contexte dans lequel l'agence morale est étudiée est principalement tourné vers le blâme.

Par ailleurs, nous l'avons évoqué, l'action morale est l'action qui entraine une responsabilité morale, et elle est à distinguer du simple acte, ou du geste, dont un individu peut se rendre responsable causalement. Tout être vivant n'est cependant pas capable d'agir moralement. Il n'est pas question de responsabilité morale, nous l'avons vu, dans les gestes, aussi intentionnels peuvent-ils sembler, d'un nourrisson. De la même manière, si j'ouvre, sans être brusque, une porte opaque et se faisant cogne quelqu'un d'autre qui se situait derrière la porte, bien que je sois responsable causalement du choc entre la porte et l'autre personne, je ne peux être tenu responsable moralement de la souffrance de cette personne si rien dans mon comportement n'est reprochable, et si je ne pouvais m'attendre à ce qu'elle se trouve juste derrière cette porte. Notons qu'un élément clé dans cet exemple est l'information dont je dispose, ainsi que l'information dont on attend de moi que je dispose. Si la porte était vitrée et que je voyais la personne de l'autre côté, un tel accident aurait été facilement évité, et si j'ouvrais la porte malgré tout et que le choc avait lieu, j'en serais évidemment responsable moralement, disposant de l'information nécessaire pour l'éviter. De la même manière, si la porte était vitrée et que, pour quelque raison que ce soit, je décidais de fermer les yeux et de ne pas regarder à travers la vitre avant d'ouvrir la porte, bien que je ne dispose pas de plus d'information au moment de pousser la poignée que dans l'exemple initial, je pourrais fort probablement être blâmé parce que j'aurais choisi moi-même de ne pas acquérir cette information et d'être dans l'ignorance de la présence de quelqu'un derrière la porte alors que le simple fait d'ouvrir les yeux m'aurait donné l'information nécessaire à éviter l'accident. Nous reviendrons plus en profondeur sur cet élément en particulier, mais l'exemple permettait ici de souligner la distinction

entre le manque d'information contingent et celui qui est délibéré, volontaire, ou tout du moins dont l'agent s'est rendu responsable.[14]

Distinction entre compatibilisme et incompatibilisme

Avant d'aller plus loin dans la construction de l'approche que nous allons défendre dans ce mémoire, il est nécessaire d'aborder une distinction majeure en ce qui concerne l'agence morale. Ce point particulier fonde le désaccord entre ceux qui défendent le compatibilisme et ceux qui favorisent l'incompatibilisme. Cette distinction concerne la façon de considérer des éléments qui semblent au premier abord paradoxaux et pourtant nécessaires au concept même de responsabilité, à savoir la liberté et le déterminisme. Si la pertinence du contexte déterministe dans le cadre du questionnement qui est le nôtre, autour de l'IA, saute aux yeux, la pertinence d'une réflexion en ce qui concerne la place de la liberté et son importance est moins évidente. La liberté n'est déjà pas une donnée évidente chez les êtres humains, et certains lui présentent des contre-arguments. Elle est donc au moins aussi peu évidente dans le cadre de l'IA. Néanmoins, une certaine notion de liberté est nécessaire à l'action. En l'absence totale de quelque notion de liberté que ce soit, il parait impossible de maintenir le concept d'action morale.

> How is the responsible agent related to her actions; what power does she exercise over them? One (partial) answer is that the relevant power is a form of control, and, in particular, a form of control such that the agent could have done otherwise than to perform the action in question. This captures one commonsense notion of free will, and one of the central issues in debates about free will has been about whether possession of it (…) is compatible with causal determinism.[15]

Il s'agit donc de déterminer quelles approches concernant la liberté sont compatibles avec notre réflexion. Au premier abord, on peut être tenté de penser que le déterminisme rend toute forme de liberté impossible et que déterminisme et liberté sont des concepts incompatibles, c'est ce que défendent les « incompatibilistes ». Mais le rapport entre ces deux éléments est plus complexe qu'il n'y parait. Tout d'abord, tous deux sont à première vue nécessaires à l'existence même du concept de responsabilité, et c'est pourquoi certains incompatibilistes adoptent une position sceptique quant à

[14] La question de la responsabilité de la qualité de l'information dont l'agent dispose est discutée plus amplement dans la suite de ce mémoire, dans le chapitre intitulé « les conditions épistémiques de la responsabilité » à la page 29.
[15] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019)., p5.

l'existence de la responsabilité morale. Pour simplifier, ils sont tous deux nécessaires à l'existence du concept de responsabilité morale parce que sans liberté il ne peut y avoir de choix et sans choix il ne peut y avoir de responsabilité, étant donné qu'une seule possibilité ne laisse pas de place à l'alternative et contraint l'agent à agir d'une telle façon, indépendamment de sa volonté.[16]

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (1983: 16)[17]

A l'inverse, un certain déterminisme est nécessaire afin de rendre le monde sur lequel portent les actions d'un agent suffisamment prévisible pour qu'il puisse être rendu responsable des conséquences de ses actes. Si rien ne permet à l'agent de prévoir les conséquences de ses actes, il est trivial de l'en tenir responsable, nous l'avons vu précédemment avec l'exemple de la porte opaque. Si le monde n'est pas prévisible, si par exemple une centrale explose lorsque j'appuie sur l'interrupteur de ma chambre, je ne peux être tenu responsable de l'explosion parce que rien ne me permettait de prévoir que mon acte, appuyer sur l'interrupteur, allait provoquer l'explosion. Pour que la responsabilité ait du sens, il est nécessaire que le monde dans lequel l'agent agit soit suffisamment prévisible. La responsabilité causale n'est pas juste différente de la responsabilité morale, elle est impliquée par celle-ci, au moins à un certain degré. Nous voyons nettement comment la nécessité de deux conditions qui semblent mutuellement exclusives peut pousser à envisager un certain scepticisme en ce qui concerne l'agence morale. Par défaut, c'est une position que nous allons nous efforcer d'éviter, étant donné qu'elle est de toute évidence peu fertile. Néanmoins, il existe plusieurs approches, aussi bien parmi les « incompatibilistes » que parmi les « compatibilistes » (ceux qui estiment que la liberté et le déterminisme ne sont pas mutuellement exclusifs), qui permettent de donner du crédit à la responsabilité morale. Nous allons discuter certaines de ces approches.

---

[16] Cette définition se veut caricaturale. Nous allons voir par la suite comment certains défendent qu'un choix peut être posé en l'absence d'alternative.

[17] P. van Inwagen, An Essay on Free Will, 1983, cité dans TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p6.

La première approche compatibiliste est celle qui rejette tout simplement l'idée selon laquelle la liberté d'agir comme on le souhaite requiert une alternative, requiert la possibilité d'agir autrement. Ce qui est incompatible avec le déterminisme, c'est la possibilité d'agir autrement, mais les compatibilistes affirment qu'il est possible d'être libre et de choisir d'agir dans un sens quand bien même il n'est pas possible d'agir autrement. D'après Moritz Schlick :

> Freedom means the opposite of compulsion ; a man is *free* if he does not act under *compulsion*, and he is compelled or unfree when he is hindered from without…when he is locked up, or chained, or when someone forces him at the point of a gun to do what otherwise he would not do (1930 {1966:59})[18]

Il y a deux arguments pour défendre cette position. Le premier argument, le plus ancien, remonte aux stoïciens, dont certains défendaient que le déterminisme ne signifie pas que nos actions sont déterminées à partir d'éléments qui nous sont extérieurs. Or, si la contrainte qui nous force à opérer telle ou telle action n'est pas une contrainte extérieure mais trouve sa source en nous-même, la volonté libre est compatible avec le déterminisme. C'est d'ailleurs une idée que l'on retrouve aussi chez Aristote, à savoir que la volonté libre permet à un être d'agir en fonction de qui il est fondamentalement, d'exprimer son essence, et non d'agir en fonction de contraintes extérieures. L'idée de contrainte intérieure n'est ici pas prise en compte, dès lors qu'un être agit en fonction de sa nature propre, il agit librement. Cette idée est compatible avec le déterminisme ainsi qu'avec une notion de liberté qui ne requiert pas de choix alternatif possible. Plus récemment, une autre approche tente de contourner le problème d'un choix alternatif en proposant qu'un choix alternatif serait possible à condition que certains éléments du passé aient été différents. Ils proposent une analyse conditionnelle de la capacité à agir autrement et disant que si un sujet a choisi l'action A, il aurait pu choisir l'action B si le passé avait été différent, si par exemple il avait, au passé, eu un désir différent qui l'aurait amené à choisir l'action B.

Ces deux arguments rencontrent un certain nombre d'objections de la part du camp incompatibiliste qui ne les trouvent pas satisfaisants. Le principal défaut du premier

[18] M. Schlick, cité dans TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p7.

argument concerne, comme nous l'avons évoqué, l'absence de considération pour certaines contraintes intérieures, dans des cas d'endoctrinement, de « lavage de cerveau », etc. Au moment du choix dans ces cas, la contrainte, bien qu'elle trouve son origine ultime à l'extérieur de l'agent, est bien une contrainte intérieure, et semble ne poser aucun souci aux compatibilistes qui risqueraient dès lors de juger un agent ayant subi ce genre de manipulation comme capable d'un jugement moral, alors qu'il est évident que quelqu'un qui a subi un lavage de cerveau peut ne pas être en mesure de prendre certaines décisions. Cet argument est particulièrement fort dans le cadre de la réflexion sur l'IA, étant donné que par essence, tout agent artificiel est un agent qui a subi une forme de « lavage de cerveau »[19]. Le second argument est moins problématique mais est pour autant jugé comme très insatisfaisant par les incompatibilistes qui estiment qu'une liberté d'agir autrement à la condition que le passé ait été différent n'est pas suffisant pour être qualifié de liberté. La possibilité d'un choix alternatif, pour les incompatibilistes, doit être possible ici et maintenant, avec un passé identique.

> More generally, incompatibilists are likely to be dissatisfied with the conditional analysis since it fails to give an account of an ability that agents can have, right here and right now, to either perform or omit an action while holding everything about the here and now, and about the past, fixed.[20]

Enfin, une troisième approche compatibiliste consiste, plutôt que de tenter de sauver la possibilité d'un choix alternatif dans un monde déterministe, en l'affirmation que la possibilité d'un tel choix alternatif n'est tout simplement pas nécessaire pour fonder la responsabilité morale, et, sur cette base, l'agence morale. Harry Frankfurt publie à ce sujet un texte en 1969 qui aura une large influence, en proposant différents exemples qui cherchent à prouver qu'un agent peut être tenu responsable de ses actes quand bien même il n'aurait pas pu agir autrement. Ces exemples sont appelés « exemples de Frankfurt » (*Frankfurt cases or Frankfurt examples*). Matthew Talbert décrit l'exemple de Frankfurt dans sa forme basique comme suit :

> An agent, Jones, considers a certain action. Another agent, Black, would like to see Jones performs this action and, if necessary, Black can *make* Jones perform it through some type of intervention in Jones's deliberative process. However, as things transpire, Black does not intervene in Jones's

[19] Les conséquences d'une telle situation seront explorées plus amplement dans la troisième partie de ce document.
[20] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p8.

> decision making since he can see that Jones will perform the action on his own and for his own reasons. Black does not intervene to ensure Jones's action, but he could have, and he would have, had Jones showed some sign that he would not perform the action on his own. Therefore, *Jones could not have done otherwise*, yet he seems responsible for his behavior. After all, given Black's non-intervention, Jones's action is a perfectly ordinary bit of voluntary behavior.[21]

Bien que cet exemple tente de démontrer que l'agent peut être responsable de son action même s'il n'aurait pu agir autrement, le contre argument qui saute aux yeux est évidemment de dire que Jones avait en fait le choix entre deux options. Soit il pouvait effectuer l'action de son plein gré, soit il pouvait effectuer la même action des suites de l'intervention de Black. Il parait absurde de ne pas reconnaitre la nuance entre ces deux évènements. Si un adulte assiste un jeune enfant dans ses exercices de calcul mental, et que l'enfant donne la réponse qu'il pense être la bonne à haute voix avant de l'écrire, et que l'adulte le corrige, ou ne le corrige pas, selon si la réponse est exacte avant que l'enfant n'écrive la réponse, on peut savoir à l'avance que l'enfant écrira la bonne réponse sur sa feuille, ce qui reste à déterminer, c'est si l'enfant la trouvera par lui-même. Que l'enfant trouve la réponse par lui-même ou pas est l'élément le plus important dans cet exemple, s'il trouve la réponse, il sera félicité, s'il se trompe, l'adulte le corrigera, dans les deux cas, il écrira ensuite la bonne réponse sur sa feuille. L'enfant ne sera fier d'avoir écrit la bonne réponse sur sa feuille que s'il l'a trouvée lui-même, s'il a été corrigé, il ne se sentira pas responsable d'avoir « trouvé » la bonne réponse. Néanmoins, ce que Frankfurt essaie de montrer avec son exemple, ce n'est pas ce qu'il se serait passé si Jones n'avait pas choisi d'agir comme Black le souhaitait, mais c'est bien que, dans le cas ou Jones agit conformément à la volonté de Black, pour des raisons qui lui sont propres, il n'est pas pertinent de se demander s'il aurait pu agir autrement pour affirmer qu'il est responsable de ses actes. A partir de cet exemple, ce qui prévaut n'est plus la possibilité d'agir autrement mais bien le fait que la personne ait agi pour ses propres raisons. On retrouve à ce sujet les arguments compatibilistes que nous avons vu précédemment, mais débarrassés de la nécessité de prouver que l'agent aurait pu agir autrement. Notons tout de même que cette approche requiert, une fois de plus, d'en savoir plus sur l'origine de la volonté d'agir, sur l'origine de l'intention au sein même de l'esprit de l'agent, et comme nous l'avons

[21] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p9.

déjà évoqué, nous verrons ultérieurement comment une telle approche peut s'avérer problématique et limitante.

Le problème de l'intention

Une autre notion que l'on ne peut esquiver lorsque l'on considère l'agence morale est la notion d'intention. Comme nous l'avons évoqué plusieurs fois déjà, la notion d'intention est problématique. Elle est particulièrement problématique dans le cadre d'un agent artificiel, et ce pour au moins deux raisons que nous allons évoquer ci-après. Notons tout d'abord qu'elle est hautement pertinente et que s'en séparer ne se fait pas à moindre coût. Ce n'est pas sans raison qu'elle est traditionnellement très présente dans ces discussions. Elle permet de distinguer sans plus d'efforts les actes posés de manière accidentelle de ceux posés avec l'intention de provoquer les résultats effectivement obtenus. Elle opère en fait la liaison entre la volonté libre et le monde de la causalité. L'intention va manifester la volonté au travers de l'acte posé. L'intention est ce qui permet de relier ces deux réalités intuitivement nécessaires à l'action et pourtant très distinctes. Mais pourquoi souhaiterions-nous dès lors écarter un tel outil ? Le problème de l'intention, de façon similaire à beaucoup d'autres concepts du même ordre, est qu'elle se retrouve être moins facilement ciblable que nous le souhaitons. Dans le cadre de ce travail, nous tentons de définir une approche aussi rationaliste que possible, aussi objective que possible, et en nous affranchissant autant que faire se peut des concepts qui laissent place à de l'interprétation, à de la subjectivité ou à du flou sémantique. De toute évidence, un tel projet ne peut se faire sans concessions, mais nous allons au moins tenter de montrer jusqu'où il peut se maintenir. Nous allons donc faire le pari que le concept d'intention n'est pas indispensable à une théorie de l'agence morale robuste. Avec l'intention, nous allons mettre de côté de manière systématique tout ce qui échappe à la certitude. En effet, si l'intention est véritablement problématique, c'est précisément parce que l'intention n'est en vérité jamais connue que, tout au plus, de l'agent lui-même. Qu'il décide ou pas de révéler ses véritables intentions à celui qui les lui demande représente un nouveau choix qui n'appartient qu'à lui, et c'est une incertitude qui rend l'intention plus problématique qu'elle n'est véritablement indispensable. Par ailleurs, il est bon de préciser que la complexité de l'esprit humain semble montrer que même l'agent honnête n'est peut-être pas toujours au clair quant à ses propres intentions. Avons-nous vraiment prononcé ce lapsus par accident ou était-ce « intentionnel » pour une partie inconsciente de notre

volonté ? Il est bien connu que le témoignage n'est pas toujours une source d'information fiable, même si l'honnêteté du témoin est établie, en particulier en ce qui concerne des évènements particulièrement forts émotionnellement, comme les questions liées à la responsabilité morale peuvent l'être.

Venons-en brièvement aux deux raisons qui nous font mettre de côté la question de l'intention dans le cadre d'un agent artificiel. Tout d'abord, les IA les plus complexes présentent des structures telles qu'il est parfois particulièrement difficile, voire simplement impossible, pour l'humain de décrypter les éléments précis qui ont mené à une telle décision. D'une manière similaire à la délibération chez l'humain, lorsqu'une décision est prise, on pense peut-être que c'est pour telle et telle raison, mais c'est en fait une myriade d'éléments qui jouent plus ou moins en faveur, qui compose notre système de valeurs à partir duquel une situation complexe est décortiquée pour, face à un dilemme, laisser place à un choix. Toute cette complexité, tous ces détails ne sont pas forcément plus accessibles chez la machine que chez l'humain, c'est pourquoi un tel concept que celui d'intention, bien que pertinent, soulève plus de nouvelles questions qu'il n'en résout. Il nous faudra donc trouver comment pallier ce manque, ainsi que les autres manques qu'une approche aussi exigeante soulèvera inévitablement. Un second élément qui nous empêche de mobiliser l'intention dans le cadre d'un agent artificiel est qu'un tel usage nécessiterait un moyen de pointer du doigt la présence ou l'absence d'intention dans les actions d'un agent artificiel, et un tel défi ne semble pas être résolu à l'instant. S'il est toujours possible de demander à un humain quelles étaient ses intentions, une telle question peut être problématique dans le cadre de l'action d'un agent artificiel. Mobiliser une telle notion, en plus d'être difficile à justifier, semble tout simplement impossible en l'état.

Différentes approches

La première approche à laquelle nous allons nous intéresser est l'approche dite « Forward Looking ». Cette approche consiste à considérer la responsabilité morale par le spectre des conséquences positives au fait même de l'attribuer.

> Forward-looking perspectives tend to emphasize one of the central points discussed in the previous section: an agent's being subject to determinism does not entail that he is subject to constraints that force him to act independently of his choices. If this is true, then, regardless of the truth of

> determinism, it may be useful to offer certain incentives to agents—to praise and blame them and generally to treat them as responsible—in order to encourage them to make certain choices and thus to secure positive behavioral outcomes.[22]

Quelle que soit la posture que l'on adopte en ce qui concerne le déterminisme, dès lors que l'on reconnait qu'un agent agit en fonction de ses choix, il peut être intéressant de l'inciter à agir d'une façon qui est jugée conforme à la morale. En lui donnant des incitations à agir positivement et, de manière générale, en assignant blâme ou félicitation en fonction des actions de l'agent, en le considérant comme un agent moral, on peut raisonnablement s'attendre à obtenir un comportement positif. Il ne s'agit pas de regarder en arrière vers ces comportements, dans une approche qui serait alors de facto « backward looking », mais bien de leur assigner une valeur morale, ainsi qu'un blâme ou une félicitation de manière à inciter l'agent à répéter les comportements qui ont été la source de félicitations et d'éviter les comportements qui ont provoqué un blâme. Dans une certaine mesure, c'est cette approche qui donne naissance à la distinction entre une justice dite « rétributive » (visant à punir) et une justice « transformatrice » (qui vise à modifier les comportements) (art. consid. Morale : à mobiliser plus tard). D'après Mortiz Schlick, l'un des principaux tenants de cette approche au début du siècle passé, la question de la responsabilité peut être résumée comme suit :

> The question of who is responsible is the question concerning the *correct point of application of the motive*…. in this its meaning is completely exhausted; behind it lurks no mysterious connection between transgression and requital…. It is a matter only of knowing who is to be punished or rewarded, in order that punishment and reward function as such – be able to achieve their goal (1930 {1966: 61};emphasis in original)[23]

Schlick insiste sur le fait que l'idée que la punition soit une vengeance naturelle pour un tort commis au passé n'a plus lieu d'être dans une société cultivée. La punition ne doit servir qu'à décourager le comportement auquel est assigné un blâme, alors que la récompense doit inciter à reproduire le comportement félicité.

---

[22] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p11.
[23] M. Schlick, cité dans TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p12.

Cette approche est aussi défendue par J.J.C. Smart pour qui un comportement blâmé est simplement un comportement évalué comme négatif, alors qu'un comportement félicité est considéré comme positif, tout en reconnaissant une responsabilité à l'agent pour son propre comportement. A ce titre, Smart considère qu'un agent responsable aurait agi autrement s'il avait été incité à agir d'une certaine façon, ou découragé d'agir d'une autre. Le problème d'une telle approche, comme le souligne Wallace, est qu'elle ouvre la porte à certaines dérives, qui ont été fortement critiquées :

> (A forward-looking approach, with its focus on bringing about desirable outcomes) is not directed exclusively toward the individual agent who has done something morally wrong, but takes account of *anyone else* who is suceptible to being influenced by our response. (Wallace 1996:56; emphasis added)[24]

Sous cet angle, une telle approche pourrait cautionner la condamnation d'individus innocents, sous le prétexte qu'une telle condamnation aurait des effets positifs sur les comportements futurs d'autres individus. C'est évidemment ce dernier point qui a valu de lourds reproches à l'approche forward-looking telle que décrite par Smart.

Après avoir discuté l'approche « forward-looking », nous allons maintenant nous intéresser à l'approche dites des « reactives attitudes » ou encore de la « quality of will ». En 1962, P.F. Strawson publie un article qui décrit une nouvelle approche en ce qui concerne la responsabilité morale qui se concentre sur les « reactive attitudes ». Strawson note l'importante différence de traitements entre deux individus ayant blessé leur voisin lorsque le premier l'a fait par accident et le second l'a fait motivé par un souhait malveillant de causer du tort à l'autre. C'est ce que Strawson nomme la « quality of will », qu'il juge fondamentale pour nos relations interpersonnelles. Pour Strawson, le fait même de tenir autrui responsable de ses actes représente une réponse « to the quality of others' will towards us »[25]. Il insiste sur « the importance that we attach to the attitudes and intentions towards us of other human beings »[26]. Cet appel aux intentions de l'agent fait naturellement écho à un paragraphe de l'introduction de cette partie dans lequel nous avions évoqué la réserve que nous avions à mobiliser le concept d'intention quand bien

[24] R. Wallace, cité dans TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p13.
[25] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p15.
[26] Idem, p15.

même nous sommes bien incapables de les connaître avec précision et certitude. Néanmoins, l'approche de Strawson se focalise sur les relations interpersonnelles, plus que sur l'introspection de l'agent, et dès lors, c'est surtout la manifestation des intentions, plus que les intentions elles-mêmes, qui, par défaut, sont considérées.

> When someone explains that the injury she caused you was entirely unforeseen and accidental, she indicates that her regard for your welfare was not insufficient and that she is therefore not and appropriate target for the negative attitudes involved in moral blame.
>
> Note that the agent who excuses herself from blame in the above way is not calling into question he status as a generally responsible agent: she is still open to the demand for due regard and liable, in principle, to reactive responses.[27]

Le point le plus intéressant de cette dernière situation est sans doute de constater que ce qui est fondamental pour l'agent, c'est d'être considéré comme un agent moral capable de réagir de manière appropriée. On retrouve dans cette approche des similitudes avec la théorie du sociologue Erving Goffman, relative à la préservation de la face. En effet, cette théorie présente un certain nombre de rapports interpersonnels comme des actions quasi rituelles que l'on effectue afin de préserver notre face, ou la face d'autrui. L'exemple le plus marquant est sans doute celui de la personne qui oublie de présenter un article à la caisse d'une échoppe et dont l'honnêteté est publiquement remise en question. Lorsque la personne en charge de la caisse attire son attention sur l'article qu'elle a mis dans son sac, la personne suspectée d'être malhonnête va exprimer de manière presque excessive sa surprise, va assurer qu'elle avait oublié cet article et qu'en aucun cas elle n'avait l'intention de partir sans payer. Il s'agit, pour elle, de ne pas perdre la face, et par extension, dans le parallèle que nous dressons ici, d'assurer l'autre qu'elle est une personne honnête et fiable, qu'elle ne mérite pas d'être traitée comme une voleuse, et que l'on peut faire confiance dans la qualité de sa moralité.

On observe assez distinctement que l'humain semble tirer une véritable fierté de sa qualité morale. Lorsqu'un individu cause du tort, il va vouloir montrer que ce tort était accidentel et qu'il devrait être excusé pour cette action. Faire partie de la communauté morale est un indicateur de valeur hautement significatif pour les êtres humains, qui, bien que prétendant parfois à une certaine bien pensance, considère avec condescendance ceux

[27] Idem, p16

qui ne font pas partie de cette communauté. Néanmoins, il faut noter que la ligne entre l'appartenance et l'exclusion à la communauté morale est plus floue qu'il n'y parait, ce n'est pas une question strictement manichéenne. Ainsi, certains comportements sont attendus d'un enfant d'un certain âge alors qu'il sera exempté de responsabilité morale pour d'autres comportements. L'enfant se trouve dès lors être partiellement un agent moral, sans pour autant l'être entièrement. Il en va de même, au cas par cas, pour bon nombre de ceux qui sont généralement exemptés de responsabilités morales, on exige tout de même bien souvent d'eux une certaine responsabilité, en fonction de leurs capacités.

Ce principe est exacerbé par ce qui semble être une tendance naturelle à désigner un coupable. L'être humain excelle dans une capacité qu'il partage avec d'autres animaux, la capacité à trouver une explication. Qu'elle soit rationnelle, scientifique, ou mythologique, religieuse, tous les évènements, en particulier ceux qui sont chargés moralement, sont interprétés et justifiés. A ce titre, on trouve un véritable attrait pour les explications qui désignent un coupable, contrairement aux explications qui remettent la faute sur la malchance, sur un contexte malheureux. Ces deux derniers éléments d'ailleurs sont parfois expliqués eux-mêmes par certaines superstitions afin de récupérer, coûte que coûte, un coupable initial. Un tremblement de terre est attribué à un être mystique, dont la colère est elle-même reprochée au manque de piété de ses fidèles. Et si une seule maison est détruite par la catastrophe, la superstition voudra que son propriétaire se soit rendu coupable de tel ou tel acte impie. Ce genre de modèle a existé de tout temps, et n'a pas du tout été effacé par les techniques modernes à même de comprendre avec plus de précision les circonstances qui expliquent les évènements naturels. Si très peu de personnes croient encore que la foudre est l'étincelle qui jaillit de l'enclume frappée par Mjolnir de la main de Thor, il n'empêche que l'esprit humain se trouve toujours très inconfortable lorsqu'un accident laisse des victimes sans coupable. C'est en particulier vrai dans deux cas. D'abord lorsqu'une personne décède, ses proches sont souvent prompts à accuser le coupable idéal, dans un mécanisme de deuil qui cherche à expliquer l'injustice en en attribuant la responsabilité à un coupable. Dans ce cas, connaître le coupable, voire parfois même le voir de leurs yeux, peut faire partie d'un processus de deuil, afin de chercher à apporter une forme de « closure » aux proches de la victime. D'une manière similaire, la recherche d'un coupable est aussi exacerbée lorsque des dégâts matériels peuvent être réparés. Le dicton « qui casse paie » exprime aussi nettement que possible la nécessité, en

cas de dégâts, de pouvoir pointer du doigt le coupable. Dans ce cas, la manière dont est traité la circonstance accidentelle du dommage variera en fonction du contexte, et généralement en fonction de la responsabilité reconnue dans la prise de risque vis-à-vis de l'intégrité de la propriété d'autrui. C'est une piste que nous explorerons davantage ultérieurement.

Cette nécessité à désigner un coupable, et parfois même un bouc-émissaire, encourage parfois à élargir les portes de la communauté morale, à certains qui s'en voient traditionnellement refuser l'accès mais qui se trouvent être les coupables idéaux dans ces circonstances. Ceci peut parfois chez certaines personnes s'exprimer sur la forme d'un paradoxe conscientisé entre la réaction théorique à adopter et la réaction effectivement adoptée : « Je sais qu'il n'y peut rien, mais je ne peux pas m'empêcher de lui en vouloir ». Une telle phrase est souvent entendue dans ces contextes, et représente bien la suspension de logique que provoque la volonté viscérale de désigner un coupable. Cette incohérence tient sans doute particulièrement à la projection de sa propre réflexion dans autrui. L'autre est considéré comme un autre moi, et dès lors il peut être difficile d'envisager qu'il ne soit pas détenteur des mêmes capacités. Si les capacités physiquement sont facilement observables, si certaines capacités mentales peuvent bien souvent l'être aussi, il en est autrement des capacités morales qui, si elles ne sont pas indéterminables, sont par nature nettement moins transparentes. Dès lors, si autrui semble parfois exhiber quelques similitudes que ce soit avec notre représentation de nous-même, il peut être problématique d'évaluer exactement ses capacités morales pour déterminer s'il doit ou non être exempté de responsabilité morale. Cette incertitude apparait dans les deux sens avec d'une part les enfants que l'on « adultise » et d'autre part les adolescents que l'on infantilise. De la même manière, les attentes vis-à-vis des personnes démentes ou handicapées mentales sont très souvent incorrectement calibrées, et laisse place à des situations soit d'attentes disproportionnées, soit de manque de respect de leur capacité à être responsable moralement, et par extension, de leur personne en général. Enfin, une autre catégorie d'êtres vivants qui fait l'objet de traitements inégaux est évidemment celle des animaux, vis-à-vis desquels les attentes varient de l'indifférence de ceux qui adoptent les théories peu contemporaines de l'animal machine de Descartes, à l'excès d'anthropomorphisme qui évalue les comportements animaux d'une manière similaire aux comportements humains. En ce qui concerne les animaux, nous pouvons assez nettement constater que

l'approche de l'animal machine, qui renie toute compétence morale à l'animal, est probablement dépassée. Certains animaux domestiques témoignent de comportements sociaux que l'on peut interpréter comme des marques de responsabilité morale. Et leurs comportements peuvent parfois être interprétés à la lumière de l'approche de « quality of will ». C'est le cas lorsqu'un chien, qui a abimé quelque chose et s'apprête à être grondé, adopte une posture que l'humain traduit comme une volonté de manifester qu'il n'avait pas l'intention de causer de la peine. Nous verrons plus en détail les similitudes entre animaux et humains sur cette question, ainsi que comment la comparaison avec l'animal nous permet d'approcher la comparaison avec l'IA.

La troisième approche que nous allons considérer est l'approche dite de la chance morale. D'après Thomas Nagel, une personne est sujette à de la chance morale si des facteurs qui ne sont pas sous son contrôle affectent l'évaluation morale dont elle fait l'objet. L'exemple classique présente un tireur qui tente d'assassiner quelqu'un mais échoue parce qu'un oiseau passe dans la trajectoire du tir et intercepte la balle. Un tel évènement produit un effet de chance morale positive sur le tireur, qui n'est dès lors pas responsable d'un assassinat mais seulement d'une tentative, ce qui serait considéré comme moins reprochable moralement. Si le tireur n'est pas jugé de la même façon en fonction d'éléments indépendants de son action, c'est que l'on adopte une posture fondamentalement conséquentialiste. C'est d'ailleurs la posture qu'adopte généralement la loi, qui si elle punit les crimes sans victimes, est bien plus sévère à l'égard de ceux qui ont eu la « malchance morale » de porter à conséquences. Il est reprochable de conduire au-delà des limitations de vitesse, mais ce n'est pas comparable avec les reproches qui seront effectués à celui dont l'excès de vitesse provoque un accident grave. Bien qu'il ne faille pas faire l'amalgame entre illégal et immoral, qui appartiennent à deux sphères nettement distinctes, la loi et son application restent souvent un indicateur de la manière dont des actes sont reçus par la société.

Néanmoins, il semble curieux que deux tireurs identiques correspondant à la description initiale mais ne différant l'un de l'autre que par la présence ou l'absence d'un oiseau dans la trajectoire du tir, ne peuvent être drastiquement évalués différemment, après tout, un évènement chanceux en dehors de leur contrôle ne devrait pas modifier profondément leur responsabilité morale. Le conséquentialisme dont une telle approche fait preuve semble en désaccord avec la logique vue dans les paragraphes précédents qui

voulait assigner une valeur morale moins à l'action (et donc à ses conséquences) qu'à l'acteur lui-même. Certains défendent que pour considérer que ces deux tireurs, dont seule la chance morale varie, sont plus proches que distants, il faut s'en tenir à leur propriétés internes, c'est-à-dire les motifs de leurs actions ainsi que leurs intentions.

> On the other hand, one might think that if the two assassins just mentioned are identical in terms of their values, goals, intentions, and motivations, then the addition of a bit of luck to the unsuccessful assassin's story cannot ground a deep contrast between these two agents in terms of their moral responsibility. One way to sustain this position is to argue that moral responsibility is a function solely of internal features of agents, such as their motives and intentions (…). Of course, the successful assassin is responsible for something (killing a person) for which the unsuccessful assassin is not, but it might be possible to argue that both are morally responsible—and presumably blameworthy—to the same degree insofar as it was true of both of them that they aimed to kill, and that they did so for the same reasons and with the same degree of commitment toward bringing about that outcome (…).[28]

On retrouve un jugement moral qui prétend s'appliquer plus à l'agent qu'à l'action elle-même, qui n'est alors que la manifestation des intentions et des motifs de l'agent. Les deux tireurs ont fait preuve des mêmes intentions et des mêmes motifs, et se sont tenus à les voir réalisés avec la même intensité, la seule différence entre ces deux tireurs leur est entièrement extérieure.

Certains vont même plus loin, et posent la question de ce qu'il en est d'un tireur potentiel, qui n'a pas tenté d'assassiner qui que ce soit mais seulement parce que des éléments contextuels précédents n'étaient pas favorables à cette action. Ce tireur potentiel aurait assassiné sa cible si le contexte avait été plus favorable. Dans ce cas, ce tireur potentiel a eu de la "chance morale" à un moment précédent, mais étant donné qu'il aurait tenté d'assassiner sa cible si le contexte y avait été favorable, on peut se demander dans quelle mesure il n'est pas aussi coupable moralement que le tireur dont la balle a été interceptée par un oiseau, ou que l'assassin qui a tué sa cible. Un tel argumentaire s'expose à une régression à l'infini de l'agence que nous avons dans les circonstances qui forment notre contexte et qui échappent à notre contrôle.

---

[28] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p36.

> As Nagel notes, once the full sweep of the various kinds of luck comes into view, "the area of genuine agency" may seem to shrink to nothing since our actions and their consequences "result from the combined influence of factors, antecedent and posterior to action, that are not within the agent's control" If it is right, then perhaps, " nothing remains which can be ascribed to the responsible self, and we are left with nothing but a …sequence of events, which can be deplored or celebrated, but not blamed or praised. (Nagel 1976 [197 : 37]) [29]

Certains, comme Galen Strawson, concluent que l'on ne peut être véritablement responsable de nos actions. Pour Strawson, le seul être capable de mettre fin à cette régression à l'infini, à cette fuite vers le passé de la responsabilité, est un agent responsable de sa propre création. Seul celui qui est responsable de sa propre création est ultimement responsable de ses actes. Si la logique derrière l'argumentaire de Strawson se tient, elle présuppose néanmoins certains éléments, comme un déterminisme relativement fort et une notion assez faible de liberté. La seule liberté au sens fort de cet argument est exercée par l'agent capable de se générer lui-même, et rien n'affirme que cette liberté puisse être attribuée à un être humain.

Les conditions épistémiques de la responsabilité

Avant de pouvoir présenter cette nouvelle approche, il nous faut apporter quelques précisions sur les conditions épistémiques de la responsabilité. De telles conditions s'avéreront cruciales pour la suite de notre réflexion. Nous l'avons déjà vu, il n'est pas comparable celui qui commet un acte considéré comme propre à être blâmé volontairement et celui qui commet un tel acte sans avoir connaissance de sa nature reprochable. Mais la question qui s'impose dans cette situation est celle de savoir dès lors si le fait même de savoir que l'acte avait une telle nature est une responsabilité de l'agent. L'ignorance n'est pas toujours sujette au blâme, simplement parce qu'il est des informations, des connaissances que l'on ne peut raisonnablement attendre d'un agent qu'il les possède. Et par extension, le fait même de savoir qu'il fallait savoir une telle information est aussi sujet au même raisonnement, etc. Pour Gideon Rosen, un agent ne peut être responsable d'un acte moralement répréhensible que s'il démontre une responsabilité originale dans au moins un acte passé. On peut s'imaginer que ce soit l'acte en question que l'agent commet en connaissance de cause, mais ce peut aussi bien être

[29] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p37.

bien plus tôt dans la chaine des causes un acte passé, ou l'absence d'un acte (ce qui, en soit, est aussi un acte), qui a amené l'agent à manquer d'informations sur les conséquences de l'acte final. L'agent est donc responsable si un tel acte passé lui est imputable, s'il a commis un « knowing sin »[30], un acte immoral dont il avait connaissance qu'il était immoral. Il n'est d'ailleurs que responsable de cet acte immoral, le reste ne fait que découler logiquement de cette erreur initiale. Si aucun acte de cet ordre ne peut être imputé à l'agent qui a commis un acte immoral par ignorance, alors il ne peut être tenu responsable d'un tel acte. L'idée qu'il ne puisse être responsable que de l'erreur initiale renvoie au même raisonnement que celui tenu dans le cadre de la chance morale. Si par exemple un individu ne sait pas qu'être raciste est reprochable moralement, il est presque logique qu'il agisse de manière raciste une fois en présence d'une personne dont il ne partage pas l'origine ethnique. Néanmoins, ce qui lui est reprochable, c'est moins le fait d'avoir agi de manière raciste, parce qu'il n'est pas forcément responsable d'être entré en contact avec cette personne, que le fait d'avoir ignoré les éléments autour de lui qui lui permettait de savoir qu'être raciste est immoral. Ainsi, cette personne est tout aussi blâmable qu'un autre individu qui partage la même ignorance de l'immoralité du racisme mais ne l'a pas exprimé parce qu'il n'est jamais entré en contact avec une personne d'une origine ethnique différente. Pour citer un exemple similaire que Matthew Talbert présente :

> A slaveowner, for example, might think that slaveholding is permissible, and so, on the account considered here, he will be blameworthy only if he is culpable for his ignorance about the moral status of slavery, which will require, for example, that he ignored evidence about its moral status while knowing that this is something he should not do.[31]

Le problème de cette approche est qu'elle ne prend pas en compte la possibilité pour un agent de comprendre l'immoralité d'une telle action une fois mis face à la situation. Mais étant donné qu'initialement, une telle approche considère un déterminisme relativement fort, le fait même d'être capable de comprendre l'immoralité de l'action démontrerait en même temps que l'agent a déjà en lui le bagage de connaissances nécessaires pour faire face à une telle situation et agir moralement. Un tel individu n'est

[30] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p47.
[31] TALBERT, M., « Moral Responsibility » in Stanford Encyclopedia of Philosophy, (2019), p48.

pas effectivement ignorant, mais a déjà acquis certaines connaissances qui lui permettent d'agir moralement en de telles circonstances.

Afin de conclure clairement cette précision, il s'agit de retenir que la question de l'information est importante dans l'attribution de la responsabilité morale. Ce que nous concluons, c'est qu'un agent ne peut être tenu responsable d'un acte qu'il a posé qu'à la lumière de l'information dont il disposait à ce sujet. C'est-à-dire que si je ne sais pas qu'appuyer sur un interrupteur dans mon bureau risque de faire exploser une centrale électrique, il ne peut m'être reproché d'avoir fait exploser la centrale en question si j'appuie effectivement sur l'interrupteur. De manière similaire, si les informations auxquelles j'ai accès me laisse penser qu'appuyer sur tel bouton provoquera une explosion meurtrière, le fait même d'appuyer sur ce bouton peut m'être reproché quand bien même celui-ci ne provoque en aucune façon quelque explosion que ce soit. Il est ici primordial de prendre en compte l'information dont dispose l'agent et à la lumière de laquelle il pose une action. Cela étant dit, il faut garder à l'esprit que le fait de s'informer peut, en soit, constituer un acte moral. Si nous reprenons l'exemple de l'ouverture de la porte que nous avons évoqué au début de cette partie, si je ne sais pas que quelqu'un se trouve derrière la porte, l'ouvrir et de ce fait bousculer la personne derrière la porte peut difficilement m'être reproché. Néanmoins, si le fait que je ne sache pas que quelqu'un se trouve derrière la porte n'est pas lié à l'opacité de la porte (ce dont je serais innocent), mais bien à ma propre imprudence de fermer les yeux plutôt que de regarder à travers la vitre de la porte, je suis moralement responsable de ne pas savoir que quelqu'un se trouve derrière cette porte, et ce peut naturellement m'être reproché. Au risque de répéter le propos, c'est d'ailleurs la seule chose qui peut m'être reprochée. Si par extension on me reprochera sans doute la bousculade, en réalité, le seul acte immoral que j'ai commis est d'avoir manqué de prudence dans la collecte d'informations avant d'ouvrir cette porte, ni plus, ni moins. C'est exactement cette prise de risque que nous allons conceptualiser dans le passage suivant.

L'approche par prise de risque

Les différentes approches que nous avons discutées précédemment, ainsi que les premiers points problématiques à partir desquels nous avons introduit notre questionnement, ont soulevé autant de questions, en ce qui concerne la façon d'approcher l'action morale et la responsabilité qui est imputable à l'agent. Dans la suite de ce chapitre,

nous allons proposer une autre façon de considérer l'action et la responsabilité, qui se base principalement sur le questionnement de la chance morale, mais qui tente d'en retourner le problème et de le voir comme une opportunité. Nous nommerons cette approche l'approche par prise de risque. Reprenons le problème du tireur dont la balle est interceptée par un oiseau. Ce qu'il a de commun avec le tireur similaire dont la balle a atteint sa cible, c'est de manière très précise le risque qu'ils ont tous les deux pris. La connaissance des potentielles conséquences de leurs actes est telle que les deux tireurs savaient sans doute, s'ils avaient pris le temps de prendre en compte le potentiel d'un improbable imprévu, qu'ils avaient un très haut pourcentage de chance de tuer leur cible. Cependant, la complexité du contexte dans lequel le tireur effectue son action est telle qu'une certitude de l'ordre de 100% est virtuellement inatteignable[32], mais la qualité de la préparation, du matériel ainsi que l'expertise du tireur lui permettent de s'approcher autant que possible de cette limite. De la même manière, si le tireur doit choisir au hasard entre un fusil chargé et un fusil non-chargé, sans qu'aucun élément ne lui permette de les différencier, mais sachant que l'un d'entre eux est chargé et que l'autre ne l'est pas, il se rendra alors au maximum responsable d'avoir mis en danger la vie de sa cible par un facteur d'une chance sur deux. Qu'il choisisse l'arme chargée ou celle qui ne l'est pas ne dépendant en aucune façon de sa propre action, vu qu'il n'y a rien qu'il puisse faire pour influencer cet évènement, ceci ne change rien à sa responsabilité. Bien entendu, si le tireur ne disposait que d'une arme dont il ne pouvait savoir l'efficacité, on peut argumenter qu'il est aussi responsable que s'il disposait d'une arme chargée, parce qu'il bénéficie d'une forme de chance morale sur le contexte qui l'amène dans cette situation[33]. Mais s'il choisit sciemment pour son tir une arme dont il sait qu'elle n'a qu'une chance sur deux d'être chargée plutôt qu'une arme dont il sait avec certitude qu'elle est chargée, dans une forme de roulette russe, il aurait dès lors pu mettre la vie de sa cible dans un plus grand danger qu'il ne l'a fait, et est dès lors responsable de la mise en danger, que l'arme s'avère avoir été chargée ou non, dans un moindre facteur que s'il avait choisi l'arme qu'il savait chargée.

L'approche par risque, qui consiste à calculer les risques d'obtenir des conséquences jugées comme reprochables moralement, permet de se concentrer sur

[32] Il n'est pas possible de prendre en compte tous les évènements possibles, du tremblement de terre à la foudre qui frappe, du passage d'un oiseau au spasme incontrôlable, etc.
[33] C'est le problème de l'information que nous avons exposé dans le chapitre précédent.

l'action de l'agent. Cette approche évite à la fois une réduction au conséquentialisme, qui juge que le tireur moralement chanceux n'a tué qu'un oiseau et est donc exempt de reproches quant au meurtre (potentiel) de sa cible, et évite aussi un recours à l'intention, dont nous savons qu'elle est problématique, et sujette à diverses complications. Il est cependant tout à fait inutile de recourir à l'intention quand nous pouvons analyser l'action à la lumière des faits dont on attend de l'agent qu'il les connaisse et des probabilités que telle ou telle conséquence découle des risques pris par l'agent. Ainsi, lorsque je frappe du pied dans un ballon de foot en direction du goal, je prends le risque de cogner le gardien de but à la tête avec la balle, en particulier si ma maitrise du ballon est peu précise, et si le gardien manque de réflexe. Mais un tel choc fait a priori partie des risques auquel doit s'attendre un gardien de but. Le goal n'est pas particulièrement l'endroit où il faut se tenir si l'on veut éviter que n'arrive vers soit des balles à grande vitesse lors d'un match de foot. Dès lors, si chacun est au courant des règles du jeu et des usages lors d'une partie de football et consent à y jouer, on peut considérer qu'il y a un contrat non-verbal entre les joueurs qui considèrent qu'un tir dans la direction globale du goal qui viendrait à cogner un joueur n'est pas la conséquence d'un acte immoral, mais simplement la conséquence de l'imperfection connue des différents joueurs dans la maitrise d'une balle qu'ils sont incapables de contrôler avec une certitude absolue. Bien que la politesse enseigne que si un joueur est assommé par la balle, le tireur prenne au moins la peine de s'excuser, nous voyons ici que celui-ci n'est en fait pas responsable d'un acte immoral. Cette absence de responsabilité du tireur a elle aussi néanmoins ses limites. Il y a un seuil de tolérance à la prise de risques qui est admise lorsqu'un groupe de personnes jouent au football, mais ce seuil est variable et peut être dépassé. Si par exemple un adulte joue au football avec de jeunes enfants, il ne sera pas raisonnable de sa part qu'il tire de toute sa puissance sur la balle, parce qu'un tel tir pourrait blesser un des jeunes enfants. Il est ici attendu de l'adulte qu'il soit à même d'évaluer la prise de risque associée à un tel tir comme immorale et décide de ne pas prendre ce risque, quel que soit l'impact sur ses chances de remporter le match. S'il choisit de prendre ce risque, le fait qu'il marque un but ou non n'aura pas d'impact sur le fait que son action imprudente soit considérée comme blâmable. De la même manière, le fait que son tir percute ou non un enfant ne changera la donne que si, dans le cas où il ne percute personne, on puisse évaluer que le tir était suffisamment loin de toute victime potentielle pour s'avérer ne pas être une prise de risque.

Ces exemples nous permettent de comprendre comment une approche par prise de risque fonctionne. Nous allons maintenant tenter d'en montrer les caractéristiques et d'en exprimer les avantages. En ce qui concerne la question du rapport entre déterminisme et liberté dans l'usage de cette approche, plutôt que de simplifier la question en optant pour un point de vue compatibiliste ou incompatibiliste, notons que cette approche requiert un certain degré de déterminisme, à savoir que l'agent doit être capable de prévoir dans une certaine mesure les conséquences potentielles et probables de ses actions, mais remarquons qu'une telle approche laisse néanmoins place à une forme d'imprévisibilité, qui n'est pas forcément pour autant un indéterminé[34]. Nous allons laisser de côté le questionnement fondamental qui cherche à savoir si le monde est ou n'est pas déterminé, parce que cette question dépasse largement le spectre de ce document, et nous contenter de constater que dans un certain nombre de situations, il n'est pas raisonnable d'attendre de l'agent qu'il sache avec certitude quelle sera la conséquence de telle ou telle action, étant donné qu'un certain nombre de variables, qu'elles soient ou non déterminées, ne sont en tout cas pas connues de l'agent. Ce qui nous intéresse dans l'angle de vue que nous prenons ici, c'est moins de percer le secret de la détermination de ses variables, que de savoir si leur connaissance est accessible à l'agent, et si oui, s'il est raisonnable d'attendre de l'agent qu'il ait connaissance de ces éléments lorsqu'il agit. Prenons l'exemple d'un lancer de pièce. Si le lancer est effectué correctement et que la pièce est correctement équilibrée, il n'est pas possible pour l'agent de déterminer s'il tombera sur pile ou sur face alors qu'elle poursuit encore son vol. Cette information n'est accessible qu'une fois que la pile a atterri. Faut-il en déduire pour autant que le résultat du lancer de pièce était indéterminé lors du vol de la pièce, rien n'est moins sûr. Un tel objet macroscopique répond aux lois de la physique, et la trajectoire de son vol ainsi que sa rotation peut certainement être connue à tout moment du vol, moyennant la connaissance de l'angle et de la puissance de l'impulsion initiale. Cependant, si la pièce laisse penser qu'elle est entièrement déterminée, et que ce n'est qu'une ignorance tout à fait acceptable qui sépare l'agent de cette connaissance, dès lors que le système est suffisamment complexe, le fait

[34] Sans entrer dans les détails de ce questionnement, nous voyons bien comment il y a une différence entre indétermination et imprévisibilité, mais comment par contre ces deux éléments sont identiques lorsque l'on s'intéresse au rapport de l'agent à l'information. La raison pour laquelle l'agent n'a pas accès à l'information, qu'elle soit parce que cette information est indéterminée ou plus simplement parce qu'elle est hors de sa portée, ne change pas la façon dont le fait même que l'agent n'ait pas accès à l'information influe son comportement et la manière dont celui-ci peut être évalué moralement.

même qu'il puisse être expliqué entièrement par quelques calculs devient plus discutable. Si je prends une centaine de petites billes et les lance dans une pièce vide, il semble déjà bien plus complexe d'estimer la position de chacune de ces billes à l'avenir. Un modèle peut nous donner une idée de ce à quoi la position de l'ensemble des billes ressemblera à l'avenir, mais ce modèle risque de ne pas résister à l'étude d'une seule bille en particulier. Il surgit ici une forme de chaos, généré par un très grand nombre de très légères modifications qui interagissent les unes avec les autres en générant toujours plus de différences. Le moindre changement dans la pression de l'air, la moindre humidité, ou quelque autre variable, aura à long terme des conséquences énormes. S'il est encore ici trop tôt pour dire que la position finale de chacune des billes est indéterminée alors qu'elles sont encore en mouvement, on peut cependant nettement voir comment au moins la connaissance de leur position future n'est pas atteignable pour un observateur lambda. Dès lors que le système est trop complexe, une connaissance parfaite du système à un instant donné devient impossible, et dès lors une connaissance approximative du système à l'avenir devient de moins en moins possible. D'où notre recours à la probabilité pour conceptualiser la prise de risque. Il est illusoire de prétendre que l'agent peut connaitre les conséquences de ses actes, il est par contre tout à fait pertinent d'imaginer qu'un agent est en mesure d'estimer la probabilité que ses actions aient telle ou telle conséquence. En l'occurrence, la réduction que nous opérons consiste à ne pas s'intéresser au système en soit, avec toute sa complexité qui le rend opaque à nos yeux, mais de le réduire au modèle dans lequel l'agent prend sa décision.

En réalité, j'argumenterais ici que toute action est toujours effectuée par un agent à la suite d'une modélisation de la réalité à l'aide de laquelle l'agent prétend estimer les conséquences de son action. L'agent prend alors en compte trois éléments fondamentaux. Tout d'abord la prédictibilité de son action dans ce modèle, ensuite sa capacité à répliquer l'action prévue avec précision dans le monde réel, et enfin l'imperfection du modèle par rapport à la réalité. Ce modèle, l'agent le sait imparfait et peut estimer son degré d'imperfection. Lorsqu'un tireur expérimenté vise une cible, il sait que lorsqu'il appuie sur la gâchette, dans le modèle qu'il a construit pour comprendre les conséquences de son action, le projectile va atteindre sa cible. Il sait, de par son expérience, qu'il est capable de reproduire le tir qu'il a modélisé avec une grande précision dans la réalité. Mais il sait aussi que ce modèle n'est pas la réalité, et en varie par un léger pourcentage. Ce léger

pourcentage explique comment un oiseau pourrait intercepter le projectile, ainsi que tout autre imprévu ayant échappé au modèle. Evidemment, ce processus de modélisation qui précède toute action est presque toujours inconscient. Il en va de même pour le rapport à l'imperfection de notre capacité de modélisation. Ceci n'empêche en rien que pour poser une action, l'agent a besoin d'avoir une idée de ce que peuvent être les conséquences, il a besoin d'estimer la manière dont son action pourrait influencer le contexte dans lequel il la pose. Cette estimation, il l'opère à l'aide de ce modèle. Pour que cette modélisation soit pertinente, le monde dans lequel l'agent agit doit être au moins partiellement déterministe. Il peut subsister une part d'indétermination, mais le socle de l'action doit pouvoir être posé sur des bases de prévisibilité suffisamment solides que pour que le concept de responsabilité puisse être pertinent. Notons brièvement qu'en ce qui concerne l'agent, le fait que ce qu'il ne peut pas connaitre lui-même soit fondamentalement indéterminé ou simplement étranger à sa connaissance n'a aucun impact sur sa capacité à poser une action. Cette question, très pertinente pour la physique, n'est ici pas particulièrement une question qui concerne l'action d'un individu lambda, l'indétermination fondamentale ou épistémique se traduisant de la même façon pour l'individu lambda, à savoir des connaissances auxquelles il n'a pas accès. Ce même rapport s'opère quant à la réalité de la liberté qu'exerce l'agent par ses choix. Que cette liberté soit apparente, c'est-à-dire que l'agent ne semble être libre que parce que nous n'avons pas accès à certaines informations, ou qu'elle soit réelle, c'est-à-dire qu'elle soit le fruit d'une indétermination fondamentale face à deux actions potentielles, n'impacte à nouveau pas l'agent qui choisit. Cette question est évidemment très pertinente pour le philosophe qui s'interroge sur la nature de la liberté, mais étant donné que nous n'avons ici pas la prétention de résoudre cette problématique, nous nous contentons de remarquer qu'en l'absence de certitude à ce sujet, les deux possibilités coexistent et ne mettent pour autant pas à mal la théorie que nous défendons.

Cette théorie soulève néanmoins son lot de questions. Qu'en est-il de la manière dont nous sommes sensés calculer la prise de risque, ou bien déterminer de quelle information un agent devrait ou non disposer. De la même manière, conformément à l'adage qui dit qu'on « ne fait pas d'omelettes sans casser des œufs », toute action un tant soit peu complexe constitue en elle-même une prise de risque. Le fait même d'être strictement immobile peut être considéré dans ce paradigme comme une action avec ses

propres risques. Il semble au premier abord qu'une telle théorie pousse à une prudence excessive, qui n'est pas un terrain propice à l'action morale. Ce n'est pas la conclusion que nous souhaitons ici tirer. Certes cette approche n'est pas favorable au fait de prendre des risques inconsidérés, mais c'est le cas de toute approche qui considère la responsabilité morale. Dès lors que la responsabilité morale est en jeu, le fait de prendre des risques inconsidérés est plus ou moins immoral. Si quelqu'un conduit sous influence, même s'il n'a pas d'accident et que son imprudence ne cause de tort à personne, le risque qu'il a pris est, dans la plupart des contextes, immoral. Pour justifier une telle prise de risque, l'agent en question devrait mettre dans la balance une justification de taille. Admettons que ce conducteur a conduit sous influence parce que c'était la seule solution par laquelle il pouvait sauver la vie de nombreuses personnes. Dans ce cas, on pourra sans doute déterminer que le risque en valait la chandelle, quel que soit le résultat de la conduite de l'agent. A l'inverse des risques inconsidérés, se trouvent donc des risques considérés. Ces risques considérés, ce sont ceux que nous prenons dans toute notre vie quotidienne, et ce sont des risques, soit très élevés d'une conséquence très peu dommageable, soit moyennement élevé d'une conséquence moyennement dommageable, soit très peu élevé d'une conséquence très dommageable. Lorsqu'un individu touche une poignée de porte dans un espace public, il y a un très grand risque qu'il salisse cette porte, par une très faible quantité de poussière, de transpiration. Ce risque est naturellement admis par tout un chacun, et personne ne considérerait, sur une poignée de porte commune, qu'il soit immoral de la salir de la sorte, alors que l'on pourrait reprocher à quelqu'un d'entrer chez quelqu'un avec des bottes couvertes de boue. C'est accepté, non pas parce que le risque est faible, mais au contraire parce que les conséquences, quand bien même le risque est effectif, sont minimes. De l'autre côté du spectre, nous avons en Belgique un certain nombre de réacteurs nucléaires, qui sont proportionnellement à d'autres centrales, très sûrs. Le risque qu'ils provoquent une explosion nucléaire est très faible. Le problème est évidemment qu'une telle explosion serait catastrophique et pourrait bien couter de nombreuses vies. Cependant, une même production énergétique par des centrales à énergie fossile rejette un tas de particules toxiques dans l'atmosphère. Discrètement, ces particules peuvent nous rendre malades, détériorer nos conditions de vies, voir nous tuer. Ce à quoi il s'agit de comparer la prise de risque du choix politique d'utiliser des centrales nucléaires, c'est au risque bien moins grand mais bien plus certain de la toxicité de l'alternative. Evidemment, ce calcul ne prend pas en compte les questions

environnementales qui sont primordiales, ni le coût de telles infrastructures, ainsi que d'autres alternatives qui pourraient être à la fois moins dangereuses et moins polluantes. Ces exemples ont simplement pour objectif de mettre en évidence le rapport entre la probabilité de réalisation du risque et la gravité du risque, qui ne peuvent être considérés qu'à la lumière l'une de l'autre.

Il reste à apporter un éclairage sur les questions que nous avions soulevées. Tout d'abord, nous n'allons pas déterminer quels risques sont ou ne sont pas acceptables, parce qu'une telle question revient à demander ce qui est ou qui n'est pas moralement acceptable. Evidemment, une théorie comme la nôtre qui prétend proposer une façon d'approcher l'action morale, sous le prisme de la prise de risque, ne prétend pas du même coup résoudre la totalité des questions éthiques. Il s'agit plutôt de proposer un cadre plus clair, et de se débarrasser de certaines préoccupations problématiques qui parasitent notre capacité à considérer les questions plus fondamentales. A titre d'exemple, la pratique du ski est un sport relativement dangereux, pourtant, dans un cadre sécuritaire suffisant, il n'est pas considéré dans notre société comme immoral d'emmener des enfants faire du ski. Ceci étant dit, il ne parait pas aberrant d'imaginer qu'une société qui valoriserait moins le divertissement et la pratique sportive puisse ne pas évaluer les avantages obtenus au prix des risques encourus comme valant la peine de pratiquer le ski, et pourrait donc considérer comme immoral d'exposer des enfants à de tels risques. Ceci renvoie à la question du relativisme moral, que nous avons discuté précédemment, et que l'on conclut d'une manière similaire à la question de l'information. Qu'il existe ou non une vérité morale absolue, celle-ci ne nous est de toute façon pas accessible, et dès lors il n'est pas pertinent de considérer que l'on puisse l'utiliser pour déterminer la validité morale d'une prise de risque. Nous sommes donc contraints de nous référer à ce qui est moralement acceptable dans une société donnée.

Enfin, nous avons évoqué la façon de calculer de tels risques, par le processus de la modélisation. Mais nous n'avons pas éclairé la façon d'évaluer l'imprécision du modèle, la précision de l'agent dans sa capacité à répliquer l'action modélisée dans le monde réel, ni le degré de prédictibilité de l'action dans le modèle lui-même. A ce titre, notons que le rapport même qu'un agent entretient avec cette capacité de modélisation et d'évaluation de son action est en soit une action. C'est-à-dire que si un agent fait preuve d'excès de confiance dans la précision de sa modélisation, il s'expose à agir de manière

inadéquate de manière répétitive, et son tort moral sera ici dans cet excès de confiance, plus que dans tout autre conséquence de celui-ci. Au-delà de cette note, il semble bien que la capacité de modélisation, en ce y compris la capacité à évaluer ces trois imprécisions, est une capacité que l'on est en droit d'attendre d'un agent moral. Dans le début de cette partie, nous avions évoqué comment l'agence morale était traditionnellement l'apanage d'un certain nombre de personnes, à savoir les adultes en pleine possession de leurs capacités mentales. En réalité, cette théorie affirme que ne sont capables d'être des agents moraux à part entière que ceux qui sont capables d'effectuer cette modélisation, de comprendre la portée de leur action, de comprendre les conséquences potentielles et d'évaluer moralement ces conséquences potentielles. Si nous prenons en compte une certaine imprécision, qui va jusqu'à être capable d'estimer l'imprécision dans le calcul même de ces imprécisions[35], c'est parce que le monde dans lequel nous vivons nous force à réaliser que les êtres humains ne sont pas des agents moraux parfaits. Il n'est d'ailleurs pas certain qu'un agent parfait puisse exister, dans le cas où l'avenir n'est pas déterminé. Quoi qu'il en soit, cette imperfection de l'agent est tout à fait acceptable et doit être admise chez l'être humain. Nous ne pouvons naturellement nous bercer de l'illusion que nous avons une maitrise complète sur les conséquences de nos actions. Par contre, nous devons tendre vers l'agent moral parfait autant que possible. La question qui occupera notre troisième partie sera notamment de tenter d'évaluer dans quelle mesure une IA peut prétendre se rapprocher de cet agent moral parfait. Si cette entreprise est vouée à l'échec comme certains le pensent, ou si au contraire, un agent artificiel pourrait être le seul agent moral parfait possible.

Introduction à quatre courants éthiques

Ce que nous avons vu jusqu'ici et en particulier l'approche par risque se présente à nous comme des outils nous permettant d'envisager le rapport entre action morale, agent moral et responsabilité avec plus de précision. Il a s'agit jusqu'ici de montrer le rapport à l'information disponible, de déconstruire le rapport entre responsabilité et conséquences

[35] Au risque d'être redondant, un agent doit être capable de se rendre compte qu'il n'est pas en mesure de calculer l'imprécision de son modèle de manière parfaite, mais qu'il ne peut que l'estimer. Cette estimation constitue en soi une imprécision que l'agent doit reconnaitre et évaluer. Il y a ici une forme de régression à l'infini. J'argumenterai que cette régression à l'infini n'est pas ici problématique vu que chaque erreur sur l'imprécision suivante n'est que toujours plus petite, et l'agent est en mesure d'estimer vers quelle « valeur » tend le calcul de son imprécision. Les lointains chiffres après la virgule n'ont que très peu d'importance.

effectives de l'action afin de souligner la logique selon laquelle les conséquences prévisibles sont les seules qui devraient être envisagées moralement. Nous avons aussi montré comment l'usage du concept d'intention, bien qu'intéressant au premier abord, desservait nos objectifs. Mais nous n'avons encore rien dit de ce qui devrait être considéré comme moral, ce qui devrait être considéré comme souhaitable. Evidemment, nous n'allons ici pouvoir qu'envisager la question de manière relativement superficielle, et proposer des pistes en ce sens, mais la question de déterminer quelle théorie éthique doit être mobilisée en particulier dans le rapport à l'IA est une question encore ouverte aujourd'hui. Nous avons sélectionné trois courants majeurs. Nous allons essayer de mettre en évidence leurs forces et leurs faiblesses, et nous montrerons dans la troisième partie comment ils peuvent interagir avec ce que nous savons de l'IA. D'abord les plus attendus, avec d'une part l'utilitarisme, dont l'omniprésence dans le discours moral moderne n'aura échappé à personne et d'autre part le déontologisme, qui, assez intuitivement, se prête particulièrement à une approche descendante[36] de l'IA. Dans la même idée, nous allons aussi rapidement évoquer l'éthique de la vertu, dont nous verrons dans la dernière partie de ce travail comment elle peut sembler pertinente pour une application à l'IA. Enfin, nous allons évoquer une théorie qui n'est généralement pas attendue dans le cadre qui est le nôtre, et qui est l'éthique du Care. Ce choix de théories se veut aussi succinct que possible tout en étant suffisamment complet que pour nous donner une bonne idée de ce que nous pouvons mobiliser pour tenter de faire progresser notre raisonnement. Il est cependant évident que chacune de ces trois théories pourrait faire l'objet d'un mémoire en lui-même, et que ce n'est pas ici ce que nous allons faire, nous allons nous contenter de citer les grandes lignes contextuelles ainsi que conceptuelles, afin de disposer des éléments nécessaires pour notre analyse.

Le premier courant dont nous allons discuter est l'utilitarisme. Nous n'allons pas ici revenir sur les détails de l'utilitarisme, ni sur les raisons pour lesquelles ce courant éthique a les faveurs d'une bonne partie de notre société, simplement nous allons ici en exposer succinctement les grands principes. L'utilitarisme, théorisé par Jeremy Bentham dans la fin du 18^{e} siècle, peut être globalement résumé par la maxime « le plus grand bonheur du plus grand nombre ». Bentham propose de la sorte ce qu'il appelle le « calcul félicifique » qui consiste à calculer la somme des bonheurs et malheurs que provoque une

[36] Voir chapitre intitulé « approche descendante et approche ascendante » à la page 63.

action. L'action choisie doit être celle qui maximise cette somme. Traditionnellement, l'utilitarisme est décrit comme une éthique conséquentialiste, vu que ce qui compte, c'est le résultat, c'est le bonheur effectivement obtenu par des personnes en résultat d'une action. Il existe par ailleurs un certain nombre de variantes de l'utilitarisme, de théories développées par la suite. L'utilitarisme de Bentham mobilise donc un calcul félicifique qui se fonde sur un certain nombre de valeurs, qu'il appelle « circonstances ». Les circonstances qualifient le bonheur attendu selon les valeurs suivantes : son intensité, sa durée, sa certitude[37], sa proximité, sa fécondité (à savoir la probabilité qu'il produise des bonheurs supplémentaires), sa pureté (c'est-à-dire le risque qu'il soit suivi de sensations opposées) et son étendue (c'est-à-dire le nombre de personnes affectées). Si certaines circonstances sont faciles à mesurer et à intégrer, comme l'étendue qui est un simple multiplicateur, d'autres le sont beaucoup moins, comme l'intensité ou la pureté. La proximité elle-même, bien que potentiellement simple à mesurer, est complexe à intégrer au calcul. Vaut-il mieux un bonheur dans une heure ou le double de celui-ci dans deux heures ? Il y a probablement une fonction logarithmique qui exprime qu'il faut de plus en plus de bonheurs au plus l'attente augmente, mais il semble impossible de pointer du doigt la fonction exacte. Le calcul félicifique, bien qu'attrayant, présente ses propres défis et ses propres faiblesses qu'il ne faut pas perdre de vue.

En travaillant en dehors du cadre de l'utilitarisme classique, nous souhaitons simplement attirer l'attention sur le fait que, contrairement à l'idée préconçue, la théorie utilitariste n'a pas d'intérêt que dans un contexte conséquentialiste. C'est certes dans ce contexte qu'il a été formulé, mais nous pouvons sans peine imaginer comment un utilitarisme probabiliste pourrait être pertinent et s'associer avec l'approche par risque que nous avons proposée dans le chapitre précédent. Dans cette idée, il s'agirait d'analyser l'action non pas en fonction de ses conséquences effectives, dont nous avons vu que le contrôle échappe partiellement à l'agent, mais en fonction des conséquences probables. Ce ne veut pas dire qu'il s'agit de réduire à la conséquence la plus probable, mais simplement qu'il faut prendre en compte la globalité des conséquences potentielles. Si un parent joue l'argent prévu pour payer les études de son enfant au casino, il a une grande probabilité de perdre cet argent, et une petite[38] probabilité de gagner plus d'argent. Pour

[37] Cette circonstance particulière est prise en compte de manière intrinsèque dans l'approche par prise de risques.

[38] Selon le type de jeu, le nombre de parties jouées et le retour sur investissement attendu du jeu.

juger de la moralité de l'action, le calcul félicifique conséquentialiste serait dépendant du résultat, et jouerait en la faveur du parent s'il a gagné, et en sa défaveur s'il a perdu[39]. Si nous intégrons un calcul probabiliste dans l'équation[40], et une approche en terme de prise de risques, nous voyons que le risque pris par le parent est plus important que le bénéfice potentiel obtenu, et que dès lors le fait de jouer l'argent au casino est immoral, indépendamment du gain ou de la perte effective d'argent. Pour qu'un tel risque soit justifié, il faudrait imaginer une situation caricaturale dans laquelle le parent doit disposer d'une certaine somme d'argent pour une cause bien supérieure à l'éducation de son enfant, et que le casino soit littéralement la seule option possible[41]. Une telle approche présentée par cet exemple pourra nous être utile par la suite. Néanmoins, comme toutes les théories qui mobilisent le calcul félicifique, la façon par laquelle nous sommes sensés attribuer une valeur au bonheur reste floue et problématique.

Généralement présentée comme à l'opposé direct de l'utilitarisme, le déontologisme est la théorie morale proposée par Immanuel Kant au cours de la fin du 18e siècle. Le déontologisme est une théorie idéaliste, et non conséquentialiste, en ce qu'elle considère les actions comme étant ou non morales en fonction de paramètres qui ne sont pas liés aux conséquences de l'action. L'éthique déontologiste, étymologiquement, signifie l'éthique du devoir, et pour Kant, il s'agit pour agir moralement de se conformer à son devoir. Le devoir est un impératif catégorique, c'est-à-dire qu'il faut agir conformément au devoir et pour le devoir. L'impératif catégorique, Kant l'exprime de plusieurs façons, d'abord sous la forme de l'universalisation de la maxime. L'universalisation de la maxime affirme que je ne dois agir qu'en fonction d'un principe

---

[39] Notons que nous ne prenons pas ici en compte le bonheur des propriétaires du casino, et ce afin de garder l'exemple suffisamment simple. Bien entendu, s'il s'agit de juger la moralité de l'action de bout en bout, il s'agirait de prendre en compte tout bonheur et tout malheur impacté, de près ou de loin, par l'action, en ce y compris le bonheur du parent suscité par le fait même de jouer, quand bien même il venait à perdre après, s'il a le goût du risque, etc.

40 L'équation deviendrait quelque chose du type : P(a).A + P(b).B + … + P(x).X = F où P(a) représente la probabilité d'un événement a (0.5 s'il a une chance sur deux d'avoir lieu), A représente le bonheur procuré par un évènement a (qui devrait lui-même être déterminé par un calcul félicifique classique partant du principe que l'évènement « a » a eu lieu), et ainsi de suite pour P(b) et B, etc. et que F représente la somme des bonheurs probables. L'objectif serait donc de maximiser F, en effectuant l'action dont la somme des bonheurs conséquentiels probables serait la plus élevée.

[41] Il faudrait ici un exemple caricatural où, par exemple, la vie de l'enfant est en jeu et le parent doit absolument disposer d'une certaine somme pour le sauver, alors que jouer au casino est la seule option possible pour sauver son enfant. Dans ce cas naturellement, le faire serait moral, mais le moins que l'on puisse dire est qu'un tel exemple est tiré par les cheveux.

d'action que je dois en même temps pouvoir ériger en loi universelle[42]. Le second impératif catégorique que Kant présente affirme que les personnes doivent être toujours considérées comme des fins en soi, et place la dignité humaine au-dessus d'autres préoccupations. Sa formulation classique est la suivante : « Agis de façon telle que tu traites l'humanité, aussi bien dans ta personne que dans toute autre, toujours en même temps comme fin, et jamais simplement comme moyen »[43]. L'approche kantienne se veut théorique, idéaliste, et pas empiriste ou intuitive. Pour Kant, il s'agit de poursuivre l'idéal d'une vérité morale, et tenter de proposer des méthodes pour la découvrir.

Bien entendu, le déontologisme kantien n'est pas exempt de critiques. Les critiques les plus classiques lui reprochent de ne pas être applicable, d'être paradoxale. L'exemple du mensonge est souvent repris. Le mensonge est supposément injustifiable pour Kant parce que si la porte est ouverte au mensonge la parole ne peut plus être fiable et perdrait dès lors tout son sens. Or il n'est pas difficile de penser une situation dans laquelle intuitivement on voit bien que d'un mensonge découlerait de meilleures conséquences que de la vérité. L'exemple classique considère que l'on cache quelqu'un d'un meurtrier et que l'on soit contraint de mentir au meurtrier au sujet de la position de la personne, et ce pour protéger sa vie. Kant refuse systématiquement cet argument et affirme, hors de toute logique conséquentialiste, que le mensonge est un tort en soi, et que rien ne peut le justifier. Une position aussi absolue peut s'avérer problématique à maintenir, mais il ne faut pas pour autant en déduire qu'il n'y a rien à apprendre du déontologisme kantien.

Pour compléter le tableau que nous tentons de dresser ici des différentes théories éthiques pertinentes pour notre raisonnement, nous allons maintenant aborder la question de l'éthique de la vertu. Cette théorie éthique est traditionnellement, dans la culture occidentale, attribuée à Aristote mais certains auteurs modernes mettent en évidence des liens avec la philosophie chinoise[44]. C'est le cas notamment de Rosalind Hursthouse, qui introduit son article intitulé « Virtue ethics » comme suit,

---

[42] La formule la plus fondamentale dit « Agis uniquement d'après la maxime qui fait que tu puisses vouloir en même temps qu'elle devienne une loi universelle » telle que formulée dans les Fondements de la Métaphysique des Mœurs (1985).

[43] Fondements de la Métaphysique des Moeurs, I. Kant, 1985.

[44] D'après l'article de Hursthouse, de telles racines peuvent être remontées jusqu'à Confucius, c'est-à-dire au 6e siècle ACN.

> Virtue ethics is currently one of three major approaches in normative ethics. It may, initially, be identified as the one that emphasizes the virtues, or moral character, in contrast to the approach that emphasizes duties or rules (deontology) or that emphasizes the consequences of actions (consequentialism). Suppose it is obvious that someone in need should be helped. A utilitarian will point to the fact that the consequences of doing so will maximize well-being, a deontologist to the fact that, in doing so the agent will be acting in accordance with a moral rule such as "Do unto others as you would be done by" and a virtue ethicist to the fact that helping the person would be charitable or benevolent.[45]

Elle affirme donc qu'un agent suivant l'éthique de la vertu aiderait quelqu'un dans le besoin parce que ce serait charitable et bienveillant. On peut tout de suite remarquer que l'éthique de la Vertu a le défaut de faire appel aux intentions de l'agent. Dans la suite de son article, Hursthouse affirme clairement qu'il y a une différence entre quelqu'un qui ne triche pas par honnêteté et quelqu'un qui ne triche pas par peur de se faire prendre. La question que nous devons ici nous poser, c'est de déterminer si un tel appel à l'intention rend caduque l'usage de l'éthique de la Vertu dans la réflexion qui est la nôtre. Nous allons tenter d'argumenter que ce n'est pas le cas. En réalité, il s'agit de se demander s'il y a une différence entre l'intention réelle et l'intention apparente. Il semble que pour l'éthique de la vertu, l'intention apparente n'a aucune importance, seule l'intention réelle compte. Mais si nous reprenons notre réflexion sur l'intention, il semble que l'intention réelle, inaccessible, puisse être réduite à l'intention apparente, ou au moins à une forme d'intention exprimée. Les seules manières d'accéder à l'intention dont nous disposons sont soit de faire confiance à ce que l'agent laisse transparaitre, « il ne triche pas, c'est qu'il doit être honnête ». Ou alors de se fier à ce que l'agent exprime de ses intentions, « il nous dit qu'il ne triche pas par honnêteté, c'est qu'il doit être honnête ». En réalité, la prudence pourrait nous mener à dire que tant qu'il ne triche pas dans un cadre où il pourrait être pris, on ne peut rien dire de son honnêteté. Par contre, dès lors qu'il ne triche pas alors qu'il sait qu'il aurait pu tricher sans être pris, alors seulement nous pouvons déterminer qu'il est honnête. En réalité, quelqu'un qui est face au choix de tricher et qui sait qu'il risque d'être pris ne peut jamais vraiment savoir lui-même comment il agirait s'il ne risquait pas d'être pris. Il peut affirmer et se convaincre lui-même que son honnêteté prendrait le dessus et qu'il ne tricherait en aucune circonstance, mais il ne saura comment

[45] HURSTHOUSE, R., « Virtue Ethics » in *Stanford Encyclopedia of Philosphy,* (2016), p1.

il agira en une telle situation qu'une fois qu'il est mis dans la situation. Et quand bien même, un tel raisonnement est essentialiste. Un agent pourrait bien ne pas tricher une fois, et tricher la suivante, ou inversement, selon tout un ensemble de circonstances.

Au-delà de la problématique incessante de l'intentionnalité, nous pouvons néanmoins retenir de l'éthique de la Vertu le principe par lequel elle prétend être cultivée. Afin de cultiver sa vertu, Aristote conseille notamment d'observer des personnes vertueuses, afin de comprendre comment elles agissent. Enfin, Hurtsthouse insiste sur le rapport à l'information. Nous l'avons vu dans un passage précédent dédié à cette question, mais le fait de disposer d'une certaine information peut rendre tel acte, qui serait admirable sans cette information, tout à fait reprochable. Si nous reprenons l'exemple de la porte, si je sais que quelqu'un est derrière la porte et que j'ouvre la porte violemment malgré cela, il sera logique que l'on me reproche mon action. L'éthique de la Vertu considère comme une forme de sagesse, qui est elle-même une vertu, le fait même de disposer d'informations adéquates permettant une action informée, mais aussi le fait de connaitre les limites de notre propre connaissance. Nous discuterons dans la troisième partie de ce travail comment de telles idées peuvent s'appliquer à notre questionnement.

Enfin, et bien qu'une telle approche éthique puisse sembler hors-propos dans le cadre d'un travail sur l'IA, nous allons présenter ici une théorie éthique proposée par Carol Gilligan, en 1982 dans son ouvrage « Une voix différente », l'éthique du Care. Cet ouvrage est un ouvrage significatif du féminisme américain, et c'est d'ailleurs à l'origine d'un biais de genre que se pose la réflexion de Gilligan. Celle-ci reproche à Kohlberg, un psychologue américain avec qui elle a travaillé, les présupposés d'infériorité de la moralité de la femme qui sont exposés dans son ouvrage sur la théorie du développement moral. Pour Gilligan, Kohlberg suit un raisonnement kantien et met l'accent sur les dimensions de justice plutôt que sur les dimensions de « care ». Pour Gilligan, le Care, à savoir la sollicitude ou le « prendre soin » doit être valorisé moralement au-dessus de l'idée de justice. Pour se faire, l'approche du Care ne se veut pas idéaliste, ni même conséquentialiste, et ne prétend pas affirmer de grands principes comme le fait le déontologisme. Au contraire, il s'agit dans le cadre du Care de se soucier de chaque situation indépendamment, au cas par cas. Pour les tenants de l'éthique du Care, il n'y a pas une solution définie à chaque problème moral, mais il s'agit de se tenir auprès de ceux

qui sont concernés par la question morale afin de comprendre leur façon de vivre ce dilemme, afin de comprendre comment les enjeux sont vécus.

Par ailleurs, l'éthique du Care insiste sur l'idée de vulnérabilité ainsi que sur l'idée d'interdépendance. Il s'agit de quitter la sphère idéaliste pour montrer que nous sommes des êtres vivants en chair et en os, avec nos forces et nos faiblesses, avec nos souffrances et nos joies. Ces souffrances et ces joies, il n'y a pas lieu de tenter de les mesurer ou de les comparer, au contraire, l'éthique du Care accepte d'être une éthique subjectiviste précisément parce que la démarche d'objectivation nous éloigne du souci concret de l'autre. Par ailleurs, le Care souligne les relations d'interdépendance, en montrant qu'il ne s'agit pas de simple dépendance, mais bien d'un échange. Cette approche montre aussi que toutes les relations sociales et toutes nos vies ne sont possibles qu'au travers de cette toile de l'interdépendance. Certaines auteures comme Joan Tronto ont suivi le mouvement de Gilligan en insistant sur le fait que le Care, bien que proche du féminisme, n'est pas une éthique réservée aux femmes, et ne prône pas une essentialisation de la femme comme étant plus propice au soin, au souci de l'autre. Le Care au contraire s'ouvre sur tous les agents moraux, et les invite à un regard plus subjectiviste sur les situations vécues. De toute évidence ce n'est pas ce type d'approche éthique que l'on s'attend à rencontrer dans un mémoire qui traite de l'IA, mais un tel choix est tout à fait volontaire. Bien que l'utilitarisme, le déontologisme et l'éthique de la vertu semblent nettement plus adaptés à la machine, nous allons, dans la troisième partie de ce travail, tenter de montrer comment le Care et d'autres approches éthiques plus subjectivistes, moins idéalistes et moins conséquentialistes, pourraient bien pallier les faiblesses de celles qui sont choisies traditionnellement.

Deuxième partie

Histoire de l'IA

Le terme intelligence artificielle (IA) a été entériné en 1956, dans une petite conférence devenue célèbre au Darmouth College à Hanover dans le New Hampshire. Néanmoins, ce champ de recherches et de questionnements précède évidemment cette date, si bien que l'on retrouve en 1950 un article d'Alan Turing qui propose que la question « Est-ce qu'une machine peut penser ? » soit remplacée par la question, plus pertinente à son sens, « Est-ce qu'une machine peut-être linguistiquement indistinguable d'un humain. C'est pour cela qu'il propose un test, le « Test de Turing » (TT) tel qu'on l'appelle actuellement. Ce test consiste à demander à un juge de comparer les réponses que lui remettent deux entités à des questions qu'il leur pose, ces entités sont une machine et une personne, et le juge ne sait pas lequel des deux est l'humain et lequel est la machine. Si le juge ne sait pas faire mieux que 50/50 au moment de déterminer lequel est l'humain et lequel est la machine, on peut déterminer que la machine est linguistiquement indistinguable d'un humain, et elle passe dès lors le Test de Turing. Ce TT influence encore aujourd'hui de manière significative la détermination de ce qu'est une IA. Le champ de recherche sur le développement de l'IA peut encore aujourd'hui être défini par ce test.

> The TT continues to be at the heart of AI and discussions of its foundations, as confirmed by the appearance of (Moor 2003). In fact, the TT continues to be used to define the field, as in Nilsson's (1998) position, expressed in his textbook for the field, that AI simply is the field devoted to building an artifact able to negotiate this test. Energy supplied by the dream of engineering a computer that can pass TT, or by controversy surrounding claims that it has already been passed, is if anything stronger than ever, and the reader has only to do an internet search via the string "turing test passed" to find up-to-the-minute attempts at reaching this dream, and attempts (sometimes made by philosophers) to debunk claims that some such attempt has succeeded.[46]

L'objectif est toujours de créer une machine capable de passer le TT, et malgré quelques étonnantes controverses à ce sujet, aucune machine n'a encore passé le TT à ce

[46] BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in *Stanford Encyclopedia of Philosphy*, (2018), p4.

jour. Par ailleurs, il est intéressant de noter que Turing, dans ce même article, proposait déjà des suggestions sur la façon dont une machine capable de passer le TT soit construite. Pour Turing, il serait nécessaire de créer des « child machines » (des machines-enfants) et que ces machines se développent d'elles-mêmes de manière à apprendre à communiquer naturellement avec une maitrise du langage du niveau d'un humain adulte. Nous verrons comment une telle suggestion n'est pas si lointaine de certaines approches modernes de l'IA. Néanmoins, si la conférence du Darmouth College de 1956 a entériné le terme IA, si Turing avant cela a proposé un article pertinent sur le sujet, il faut remonter bien avant cela pour pointer le début des questionnements philosophiques sur l'intelligence mécanique. Par exemple, Descartes déjà proposait en 1637 une première version de ce qu'on appelle aujourd'hui le Test de Turing :

> If there were machines which bore a resemblance to our body and imitated our actions as far as it was morally possible to do so, we should always have two very certain tests by which to recognise that, for all that, they were not real men. The first is, that they could never use speech or other signs as we do when placing our thoughts on record for the benefit of others. For we can easily understand a machine's being constituted so that it can utter words, and even emit some responses to action on it of a corporeal kind, which brings about a change in its organs; for instance, if it is touched in a particular part it may ask what we wish to say to it; if in another part it may exclaim that it is being hurt, and so on. But it never happens that it arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do. And the second difference is, that although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others, by which means we may discover that they did not act from knowledge, but only for the disposition of their organs. For while reason is a universal instrument which can serve for all contingencies, these organs have need of some special adaptation for every particular action. From this it follows that it is morally impossible that there should be sufficient diversity in any machine to allow it to act in all the events of life in the same way as our reason causes us to act. (Descartes 1637, p.116)[47]

Si l'approche de Descartes est naturellement dépassée sur certains détails, force est de constater qu'au 17e siècle déjà, il pointait du doigt les plus grands challenges auxquels le domaine de recherche de l'IA fait encore face aujourd'hui. Si la science-fiction laisse

[47] Cité dans BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in *Stanford Encyclopedia of Philosphy*, (2018).

penser que ce n'est qu'une question de temps avant que ces challenges soient dépassés, et que les tenants de la singularité sont convaincus que l'IA rattrapera et dépassera l'intelligence humaine, il y a en fait assez peu d'éléments concrets qui garantissent que le point de vue de Descartes soit un jour forcément mis à mal.

> At the moment, Descartes is certainly carrying the day.[8] Turing predicted that his test would be passed by 2000, but the fireworks across the globe at the start of the new millennium have long since died down, and the most articulate of computers still can't meaningfully debate a sharp toddler. Moreover, while in certain focussed areas machines out-perform minds (…), minds have a (Cartesian) capacity for cultivating their expertise in virtually any sphere. (…) AI simply hasn't managed to create general intelligence; it hasn't even managed to produce an artifact indicating that eventually it will create such a thing.[48]

Certes l'IA est capable de vaincre l'humain sur certains défis comme la fameuse victoire de Deep Blue contre Gary Kasparov aux échecs et l'avènement depuis lors de nouvelles IA comme Stockfish et AlphaZero qui ont révolutionné la manière dont les meilleurs joueurs d'échecs préparent leurs tournois. Mais toute proportion gardée, le jeu des échecs est finalement extrêmement simple, si on le compare à des choix plus complexes de la vie réelle. Il existe en effet un très grand nombre de positions atteignables aux échecs, mais ce nombre est étonnamment faible quand on le compare à la complexe nuance des choix qu'un être humain effectue chaque jour. Il n'y a que 18 premiers coups possibles pour l'ouverture d'une partie d'échec, et chacun de ces coups peut être rencontré par 18 réponses possibles de la part des pièces noires. Après un coup de chaque camp, il n'y a « que » 324 positions possibles. Si ce nombre parait énorme, il est en fait particulièrement petit par rapport à toute la nuance des possibilités d'une interaction sociale, ou d'un dilemme moral. Chaque froncement de sourcil, chaque mouvement la tête, chaque manière de prononcer un mot, sans même parler du choix des mots et du sens du message véhiculé font que chaque phrase prononcée dans une discussion est une phrase parmi une virtuelle infinité de possibilités d'interactions sociales. Prenons une interaction simple comme la salutation d'un voisin au sortir de sa maison. Lorsqu'une personne sort de chez elle et aperçoit son voisin, elle peut interagir avec celui-ci d'une virtuelle infinité de façons, et chacune de ces approches laisse place à autant de réponses possibles. A peine

[48] BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in *Stanford Encyclopedia of Philosphy*, (2018), p6.

sorti de chez soi depuis moins de 5 minutes, une IA qui n'aurait aucun mal à visualiser instantanément les 324 positions possibles d'une partie d'échec après un coup de part et d'autre, se trouve déjà complètement dépassée par la densité de l'interaction à travers laquelle une personne lambda navigue sans le moindre accroc. Cette personne lambda pourtant risquerait, dans une partie d'échec, de commettre une erreur dès les premiers coups, en faisant un choix sub-optimal, garantissant la victoire à son adversaire IA. L'intelligence naturelle est le fruit d'une longue évolution, nous ayant permis d'acquérir un certain nombre de capacités pertinentes pour notre contexte. La capacité d'effectuer des calculs et de planifier des centaines de positions d'échecs potentielles différentes n'est pas particulièrement incroyable, en comparaison à d'autres capacités dont l'intelligence humaine dispose, seulement, cette capacité est rare chez l'humain, et requiert de l'entrainement parce qu'elle n'est simplement pas aussi pertinente que d'autres capacités d'un point de vue évolutionniste.

Cette approche ci-dessus se concentre sur le point de vue selon lequel l'IA doit être capable de penser ou au moins de se comporter comme un être humain, comme une intelligence naturelle. Mais il est pertinent de se demander s'il n'est pas absurde de cantonner l'IA à la comparaison avec l'intelligence naturelle, alors qu'elle ne cherche pas du tout à relever les mêmes défis, qu'elle n'a pas les mêmes objectifs. L'IA n'a aucun gène à transmettre, aucune espèce à préserver, aucun instinct de reproduction. Il est évidemment plus que caricatural de réduire l'intelligence naturelle à ces éléments, et ce n'est pas ce que nous faisons ici, mais il n'empêche que ce sont des éléments qui ont un impact significatif sur la façon dont l'intelligence naturelle s'est construite, s'est développée. La tradition philosophique grecque faisait la part belle à un dualisme du corps et de l'esprit, avec certaines interprétations qui considéraient les désirs du corps comme autant de risques de se détourner de la raison, de l'esprit. Dans le Cratyle, Socrate dit du corps qu'il est « le tombeau de l'âme ». Toute une tradition philosophique, qui a ses défenseurs aujourd'hui encore, est convaincue que les ressentis du corps éloignent notre esprit de la raison, de l'intelligence la plus noble. Evidemment, un courant opposé défend que c'est précisément notre capacité à ressentir, à vivre des émotions, à souffrir, qui rend notre intelligence noble, mais il n'empêche que l'idée d'une intelligence rationnelle désincarnée en séduit plus d'un. Naturellement, cette intelligence désincarnée trouve tout son sens lorsque l'on parle d'IA. Et c'est d'ailleurs l'un des attraits du grand public pour

l'IA, celle-ci est sensée être capable de prendre des décisions justes, en dehors de tous biais, d'où la croissance de l'usage de l'IA dans le cadre des gestions des ressources humaines dans certaines entreprises, ou même dans les systèmes judiciaires de certains pays comme la Chine. L'espoir placé dans l'IA est qu'un décisionnaire qui purement rationnel, purement désintéressé, et qui n'est pas limité par les mêmes limites que l'être humain pourrait faire surgir un monde plus juste pour tous. Hélas, le rêve d'une IA sans le moindre biais, juste en tous points, est encore lointain, comme l'a démontré Amazon avec son IA de gestion des Ressources Humaines qui fonctionnait sur base d'un set de données issues des pratiques de recrutement des années précédentes. Cette IA a manifestement découvert une tendance à ce que les hommes soient plus fréquemment recrutés que les femmes, semble en avoir déduit que le fait d'être un homme était un point positif pour être recruté et a donc perpétué le schéma en favorisant le recrutement d'hommes. Doit-on déduire que cette IA était sexiste ? Probablement pas. Par contre, nous observons une fois de plus qu'une intelligence désincarnée, incapable d'empathie pour ces femmes, n'a pas été en mesure de constater le sexisme du set de données qui lui avaient été fournies. Les données étaient sexistes, les conclusions qu'en a tiré une IA étaient tout aussi sexistes. Evidemment, bon nombre de recruteurs en chair et en os, supposés être capables d'empathie, être capables de comprendre ce qu'est le sexisme, continuent pourtant de perpétuer ce sexisme. Il ne s'agit pas ici de dire que l'IA est le problème, néanmoins il semble évident de constater qu'elle n'a pas été la solution, et pire, qu'elle ne pouvait pas l'être. Si l'équipe en charge du recrutement, équipe qui manifestement produisait des résultats sexistes, avait été remplacé non pas par une IA mais par une nouvelle équipe d'êtres humains, ceux-ci auraient eu l'opportunité de mettre le doigt sur le problème et de le résoudre. Rien ne dit qu'ils l'auraient fait, mais ils auraient été bien plus en mesure de le faire que n'aurait pu l'être l'IA qui a été utilisée. Celle-ci manquait d'une capacité d'adaptation et surtout d'une capacité à intégrer des valeurs morales. Cette capacité à intégrer des valeurs morales, l'être humain en dispose, certes pas d'une façon parfaite, mais via notamment sa capacité, en temps qu'être incarné, à souffrir et donc à comprendre la souffrance d'autrui. Faire preuve d'empathie est un défi de taille pour un être qui ne souffre pas.

C'est en réalité le terme intelligence, qui est ici équivoqué. Lorsque l'on parle de l'intelligence humaine, ou de l'IA telle qu'elle pense ou du moins se comporte comme un

humain, on ne parle pas forcément de la même chose que lorsque l'on parle d'une IA purement rationnelle. Le terme IA, intelligence artificielle, n'est d'ailleurs que la traduction du terme anglais « AI » (Artificial intelligence), alors que le terme *intelligence* en anglais n'a pas exactement la même signification que le terme intelligence dans la langue française. Il y a du coup au moins un risque d'amalgame culturel qui peut être fait lorsque l'on parle d'IA ou d'AI, simplement parce que la tradition francophone veut parler d'une intelligence au sens francophone du terme, alors que l'AI désigne une *intelligence* au sens anglophone du terme. Là où le français désigne principalement une capacité, le terme anglais est plus équivoque. *Intelligence* en anglais signifie aussi « renseignement » ou « information » dans le cadre militaire par exemple, ou le cadre de l'espionnage. Ce sens n'apparait pas du tout en français, et dès lors si l'AI se rapproche nettement de l'IA et partage le fantasme d'un robot autonome et intelligent, elle insiste dans la langue anglaise sans doute plus nettement sur le rapport au traitement de données, à l'utilisation de l'information en grande quantité, de manière intelligente, par du croisement, en mettant en évidence des liens, des corrélations.

Ces réflexions trouvent leur écho dans la description de ce qu'est l'AI, en tant que champ de recherche, d'après la description qu'en propose les trois itérations de « *Artificial Intelligence : a Modern Approach* » par Russel et Norvig publiés entre 1995 et 2009. Dans ces ouvrages, les auteurs montrent qu'il y a quatre approches à l'IA, avec deux mesures ayant chacune deux possibilités. Pour eux, d'une part l'IA vise soit l'humain, soit la rationalité, et d'autre part elle cherche à produire un comportement ou bien à produire un raisonnement. On obtient donc quatre angles d'approche de ce que cherche à construire l'IA : soit des systèmes qui agissent comme les humains, soit des systèmes qui agissent rationnellement, d'une part, et d'autre part soit des systèmes qui pensent comme des humains, soit des systèmes qui pensent rationnellement. Ces angles d'approches peuvent être représentés dans le tableau[49] suivant :

[49] BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in *Stanford Encyclopedia of Philosphy*, (2018), p13.

	Human - Based	Ideal - Based
Reasoning - Based	Systems that think like humans.	Systems that think rationally.
Behavior - Based	Systems that act like humans.	Systems that act rationally.

Ces quatre catégories semblent, à l'heure actuelle, recouvrir l'ensemble des approches en ce qui concerne ce que tend à être l'IA. Pour inscrire ces catégories dans le contexte de ce que nous avons vu dans les paragraphes précédents, Alan Turing et le TT s'inscrivent résolument dans la catégorie « Agir comme un humain ». C'est précisément le but du TT que de pouvoir déterminer si une machine agit comme un humain « aussi bien qu'un humain ». Une machine qui passerait le TT serait l'aboutissement de cette catégorie. Norvig et Russel, en revanche, sont de ceux qui défendent l'objectif pour la machine d'agir rationnellement. C'est une approche qui a le mérite de prendre en compte les limites de la machine, notamment son incapacité à être incarnée de la même façon qu'un être vivant, au sens fort, de connaitre la souffrance, la faim, la mort, etc.

Avant de discuter plus précisément lequel de ces deux points de vue a le plus de sens pour notre réflexion, prenons un instant pour discuter l'autre distinction, qui ne différencie pas Turing de Norvig et Russel, à savoir la distinction entre une machine qui agit et une machine qui pense. Nous nous sommes volontairement concentrés sur les approches qui sont relatives à l'action, parce que les approches qui s'intéressent à la pensée rencontrent le même problème que nous avons souligné dans la partie précédente en ce qui concernait l'intention. En effet, la structure même de l'IA peut rendre obscure l'accès à sa pensée. Que ce soit parce que, naturellement, elle ne s'exprime pas dans le même langage que nous, ou plus fondamentalement, parce que la quantité de données qu'elle utilise et retransmet est telle qu'il est impossible d'en distinguer une pensée claire, prétendre avoir accès à la pensée de l'IA ouvre un certain nombre de problèmes. Par ailleurs, l'idée même de dire qu'une IA doit penser comme un humain requiert que nous puissions déterminer de quelle manière un humain pense, et si nous pouvons avoir le sentiment d'être enfermés avec nos pensées toute notre existence, c'est encore un fameux saut en avant que de dire que nous pouvons décrire la nature de ce que sont des pensées

humaines. Enfin, et c'est sans nul doute le point le plus important, ce travail s'inscrit dans une réflexion fondamentalement éthique, et penche donc naturellement du côté de l'action. Par ailleurs, l'approche que nous prenons pour cette réflexion reste de tenter de mettre de côté autant de problèmes que possible, sans rendre trivial le résultat final. Ce n'est évidemment pas qu'une réflexion sur la façon dont l'humain pense, ou la façon dont la machine puisse penser ne soit pas éminemment intéressante, mais c'est simplement qu'une telle réflexion risquerait de noyer notre sujet dans des détours labyrinthiques. Si de très inspirés penseurs ont déjà pu montrer comment de telles réflexions valaient la peine d'être menées à bien, il n'empêche que les conclusions de ces réflexions restent, par leur nature, sujettes à la discussion. En nous contentant de parler de l'action, et non de la pensée, nous tentons de rester dans un domaine relativement objectivable, ou tout du moins toujours plus objectivable que le registre de la pensée.

Cette distinction étant faite, il nous reste à discuter si nous nous dirigeons plutôt vers l'approche défendue par Norvig et Russel, à savoir d'une IA qui agisse rationnellement ou plutôt vers l'approche d'Alan Turing, d'une IA qui agisse comme un humain. Nous faisons face ici à un dilemme. Au premier abord, l'approche rationaliste est séduisante parce qu'elle se veut objective, et parce que, comme nous l'avons déjà précisé, elle respecte les limites fondamentales d'une IA qui n'est pas un être vivant incarné. Nous n'avons aucun mal à penser qu'une IA soit capable de prouesses logiques et mathématiques, alors qu'imaginer une IA agir tel un humain, en fonction d'émotions, de ressentis, parait plus complexe. En réalité, il semble bien que Norvig et Russel défendent un point de vue plus accessible, moins ambitieux, et potentiellement plus pertinent qu'une copie mécanique de l'humain. Mais, quelque soit l'accessibilité de ces points de vue, leur approche présente au moins un défaut, le rapport à la moralité, aux valeurs. Pour qu'une telle machine y ait accès, il faudrait alors que l'éthique soit rationalisable, et bien que Leibniz ait rêvé un calcul moral purement objectif, un tel idéal est loin d'être atteint.

> When controversies arise, there will be no more need for a disputation between two philosophers than there would be between two accountants [computistas]. It would be enough for them to pick

> up their pens and sit at their abacuses, and say to each other (perhaps having summoned a mutual friend): 'Let us calculate.'[50]

Plus nettement même, ce rêve est considéré par beaucoup comme n'étant tout simplement pas possible. Nous faisons ici référence aux défenseurs d'un certain relativisme moral, comme David Wong notamment. Ceux-ci considèrent qu'une éthique universalisable n'est tout simplement pas possible, que les points de vue, qu'ils soient culturels ou religieux, varient trop fortement que pour être réconciliés. Quelle que soit la conviction que l'on ait à ce sujet, il faut au moins reconnaitre que l'idée même de valeurs morales universelles parait encore plus lointaine que l'idée d'un agent moral artificiel. Cet agent moral artificiel devra, au moins temporairement rendre des comptes à la moralité d'une communauté locale, au sens large du terme. En effet, comme nous l'avons déjà préfacé, les valeurs morales humaines sont profondément inscrites dans notre façon de nous incarner, en chair et en os, dans une vie finie et fragile, empreinte de joies et de tristesses, de désirs et de souffrances. Ce n'est que parce que nous rions et nous pleurons, parce que nous éprouvons de l'empathie, parce que nous souffrons avec ceux qui souffrent, parce que nous pouvons au moins partiellement envisager la souffrance d'autrui, que nous sommes capables, en communauté, de générer un code moral, un ensemble de valeurs. Ces valeurs, on ne peut s'attendre qu'elles surgissent naturellement d'une machine qui ne peut pas, de par son incarnation, partager notre expérience de vivants. Bien que tout ceci nous pousse vers l'objectif d'une IA plus humaine dès lors qu'on attend d'elle qu'elle agisse moralement, il faudra prendre en compte les limitations qui sont propres à l'IA et faire face au défi de tenter d'envisager une IA qui puisse dépasser ces limites sans pour autant être mortelle, être incarnée et souffrir de la même façon que l'humain souffre. En bref, l'IA est par essence moins humaine que l'humain, plus rationaliste que l'humain, mais il nous faudra tenter de dépasser cette problématique si l'on veut prétendre à ce qu'elle agisse moralement de manière idéale.

Description et Fonctionnement de l'IA

L'ensemble de ce qui est appelé IA est assez vaste, et les prérequis pour pouvoir être décrit comme une IA sont relativement faibles. Pour pouvoir parler d'agent intelligent

[50] BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in *Stanford Encyclopedia of Philosphy*, (2018), p59.

artificiel, il suffit d'avoir un agent artificiel qui est capable d'obtenir des perceptions d'un environnement à partir de percepteurs, et après un procédé de délibération (aussi simple ou complexe soit-il), d'agir sur cet environnement via des « actionneurs ». Le terme « actionneur » désigne ici tout moyen par lequel l'agent est capable d'avoir une influence sur l'environnement initial, qu'il s'agisse de l'application d'une fonction mathématique, ou la sélection de données, pour des cas virtuels, ou très concrètement la stabilisation automatique du vol d'un avion, ou, pour prendre un exemple bien plus simpliste, l'allumage automatique d'une lampe munie d'un détecteur de mouvement. Aussi trivial que cela paraisse, il s'agit bien d'un agent intelligent artificiel, qui perçoit le mouvement à l'aide de ses percepteurs, détermine si le mouvement est suffisant pour justifier l'allumage de la lampe, et allume la lampe le cas échéant. Nous sommes ici évidemment très loin de parler d'un agent moral. Cette agence artificielle au sens le plus basique, notre vie en est déjà envahie dans presque tous ses aspects, que ce soit tout ce qui concerne l'informatique, la bureautique que nous utilisons quotidiennement, que ce soient les bien-nommés smartphones qui accompagnent nos journées, ou même plus simplement encore les thermostats, qui mesurent la température et adaptent le chauffage en fonction. Tout ces objets qui entourent nos vies sont autant d'agents artificiels intelligents. Néanmoins, les plus simples d'entre ceux-ci sont bien différents des plus complexes. Les plus simples d'entre eux sont ce que l'on nomme des agents réflexes simples. Un agent réflexe est tout simplement un agent artificiel qui, comme nous l'avons décrit et sans plus de fioritures, capte son environnement, obtient une « image » de « comment l'environnement est actuellement », vérifie si cette « image » rencontre les conditions qui régissent son action, et agit le cas échéant. Comprenons bien que le terme « image » est ici utilisé quels que soient les types de percepteurs utilisés par l'agent, pour désigner les données perçues par les percepteurs de l'agent et qui lui permettent d'obtenir une information sur son environnement, comme la température pour le thermostat.

Un agent réflexe un peu plus sophistiqué que la lampe munie d'un détecteur de mouvement serait un agent capable de se faire une idée, non pas juste du monde au moment de la perception, mais de la façon dont le monde évolue sur une certaine durée, et de la manière dont ses actions l'impactent. Nous avions donné l'exemple des IA capables de jouer aux échecs, c'est de ce type d'agent qu'il s'agit. En effet, aucun premier coup ne permet de gagner directement la partie, et l'IA doit être capable de comprendre

son environnement et la manière dont chaque coup qu'elle joue l'impacte pour être en mesure de planifier une stratégie gagnante. Mais ces IA travaillent dans un monde dans lequel elles maitrisent suffisamment la suite des états possible que pour les envisager. Evidemment, une IA ne peut pas connaitre tous les coups possibles sur une suite d'un grand nombre de coups, parce que ce nombre devient exponentiellement hors de contrôle, mais le futur proche est suffisamment certain pour cette IA. Par ailleurs, l'IA sait au moins que la tour ne se déplacera pas en diagonale, ni le fou à l'horizontal. D'autres environnements sont bien plus complexes et changeants, nous avions évoqué l'environnement dans lequel le pilote automatique d'un avion opère, par exemple. Pour ces IA, il s'agit d'être capable de s'adapter au changement et à l'incertitude, et pour se faire, ces IA mobilisent le calcul de probabilités afin de prendre en compte les différentes possibilités derrière leurs incertitudes. Cet élément est particulièrement pertinent au vu des conclusions de la partie précédente, nous le réexplorerons plus précisément dans la troisième partie. Enfin, et il va sans dire que c'est un élément majeur, les IA parmi les plus complexes disposent encore d'un avantage de taille, la capacité d'apprendre. Cette capacité est atteinte en mobilisant un set de fonctions parallèles à la délibération, qui évalue l'action par rapport à un standard de performances et génère des changements potentiels. La façon précise dont ceci se produit varie selon l'architecture propre de l'IA. Nous parlerons plus précisément de « Machine Learning » dans la suite de ce chapitre.

Prenons un instant pour revenir sur une IA relativement simple comme celle qui est programmée pour déterminer le meilleur coup possible dans une partie d'échec. Cet IA fonctionne, comme nous l'avons déjà évoqué extensivement, dans un environnement très simple, en comparaison tout du moins à d'autres environnements plus complexes, et le nombre de possibilités d'une partie d'échec, s'il est particulièrement grand, reste fini. D'ailleurs, Claude Shannon (voir source) a estimé dès 1950 le nombre de parties d'échecs ayant un « sens échiquéen » à 10^{120}. Cette estimation a été affinée depuis, mais l'idée générale persiste. Si ce nombre est fini, il reste très élevé. Il est supérieur aux estimations concernant le nombre d'atomes dans l'univers, qui seraient de 10^{80}. La grandeur de ce nombre a une conséquence directe sur la capacité d'une machine à effectivement jouer le meilleur coup possible. En théorie, le raisonnement appliqué par la machine devrait être capable de calculer tous les coups possibles afin d'isoler le coup qui réduit au maximum les chances d'être contrecarré par l'adversaire, afin de maximiser les chances de victoires.

Néanmoins, un tel procédé n'est généralement pas possible. A moins que le nombre de pièces sur le plateau se soit suffisamment réduit que pour rendre la durée de l'analyse suffisamment courte, l'IA prendrait un temps bien trop long pour calculer tous les coups possibles. Pour donner la réponse parfaite, l'IA devrait opérer de manière infiniment rapide, ce qui est évidemment impossible. Dès lors, les IA qui sont programmées pour ces tâches sont pensées en fonction de cette contrainte et disposent d'un certain nombre de moyens pour réduire la durée de réflexion, et proposer des réponses temporaires, tout en continuant de les affiner. La première méthode consiste évidemment à réduire la profondeur (« depth » en anglais) de la réflexion, c'est-à-dire d'évaluer le coup non pas relativement à la totalité des coups à suivre, ce que nous avons montré être impossible dans un laps de temps réaliste, mais en fonction de la qualité de la position obtenue seulement quelques coups après. Evidemment une profondeur trop faible donne un résultat trivial, étant donné qu'il est bien connu qu'il faut avoir au moins un coup d'avance aux échecs, et que si je capture un cheval avec ma reine en mettant en danger ma reine, bien qu'après mon coup, j'ai un cheval d'avance, la perte de ma reine est évidemment souvent plus problématique. Mais dès que la profondeur dépasse quelques coups, ce qui peut se faire dans un laps de temps très raisonnable, le coup préconisé devient évidemment bien plus pertinent. Les moteurs modernes atteignent une prévision d'une quinzaine de coups en quelques secondes, alors qu'un tel calcul prendrait évidemment bien plus de temps à n'importe quel expert des échecs.

Cette procédure pose néanmoins un problème significatif, à savoir que la machine doit disposer d'un moyen d'évaluer la partie au fur et à mesure qu'elle se déroule afin de déterminer quelle situation doit être favorisée à quelle autre. En effet, en dehors de l'échec et mat ou d'une situation qui permet une séquence de coups forçant un échec et mat, la victoire que vise le moteur et qui est évidemment l'objectif final ne peut pas être envisagée à suffisamment court terme dans les premiers instants de la plupart des parties. Dès lors, il y a deux mécanismes qui sont mobilisables pour évaluer une position. Le premier mécanisme consiste à tenter d'évaluer directement la position. Pour ce faire, une première solution consiste à compter les valeurs des pièces. Ainsi à chaque pièce est attribuée une valeur, qui est évidemment indicative et relative. Dans ce système de cotation, en dehors de tout contexte, chaque pion vaut 1 point, chaque pièce mineure (cavalier ou fou) vaut 3 points, les tours valent 5 points et la reine vaut 9 points. Evidemment, le Roi n'a pas de

valeur attribuée vu que sa valeur est théoriquement infinie. Ce système, bien que relativement arbitraire, décrit suffisamment bien le rapport de force entre les différentes pièces que pour être utilisable. Cependant, chaque pièce peut gagner ou perdre en valeur selon le contexte de la partie, un pion sur le point d'être promu en reine qui force une tour à être sacrifiée pour le stopper vaut bien plus d'un point, alors qu'un cavalier dont tous les mouvements possibles seraient obstrués vaut sans doute moins de trois points. Au contraire, si ce même pion ne peut être défendu et sera facilement perdu parce qu'il s'est aventuré trop loin en territoire ennemi, il peut valoir moins d'un point, voire n'avoir presque aucune valeur. Tout ces détails afin d'expliquer le type de raisonnement complexe qu'une machine doit effectuer pour tenter d'évaluer une position, quand aucun mat clair n'apparait encore. L'autre méthode qui peut être mobilisée pour évaluer une position est beaucoup plus rapide, mais relativement moins fiable. La machine peut avoir accès à une base de données de parties jouées et, pour peu qu'une partie précédemment jouée ait atteint le même stade, s'intéresser au résultat de ces parties. Si les parties comprises dans la base de données sont des parties de qualités, ne prenant pas en compte celles où un joueur a commis une erreur évidente, un tel résultat peut être assez indicatif de la qualité d'une position. Cette seconde méthode est particulièrement pertinente dans le début de la partie, quand les positions atteintes ont quasi systématiquement déjà été rencontrées, et quand les projections futures ne laissent pas entrevoir de gain ou de perte de matériel pour l'un ou l'autre des joueurs. Ainsi, les IA peuvent produire leur propre base de données, en expérimentant certains coups dans certaines positions, afin de tenter de voir dans quel sens un tel coup fera pencher la balance de victoire sur un très grand nombre de parties. Ce genre d'expérimentation et de construction de base de données correspond au calcul des coups futurs sur une profondeur maximale, mais le fait de telle sorte à ce que des informations pertinentes en soit retirées et soient conservées à l'avenir. Evidemment, l'IA n'est pas en mesure de jouer toutes les variantes possibles, et doit se contenter de jouer celles qui semblent les plus favorables selon la première méthode d'évaluation du score. Ces deux méthodes permettent logiquement de garantir à tout moment de la partie une évaluation de la position permettant à la machine de choisir la position qui lui est la plus favorable. Si la seconde méthode est nécessaire, c'est parce que toutes les positions ne peuvent pas être connues dans un laps de temps réaliste par l'IA. Si nous prenons le jeu du « Tic-Tac-Toe », le nombre de parties possibles est bien plus faible, il est même inférieur à 362 880 (9!) vu que ce nombre représente tous les coups possibles à chaque

tour, et que certains de ces coups ne seront jamais joués si la partie est déjà terminée. Bien que 9! puisse paraitre élevé, une machine moderne est capable de calculer toutes les parties possibles en un temps raisonnable, et d'ailleurs même un humain est capable de jouer parfaitement au Tic-Tac-Toe, tellement le jeu est simple. Un tel jeu ne requiert donc rien d'autre qu'un chemin direct ouvrant le plus de possibilités de victoires, et assurant de ne pas perdre, ce qui est possible si bien que si les deux joueurs jouent parfaitement, toute partie se finit par une égalité. L'ordre de grandeur du nombre de parties aux échecs est incommensurablement plus grand, si bien qu'il est impossible d'espérer faire le tour de l'arbre des possibilités. C'est pourquoi des alternatives sont nécessaires.

Cette complexité, on peut la retrouver dans bien d'autres domaines que dans celui des échecs, et de manière bien plus significative encore, si bien que l'IA moderne est capable de jouer bien mieux que l'humain aux échecs, et pourtant se retrouve bien incapable de comprendre correctement la langue française (ou une autre langue, comme l'anglais, en l'occurrence). Si Deep Blue, qui a battu Garry Kasparov aux échecs, avait reçu les instructions du jeu d'échec écrites en français, elle aurait été bien incapable d'en faire quoi que ce soit. Or pourtant, la lecture est l'une des manières les plus efficaces de transmettre la connaissance, l'information. Certes beaucoup d'informations peuvent être transmises sous forme de données, mais il n'en est rien des raisonnements complexes, des réflexions morales, par exemple. Attention, une IA peut être capable de confirmer qu'une langue est écrite ou même parlée correctement. Elle peut retenir les règles d'orthographe et de grammaire, et même apprendre ce qui différencie une écriture élégante d'une écriture maladroite. Alors même que ces lignes sont écrites sur un programme de traitement de texte aussi classique que Word, un certain nombre de corrections et même de suggestions y ont été implémentées directement grâce au correcteur orthographique. Mais il y a une nuance énorme entre la maitrise des règles de la langue et la compréhension de son contenu. Il s'agit de faire face à la difficulté de transmettre le sens des mots, sans disposer pour autant d'un autre référant sur lequel s'appuyer. Tenter d'apprendre une langue humaine à l'IA, c'est presque comme tenter d'entretenir une communication épistolaire avec une personne avec qui on ne partage pas le moindre mot de langage. Imaginez qu'un peuple rentre en contact avec vous, qu'ils soient capables d'écrire mais que vous n'ayez pas un mot de langage en commun. Le seul médium restant serait sans doute de passer par des dessins, mais si vous êtes contraints de communiquer avec l'alphabet classique de

votre langue, la cause semble perdue d'avance. C'est cette montagne qu'affronte la recherche en IA lorsqu'elle tente de faire comprendre le sens des mots à une IA. Ce problème complique particulièrement la transmission d'informations complexes de l'humain à l'IA, alors que le modèle principal de stockage de la connaissance dont nous disposons est et reste l'écriture. Tant que nous ne pouvons apprendre à une IA la lecture, c'est autant de possibilités de développer son intelligence qui ne sont pas à sa disposition. L'autre façon dont bons nombres de connaissances nous sont accessibles, c'est via l'expérience personnelle. Pour prendre un exemple caricatural, si l'enfant comprend ce que cela veut dire que de dire que l'eau mouille, c'est généralement plus par sa propre expérience, par son observation, que par une leçon donnée par un adulte. Or, si l'IA peut disposer de capteurs capables d'extraire des données d'un environnement, elle n'est pas du tout capable de comprendre ce que sont ces données dans l'absolu, et elle n'est pas capable de sortir du cadre très précis que ses capteurs prévoient. Si elle dispose d'un anémomètre et qu'elle utilise la vitesse du vent pour adapter la puissance de moteurs dans un avion afin de maintenir l'appareil dans une position stable, elle peut bien disposer des appareils de mesure les plus précis, et des moyens de calculs lui permettant une finesse que l'humain ne peut égaler pour ajuster les moteurs, elle n'a aucune idée de ce que le vent peut bien être, en soi. Et quel que soit le nombre d'heures de vol, elle ne comprendra jamais plus ce que le vent est. Un enfant, incapable de calculer la vitesse du vent, passe quelques instants dehors, ressent le vent, et comprend petit à petit ce dont il s'agit. En grandissant, il lira peut-être sur le sujet, découvrira ce qu'est la pression atmosphérique ainsi que d'autres phénomènes météorologiques produisant le vent au sujet desquels certaines autres personnes ont écrit, et il acquerra de cette façon une partie au moins du savoir dont l'humanité dispose en ce qui concerne le vent. Toutes ces connaissances, bien que peut-être simples du point de vue du météorologue, sont absolument inaccessibles à l'IA que ce même météorologue utilise pourtant pour développer des modèles de prédictions. Ces différentes difficultés nous montrent nettement comment, bien que l'IA soit capable de prouesses dans de nombreux domaines, elle reste bien souvent cantonnée à la résolution de problèmes déterminés, et nous sommes encore bien incapables de lui permettre d'apprendre certaines connaissances complexes que pourtant nous parvenons sans mal à apprendre à un jeune enfant.

Approche descendante et approche ascendante

Nous l'avons déjà évoqué, il existe des IA avec des architectures très différentes. Certaines IA prennent des approches diamétralement opposées à d'autres IA pour atteindre les mêmes résultats. Ceci vient du fait qu'il y a deux types d'approches en ce qui concerne le fonctionnement de l'IA. La première approche, l'approche descendante (top-down), est aussi celle de la programmation traditionnelle. Son fonctionnement consiste, pour le programmeur, à encoder directement la façon dont il veut que l'IA se comporte, en prenant en compte, autant que faire se peut, toutes les situations qu'une telle machine pourrait rencontrer. Ce type de programmation est évidemment très efficace en ce qui concerne les situations prévues, et particulièrement prévisibles, vu que chacune de ces réactions sont prévues par la programmation. La question peut d'ailleurs se poser de savoir si le terme « intelligence » est aussi adapté que dans le cas de l'autre type d'approche. En effet, ce type de programme ne fait qu'opérer les actions pour lesquelles il a été programmé, et à ce sens ne diffère que très peu de la meule de pierre du moulin qui tourne pour broyer le grain à mesure que la roue du moulin, qu'il soit hydraulique ou éolien, la fait tourner. Certes la machine peut paraitre plus intelligente à l'utilisateur parce qu'elle effectue des actions complexes et que, pour l'utilisateur lambda, les raisons pour lesquelles elle les effectue semblent complexes, mais en réalité la seule intelligence en présence est sans doute plus l'intelligence du programmeur, qui a encodé tous les comportements dans la machine. Le problème de cette approche très fiable en ce qui concerne les situations que le programmeur a prévues, ce sont évidemment les situations qui n'ont pas été prévues. Ainsi, un aspirateur automatique programmé par un programmeur peu prévoyant pourrait réagir d'une manière problématique en la présence d'un animal domestique. Un programmeur prévoyant aura imaginé une solution permettant d'éviter un inconvénient majeur, mais il sait que sa machine est bien incapable de trouver une solution par elle-même, et plus incapable encore de comprendre ce qu'est un animal. Cette représentation que nous proposons aussi est légèrement caricaturale, mais il n'empêche que ce sont les grandes lignes du fonctionnement de l'approche descendante.

De l'autre côté du spectre des IA se trouvent les IA construites avec une approche ascendante (bottom-up). De ce type d'IA, nous allons discuter uniquement ici de celles

qui mobilisent des « réseaux de neurones ». Comme le nom l'indique, les « réseaux de neurones » fonctionnent d'une manière similaire aux neurones que l'on retrouve chez tous les organismes biologiques intelligents. La recherche en matière d'IA n'échappe en effet pas à l'efficacité de la bio-ingénierie, au moins en tant que source d'inspiration. Si la transmission d'information et la délibération se fait chez les animaux à l'aide de réseaux de neurones, c'est bien que c'est une méthode qui fonctionne, et il est logique de tenter de l'exploiter dans le cadre de l'IA. Bien entendu, c'est d'autant plus attendu qu'une partie des chercheurs en IA ont pour objectif de reproduire l'intelligence humaine. Quoi de plus normal dès lors de s'inspirer de son fonctionnement. Les réseaux de neurones sont principalement utilisés pour leur capacité d'apprentissage dans un système de répétition de l'expérience. Le principe de ces réseaux est de mobiliser un grand nombre de nœuds, qui représentent des neurones, et qui sont reliés par des liens représentant des dendrites. Chacune de ces dendrites se voit attribué un « poids » numérique. Ce poids permettra de déterminer quel lien doit être favorisé. Dans un système d'expérience itérative, les poids initiaux sont aléatoires et sont modifiés entre chaque itération en fonction des résultats obtenus, afin qu'à chaque itération, le poids de chaque lien se trouve plus proche de l'optimum permettant de réaliser la tâche attendue de la meilleure manière. Ceci se fait de manière supervisée, en donnant des objectifs à l'IA et plus ou moins de renforcement positif lui permettant de déterminer si elle est sur la bonne voie. Les validations doivent cependant être suffisamment éloignées pour que l'IA puisse déterminer elle-même la meilleure manière d'aller d'un renforcement à un autre. Ce système de réseau de neurones brille particulièrement dans les situations où les approches traditionnelles, basées sur la logique, ne semblent pas porter de fruit, ou bien tout simplement ne sont pas réalistes au vu de la complexité de la tâche. Deux exemples de modèles ou ces IA brillent sont d'une part les jeux vidéo de type RTS (Real-Time Strategy) qui sont un environnement particulièrement intéressant pour l'expérimentation sur ce genre d'IA, et d'autre part, des problèmes comme la reconnaissance de l'écriture à la main. Ces problèmes ont pour particularité de ne pas pouvoir être formalisés clairement et pourtant de représenter des tâches réalisables pour le cerveau humain (moyennant suffisamment d'entrainement, évidemment).

Pour souligner la similitude de fonctionnement avec le cerveau animal, citons rapidement un article publié le 18 novembre 2015 dans Plos One par R. M. Levenson et

al. intitulé « Pigeons (Columba livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images ». Cet article présente les résultats d'une étude effectuée sur une cohorte de pigeons dont le but était d'évaluer leur capacité à distinguer des images de radiologie présentant une tumeur maligne d'une tumeur bégnine dans le sein. Les pigeons sont exposés à de telles images sur un écran, entouré d'une bande de couleur jaune d'un coté et d'une bande de couleur bleu de l'autre, chacune de ces couleurs représentant « tumeur bégnine » ou « tumeur maligne ». Si le pigeon tapote du bec sur la bonne couleur, il reçoit automatiquement une récompense. Après une période d'entrainement et

un temps d'adaptation au système, le groupe de pigeons est parvenu à d'étonnamment bons résultats sur des images auxquels ils avaient déjà été exposés. Evidemment, il s'agissait de déterminer s'ils se basaient sur leur mémorisation des images, ou bien s'ils étaient capables de distinguer des éléments leur permettant de déterminer si la tumeur était maligne ou bégnine.

> While the birds readily succeeded with this classification task, it was crucial to determine whether they were relying on just their ability to memorize image classification status, or whether they had managed to detected feature-based cues that allowed them to accurately classify previously unseen, novel, images. Accordingly, during a 5-day period after the end of training at each magnification level, pigeons were given a small number of novel benign and malignant breast tissue images intermixed with the full set of familiar training images. Fig 7 shows indeed that the birds had gained the ability to accurately classify novel as well as familiar benign and malignant images, and with equal accuracy, averaging 87% and 85% correct on familiar and novel examples, respectively, a non-significant difference. In each case, accuracy exceeded that expected by chance, one-tailed binomial tests, $p < 0.001$. The pattern of responding to the familiar and novel slides over the next 4 days of non-differentially reinforced testing was similar to that seen on Day 1, averaging 85% and 83% correct, respectively. Moreover, the birds demonstrated equal learning prowess at all magnifications tested.[51]

Il est intéressant de remarquer comment l'entrainement dont ces pigeons ont été les sujets est très similaire aux méthodes utilisées pour obtenir le même genre de résultats d'une IA.

[51] LEVENSON, R.M., et al., « Pigeons (Columba livia Columba livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images » in *Plos*, (08/2015), p7.

C’est bien évidemment parce que ces méthodes peuvent être modélisées, idéalisées et servir d’objectif pour la construction du système même du réseau de neurones. Ce n’est pas une heureuse coïncidence si le fonctionnement est similaire, c’est tout simplement le résultat direct du processus de mise en forme de ces réseaux de neurones.

Evidemment, au-delà de l’opposition théorique entre ces deux paradigmes, à savoir IA ascendante et descendante, c’est le système qui mobilise simultanément ces différents paradigmes qui produit les meilleurs résultats. Ce n’est pas une évidence, étant donné que ces deux systèmes sont diamétralement opposés, et sont les fruits d’approches et de compréhension de ce qu’est l’IA qui sont tout aussi opposées. L’approche descendante se veut purement logique, alors que l’approche ascendante présente une dimension organique, et suppose qu’un objectif complexe peut être atteint par des intermédiaires relativement simples. Cette idée se fonde sur une certaine théorie de l’émergence, le travail de chaque neurone n’ayant pas de sens apparent mais celui de l’ensemble du réseau se mettant à en avoir. Cela étant dit, ces deux méthodes sont en fait plus compatibles qu’elles n’y paraissent. Si nous reprenons l’exemple des échecs, ou de tout autre jeu de plateau qui a attiré l’œil de la recherche en IA, les meilleurs IA mobilisent simultanément les deux méthodes. Dans un premier temps, le système le plus efficace consiste à mobiliser le « Monte Carlo tree search » (MCTS), à savoir un système qui simule automatiquement un grand nombre de mouvements possibles en jouant leurs suites de manière aléatoire et en voyant le résultat. Mais ce système seul est loin d’être le plus efficace. A celui-ci sont associés différents réseaux de neurones qui permettent d’évaluer, comme nous l’avions expliqué, l’état de la partie, mais aussi de simuler de manière plus précise que l’aléatoire la façon de jouer de joueurs expérimentés, afin de rendre les résultats plus pertinents.

> Google DeepMind’s AlphaGo is another example of a multi-paradigm system, although in a much narrower form than Watson. The central algorithmic problem in games such as Go or Chess is to search through a vast sequence of valid moves. For most non-trivial games, this is not feasible to do so exhaustively. The Monte Carlo tree search (MCTS) algorithm gets around this obstacle by searching through an enormous space of valid moves in a statistical fashion (Browne et al. 2012).

> While MCTS is the central algorithm in AlpaGo[52], there are two neural networks which help evaluate states in the game and help model how expert opponents play (Silver et al. 2016). It should be noted that MCTS is behind almost all the winning submissions in general game playing (Finnsson 2012).[53]

Par ailleurs, au-delà de l'application à des jeux de plateau, l'IA s'est rendue indispensable dans certains domaines de la recherche scientifique pour sa capacité notamment à analyser de grandes quantités de données. Dans ce domaine, c'est en particulier le développement du champ du « machine learning » qui brille en la matière. Ce champ en question peut, de manière générale, être subdivisé en trois parties. D'une part, on retrouve l'apprentissage supervisé (Supervised Learning), dans lequel la tâche à accomplir est strictement définie et la fonction permettant d'accomplir la tâche est donnée pour un domaine précis. Par ailleurs, d'après Gringsjord et Govindarajulu, ce type de supervision domine actuellement le machine learning : « Supervised learning dominates the field of machine learning and has been used in almost all practical applications mentioned just above »[54]. Le second type de machine learning est naturellement l'apprentissage non-supervisé. Comme son nom l'indique, la machine est ici laissée à elle-même, sans avoir d'objectif clair et encore moins de façon d'atteindre quelque objectif. Au contraire, il s'agit d'exposer la machine à des données afin de faire ressortir des motifs intéressants, des comportements particuliers de l'échantillon de données qui peuvent être pertinents à être isolés. Enfin, le troisième type de machine learning consiste en apprentissage avec renforcement, où la machine ne reçoit pas de consigne mais interagit et perçoit son environnement et reçoit occasionnellement un signal de validation ou d'invalidation de ce qu'elle effectue. La seule consigne dont dispose la machine, c'est qu'elle doit maximiser ces validations. Ce type de machine learning est précisément celui qui ressemble le plus à la façon dont des pigeons ont été entrainés à reconnaitre les tumeurs bégnines des tumeurs malignes dans le chapitre précédent. L'usage de ces méthodes est très varié et un certain nombre d'applications

---

[52] Il semble qu'une coquille se soit glissée dans la publication, a priori l'auteur doit vouloir dire « AlphaGo », mais la citation est reproduite ici en correspondant exactement à sa source.
[53] BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in *Stanford Encyclopedia of Philosphy*, (2018), p40.
[54] Idem, p43.

permettent d'obtenir des résultats dans des domaines pour lesquels on ne pourrait espérer y parvenir sans machine learning.

> In addition to being used in domains that are traditionally the ken of AI, machine-learning algorithms have also been used in all stages of the scientific process. For example, machine-learning techniques are now routinely applied to analyze large volumes of data generated from particle accelerators. CERN, for instance, generates a petabyte (bytes) per second, and statistical algorithms that have their origins in AI are used to filter and analyze this data. Particle accelerators are used in fundamental experimental research in physics to probe the structure of our physical universe. They work by colliding larger particles together to create much finer particles. Not all such events are fruitful. Machine-learning methods have been used to select events which are then analyzed further (Whiteson & Whiteson 2009 and Baldi et al. 2014)[55]

« Strong » AI vs « Weak » AI

Il est difficile de parler d'intelligence artificielle dans le cadre d'un mémoire en philosophie sans approcher au moins la question du face à face entre les tenants de l'IA « forte » et ceux de l'IA « faible ». Pour simplifier la question, défendre la possibilité de l'existence de l'IA faible revient à défendre qu'il peut exister une machine qui agit comme un humain, c'est-à-dire passer le Total Turing Test (TTT), à savoir une version du TT qui ne se contente pas de l'indistinguabilité linguistique mais y ajoute tous les autres comportements humains. Défendre la possibilité de l'IA forte va plus loin que cela et va jusqu'à prétendre à la possibilité de créer de véritables personnes artificielles, disposant de toutes les capacités mentales des personnes normales, en ce y compris la conscience phénoménale. De toute évidence, cette seconde option est bien plus ambitieuse, et sans surprise, un certain nombre de personnes considèrent que l'IA forte ne peut être atteinte, seule l'IA faible le peut. Nous allons décrire l'argument principal contre l'IA forte, parce qu'il met en lumière certaines réserves que nous avions évoquées dans les propos introductifs, avec lesquels, en l'absence de certitude sur le sujet, nous devons fonctionner. Cet argument a été formulé par John Searle en 1980 et s'intitule « Chinese Room Argument » (CRA). Le CRA met en scène Searle lui-même, dans une pièce en

[55] BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in *Stanford Encyclopedia of Philosphy*, (2018), p44.

communication écrite avec l'extérieur avec des personnes communiquant dans une langue chinoise alors que Searle lui-même ne comprend pas un mot de cette langue. Il démontre que s'il recevait des communications écrites de l'extérieur, dans cette langue inconnue, et qu'il avait à sa disposition une boite de réponses à transmettre, en suivant un set de règles qu'il connait, lui permettant de répondre de manière appropriée aux communications extérieures, cela ne démontrerait en rien qu'il soit lui-même capable de comprendre la conversation. L'idée de l'argument est de dire que bien que Searle dispose de toutes les capacités dont peut disposer une IA, il ne comprend malgré tout toujours pas les communications humaines, en revanche, il est capable de produire un output adapté à l'input qu'il a reçu. Cet argument subsiste encore à l'heure actuelle, et la question n'a toujours pas été tranchée. Il n'est pas possible de dire aujourd'hui si une IA pourra effectivement un jour comprendre ce qu'elle communique, ou si elle devra se contenter de faire semblant de comprendre. Cependant, nous avions évoqué une problématique similaire en introduction et avions souligné que, bien que rien ne nous permette d'imaginer que ça ne soit pas le cas, nous ne pouvons pas disposer de la certitude que les autres personnes expérimentent leur conscience et comprennent le sens de leurs mots de la même manière que nous avons la conviction de le faire. En d'autres mots, bien que rien ne semble l'indiquer, rien ne m'empêche de croire que je suis la seule véritable personne et que tous les autres sont aussi conscients que des machines et se contentent de faire semblant. Ce n'est pas du tout une position que nous allons défendre ici, ce pourquoi nous avons souligné que rien n'indique que ce soit une piste intéressante à suivre, mais il s'agit de remettre en perspective les questions que nous pouvons nous poser en ce qui concerne l'IA par rapport à ce que nous savons des autres personnes. Certes nous savons ce que nous ressentons, ce que nous expérimentons, ce que nous comprenons, mais nous ne savons rien de ce que l'autre ressent, expérimente, comprend, et il parait pertinent de se demander s'il y a dès lors le moindre sens à prétendre vouloir savoir tout cela au sujet d'une machine, d'une potentielle personne artificielle si nous ne le savons déjà pas au sujet des personnes les plus naturelles à qui nous n'avons pourtant aucun mal à reconnaitre le statut de personnes. Toute prise de position sur ce sujet a néanmoins d'importantes conséquences lorsque l'on s'intéresse à la question de l'agence morale, comme nous allons le faire dans la partie suivante, et nous verrons comment nous pouvons approcher ce questionnement.

Troisième partie

Propos introductif

Dans la troisième partie de ce travail, nous allons, comme prévu, joindre les deux mondes que nous avons explorés jusque-ici. D'un côté, nous avons l'agence morale et les diverses conceptions éthiques que nous avons abordées dans la première partie de ce travail, et d'un autre côté nous avons l'IA et les agents artificiels. Pour faire le lien entre ces deux mondes, nous allons partir d'abord de l'article de John P. Sullins paru en 2006 dans la « International Review of Information Ethics » et intitulé « When a robot is a moral agent ? ». Dans cet article, Sullins présente une approche assez minimaliste des prérequis pour pouvoir attribuer l'agence morale à un agent artificiel, et il met en évidence trois conditions que sont l'autonomie, l'intentionnalité et la responsabilité. La mention de l'intentionnalité peut ici faire craindre qu'elle soit en désaccord avec le problème de l'intentionnalité que nous avons soulevé dans la première partie de ce travail, mais c'est en réalité tout le contraire. Sullins partage le point de vue que nous avons défendu selon lequel l'intentionnalité est problématique parce que son usage requiert un accès à certaines informations dont nous ne disposons pas. Ce dont il parle quand il mentionne l'intentionnalité, c'est de l'apparence de l'intentionnalité. En effet, comme nous allons le voir, Sullins propose des conditions minimales, et dès lors, ces conditions doivent être valides pour une conception faible de l'IA. Les deux autres conditions que sont l'autonomie et la responsabilité sont traités avec la même méthode. Il s'agit moins d'attendre de l'agent artificiel qu'il soit véritablement autonome au sens le plus essentiel du terme que de montrer qu'il présente l'apparence d'une certaine autonomie. Une telle approche est certainement décevante pour tout qui s'attendrait à des réflexions prospectives dignes de la science-fiction, mais elle est beaucoup plus pertinente dans notre domaine, et plus fertile à une réflexion logique et aussi objective que possible sur le sujet qui est le nôtre.

Dans un second temps, nous allons reprendre les quatre courants éthiques que nous avons évoqués dans la première partie[56] et tenter de montrer comment ces théories éthiques pourraient être mobilisées pour permettre à l'IA de devenir un agent moral plus adéquat. Chacune de ces théories montrera bien entendu plus ou moins de limites et

[56] Pour rappel il s'agit du déontologisme, de l'utilitarisme, de l'éthique de la vertu et de l'éthique du care.

ouvrira plus ou moins de problématiques. Gardons à l'esprit que l'idée de quelque chose comme une vérité morale universelle est loin d'être accessible à l'humain, et qu'il est donc déraisonnable d'en attendre autant d'un agent artificiel. Chaque théorie éthique fait face, de manière parfois très différente, à des dilemmes particuliers qui posent plus ou moins de problèmes, et nous allons tenter de montrer comment l'IA pourrait mobiliser des éléments de ces différentes théories. Naturellement, comme c'est le cas pour l'agent humain, un agent ne doit pas forcément se contenter d'une seule approche éthique d'un dilemme, et gagne généralement à mobiliser différents points de vue pour enrichir sa réflexion.

Dans un troisième temps, nous allons reprendre un certain nombre de concepts que nous avons introduits dans la première partie de ce travail, comme l'approche « forward-looking » ou la question de la « quality of will », ainsi que l'idée d'approche par prise de risque et allons montrer comment ils interagissent avec l'idée d'agent artificiel et avec certaines structures de l'IA. Certains éléments de l'apprentissage de la moralité chez l'humain se font de manière extrêmement similaires à l'apprentissage que l'on retrouve chez l'IA dans certaines formes de « machine learning » supervisé. Nous allons aussi soulever certaines attentes que l'on est en droit d'avoir vis-à-vis de l'IA. Cette attente, nous allons enfin discuter dans un quatrième chapitre comment elles pourraient être rencontrées, en discutant les limites de l'IA, un agent moral imparfait, qui sont, dans une certaine mesure, des limites que nous lui imposons. Nous allons discuter certaines solutions pratiques à des problèmes que l'IA peut soulever, et conclure en abordant succinctement la question de l'incarnation.

L'objectif de cette partie est avant tout de mettre à disposition du lecteur les éléments de réponses dont nous disposons pour répondre à la question initiale des conditions de possibilité de l'agence morale artificielle. Il s'agit aussi dans cette partie de discuter les implications d'une telle situation, ainsi que de montrer comment certains éléments, comme l'incarnation et une forme plus complète de liberté sont autant d'arguments qui limitent l'agence d'un agent artificiel.

Un agent moral artificiel

La première condition soulevée par Sullins est naturellement celle de l'autonomie. Nous l'avions déjà discuté dans la première partie de ce travail, le rocher qui tombe alors

qu'il a été jeté n'est naturellement pas responsable de sa chute. Il n'est pas un agent dans la mesure où il n'est pas capable de poser un acte. Le concept d'autonomie est un concept qui gravite autour du concept de liberté. Si l'on reprend l'une des définitions que nous avions vues en ce qui concernait l'action, il s'agit de dire que l'action est libre si elle n'est pas restreinte par une contrainte extérieure. Ainsi, il s'agit de distinguer l'agent autonome de l'outil. Lorsqu'un bûcheron frappe un rondin de sa hache, la hache n'a aucune responsabilité dans l'entaille qu'elle a formée, parce que son mouvement trouvait son origine dans une cause extérieure à elle-même. Par contre, si une machine est construite de telle sorte qu'elle détecte la présence du rondin et tranche celui-ci à l'aide d'une lame lorsqu'il passe dans son angle d'action, l'action trouve son origine à l'intérieur de la machine, dans sa programmation, qui détecte le rondin et agit.

> The first question asks if the robot could be seen as significantly autonomous from any programmers, operators, and users of the machine. I realize that 'autonomy' is a difficult concept to pin down philosophically. I am not suggesting that robots of any sort will have radical autonomy; in fact I seriously doubt human beings have that quality. I mean to use the term 'autonomy,' in the engineering sense, simply that the machine is not under the direct control of any other agent or user. The robot must not be a telerobot or be temporarily behaving as one. If the robot does have this level of autonomy, then the robot has a practical independent agency. (…) When that agency causes harm or good in a moral sense, we can say the machine has moral agency. Autonomy as described is not sufficient in itself to ascribe moral agency. Thus entities such as bacteria, or animals, ecosystems, computer viruses, simple artificial life programs, or simple autonomous robots, all of which exhibit autonomy as I have described it, are not to be seen as responsible moral agents simply on account of possessing this quality. They may very credibly be argued to be agents worthy of moral consideration, but if they lack the other two requirements argued for next, they are not robust moral agents for whom we can credibly demand moral rights and responsibilities equivalent to those claimed by capable human adults.[57]

Néanmoins, comme le précise Sullins dans le passage précédent, une telle autonomie n'est pas suffisante pour pouvoir assigner l'agence morale à un agent artificiel. Il y a d'autres conditions qui doivent elles aussi être rencontrées.

---

[57] SULLINS, J.P. , « When is a robot a moral agent » in *International Review of Information Ethics*, Vol.6, (12/2006), p28.

Il semble qu'il soit possible de pousser la réflexion relative à l'autonomie un peu plus loin. Sullins adopte, comme nous l'avons discuté en introduction, volontairement une posture minimaliste, et tente de déterminer, avec le moins de suppositions possibles, quels éléments sont nécessaires afin d'attribuer l'agence morale. Mais son approche concernant l'autonomie peut être complétée par une discussion autour de la liberté, de la contrainte, et de la capacité à agir autrement. Ces trois éléments, nous les avons discutés dans la première partie de ce travail, et avons tenté de montrer les différentes façons dont ils peuvent être approchés, et les différentes théories, plus ou moins contradictoires, à leur sujet. Si nous reprenons la théorie la plus large, qui considère qu'un acte est libre s'il n'est pas contraint par une contrainte extérieure, alors nous ne rencontrons aucun problème dans le cas d'un robot, même relativement simple. Mais il y aurait de bonnes raisons de trouver insatisfaisant un tel raisonnement. D'aucuns pourraient comparer un tel robot à une personne ayant subi un lavage de cerveau. Si un agent n'est littéralement pas capable de faire autrement que d'agir d'une telle façon précise, et que cette limite est le résultat de l'expression de la volonté d'un autre agent qui a fait en sorte qu'il soit limité de la sorte, nous sommes en droit de nous demander qui est effectivement un agent, et qui n'est qu'un outil.

Une façon intéressante d'apporter un éclairage sur ce questionnement, est de le comparer à un questionnement similaire sur l'autonomie d'un subalterne soumis à l'autorité de son supérieur[58]. Il semble que deux éléments déterminent si l'autonomie peut être conservée dans une situation d'obéissance. D'une part, la question du choix volontaire. Le subalterne a-t-il décidé de lui-même de soumettre son autonomie à son supérieur ? S'il l'a effectivement fait en étant correctement informé sur les implications d'un tel choix, on peut penser qu'il conserve une certaine forme d'autonomie, ou du moins suffisamment d'autonomie que pour être responsable de ses actes[59]. D'autre part, dans quelle mesure le subalterne a-t-il la possibilité de refuser l'autorité de son supérieur, et de ne pas agir conformément à ce qui lui est ordonné ? Ce point-ci est légèrement plus complexe, il y a plusieurs éléments qui doivent être considérés. Tout d'abord et avant de rentrer dans le sujet lui-même, ce questionnement renvoie au questionnement sur la nécessité d'une alternative pour être libre. Qu'une alternative ne soit pas disponible n'est,

[58] ROCHA, J., « Autonomy Within Subservient Careers » in *Ethic Theory Moral Prac*, (2011).
[59] Cette question sera explorée plus en détail dans le chapitre sur la responsabilité.

pour certains, pas suffisant pour affirmer que l'acteur n'est pas libre. Un soldat qui refuse les ordres est virtuellement capable de refuser ses ordres, ceux qui l'ont fait en sont d'ailleurs la preuve historique, mais les conséquences qu'il aura à affronter sont souvent décourageantes. Il en va de même, dans une moindre mesure, pour un employé qui refuse d'agir comme son patron le demande et qui se faisant met son emploi en danger. Ces situations se rapprochent plus qu'il n'y parait au premier abord à des actions effectuées sous la contrainte. Dans ces cas, bien qu'en apparence il n'y ait pas de contraintes directes sur l'action de ces agents, bien que personne ne force physiquement l'employé à suivre les ordres, il y a des contraintes indirectes qui poussent l'agent à obéir, ou au moins qui pèsent dans la balance décisionnelle de l'agent. On voit assez bien que dans notre vie quotidienne, ces contraintes indirectes sont omniprésentes. Notre balance décisionnelle nous pousse à agir d'une manière ou d'une autre, en fonction des conséquences attendues de nos actions, afin d'atteindre ce à quoi nous aspirons inconsciemment[60].

Si l'on tient un agent artificiel aux mêmes standards que ceux auxquels l'on tient un agent humain, on pourrait craindre qu'il soit incapable de délibérer à l'encontre de ce qui est attendu de lui et qu'il se rapproche plus de l'outil que de l'agent. Néanmoins, ce que nous ne devons pas perdre de vue, c'est qu'un agent artificiel suffisamment complexe peut lui aussi être capable de délibération en vue d'atteindre certains objectifs ultimes. Si une voiture autonome est contrainte de suivre le code de la route mais décide de traverser une ligne blanche pour éviter d'écraser un piéton, elle-même se rebelle contre l'injonction du code de la route pour poursuivre un objectif supérieur. C'est la même chose qui se passe lorsqu'un système automatique de freinage fait ralentir une voiture pour éviter un accident alors que son conducteur accélérait encore. On voit avec ces exemples assez nettement qu'une IA peut être capable de refuser d'obéir à des ordres si sa délibération lui permet de déterminer qu'elle doit poursuivre un objectif supérieur auquel, par essence, elle aspire. La formulation de cette dernière phrase laisse volontairement entendre deux éléments. Le premier élément est que l'on peut comparer l'objectif fondamental d'une IA aux objectifs de vie inconscients des agents humains. Une différence majeure entre ces deux types d'objectifs est qu'il semble a priori que ceux des agents humains ne soient pas

[60] L'expérience nous montre bien souvent qu'il y a une différence entre les objectifs que nous croyons avoir dans notre vie et ceux pour lesquels nous sommes effectivement prêt à sacrifier beaucoup. Quelques soient les raisons qui amène à cette distinction, il faut noter que nos « véritables » objectifs de vie sont parfois plus inconscients que conscients.

déterminés par des agents extérieurs, alors que ceux de l'IA sont calibrés par des agents humains. Néanmoins, de la même manière que l'on apprend à une IA capable d'apprendre qu'elle doit cibler tel ou tel objectif, on peut penser que l'éducation d'un enfant, ainsi que toutes les expériences de vie qui façonnent de manière très imprévisible les objectifs de vie de tout agent humain sont au moins partiellement liés à des décisions d'autres agents. L'autonomie humaine semble elle-même régresser à l'infini si l'on suit ce raisonnement. Plutôt que de plonger dans cette régression à l'infini, nous avons ici l'occasion de reconnaitre que l'autonomie humaine est plus ténue que l'on pourrait le croire, qu'il est parfois difficile d'affirmer qu'un agent extérieur n'a pas contraint de façon indirecte nos objectifs de vie à être tels qu'ils sont. Si l'on accepte cette limite de la nature humaine, on ne doit pas pour autant estimer que l'humain est incapable d'autonomie, mais reconnaitre que cette autonomie est limitée par un certain nombre de facteurs, et nous n'avons pas ici mentionné les facteurs biologiques comme le désir de se reproduire, la peur de la mort, la faim, la soif, etc. qui limitent encore plus notre autonomie. Certes l'agent artificiel poursuit des objectifs qui lui ont été inculqués par un autre agent, il ne peut être tenu responsable des objectifs qui sont les siens parce qu'il n'est pas responsable de son essence, de la même manière qu'un humain ne peut se voir reprocher d'avoir faim. Par contre, l'agent artificiel est moins tenu que l'agent humain à certaines considérations qui troublent nos jugements, comme les considérations biologiques. Ces éléments biologiques sont certes des atouts dans le questionnement moral à certains égards, comme nous l'avons discuté précédemment, mais ce sont aussi parfois des obstacles qui mènent à l'égoïsme.

Pour conclure ce passage sur l'autonomie, retenons certains éléments. Tout d'abord l'approche de Sullins, qui dit clairement qu'un agent artificiel n'a pas besoin d'être véritablement autonome mais doit au moins sembler autonome. Par ailleurs, il est problématique d'excuser quelqu'un pour un comportement immoral sous prétexte qu'il a reçu une mauvaise éducation dans son enfance, parce que cette excuse affirme en même temps qu'il n'est pas un agent moral capable d'assumer ses actes. De la même façon, il serait problématique de ne pas reconnaitre comme agent moral une IA capable de décisions morales autonomes sous prétexte qu'elle poursuit un objectif qui lui a été inculqué par sa propre éducation, alors qu'il est incertain d'affirmer que nos propres objectifs viennent de nous-mêmes. Si nous attribuons sans mal l'agence morale à un humain, malgré ses limitations, il n'est pas logique de ne pas attribuer la même agence à

une IA pour des limitations du même ordre. Notons enfin que la discussion sur l'autonomie s'approche régulièrement de la discussion sur la responsabilité, et que nous préciserons cet élément dans le troisième temps de cette réflexion.

Le second élément que Sullins souligne est l'intentionnalité. Nous avons déjà abordé la question de l'intentionnalité dans la première partie de ce travail, et avions conclu qu'en l'absence de la possibilité de lire les pensées de l'agent en question, cette dimension s'avérait peu utile à notre réflexion. Cependant, ce que Sullins discute dans son passage sur l'intentionnalité est une fois de plus une approche minimaliste de l'intentionnalité, à savoir non pas l'intentionnalité elle-même mais bien l'apparence de l'intentionnalité. Une telle approche est bien plus proche de la discussion que nous avions ouverte en ce qui concerne la « Quality of Will ».

> The second question addresses the ability of the machine to act 'intentionally.' Remember, we do not have to prove the robot has intentionality in the strongest sense, as that is impossible to prove without argument for humans as well. As long as the behaviour is complex enough that one is forced to rely on standard folk psychological notions of predisposition or 'intention' to do good or harm, then this is enough to answer in the affirmative to this question. If the complex interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial, and the actions are seemingly deliberate and calculated, then the machine is a moral agent. There is no requirement that the actions really are intentional in a philosophically rigorous way, nor that the actions are derived from a will that is free on all levels of abstraction. All that is needed is that, at the level of the interaction between the agents involved, there is a comparable level of personal intentionality and free will between all the agents involved.[61]

Ce qui compte, c'est bien la capacité à manifester une intention, plus que l'intention elle-même. Ceci nécessite assez simplement que l'agent soit capable de démontrer que ses actions ne sont pas inconsidérées, et sont l'objet d'une délibération, et non d'une simple pulsion. A ce titre une fois de plus, il semble bien que l'action de l'humain puisse parfois être ramenée à une pulsion, et dès lors nous pouvons sans difficulté imaginer qu'un agent artificiel soit capable d'être au moins autant capable de démontrer qu'il agit de manière qui semble intentionnelle.

---

[61] SULLINS, J.P. , »When is a robot a moral agent" in *International Review of Information Ethics*, Vol.6, (12/2006), p 28.

Enfin, le troisième élément que Sullins requiert afin de qualifier un agent artificiel d'agent moral est la responsabilité. Nous avons déjà approché la question de la responsabilité dans le passage précédent sur l'autonomie, et par ailleurs nous avons extensivement discuté celle-ci dans le questionnement qui a été le nôtre dans la première partie de ce travail. En réalité, la responsabilité se trouve au cœur de l'attribution de l'agence morale à un agent. Est responsable de ses actes l'agent qui est capable d'en assumer les conséquences, de les avoir au moins partiellement envisagées et de recevoir blâme ou félicitations pour ses actes. Notons tout d'abord que Sullins, dans son article, ne parle pas à proprement parler de la responsabilité, de la même manière qu'il ne parlait pas de l'intentionnalité, il ne parle que de l'apparence de responsabilité.

> Finally, we can ascribe moral agency to a robot when the robot behaves in such a way that we can only make sense of that behaviour by assuming it has a responsibility to some other moral agent(s). If the robot behaves in this way and it fulfils some social role that carries with it some assumed responsibilities, and only way we can make sense of its behaviour is to ascribe to it the 'belief" that it has the duty to care for its patients, then we can ascribe to this machine the status of a moral agent.[62]

Doit-on se contenter dès lors de l'apparence de responsabilité pour considérer qu'un agent artificiel peut être un agent moral ? C'est sans nul doute un prérequis minimal et nécessaire, mais il est incertain qu'il s'agisse en même temps d'un prérequis suffisant. Cependant, gardons bien à l'esprit que Sullins lui-même n'affirme pas que ce soit un prérequis suffisant, mais bien que la conjonction de l'autonomie, l'intentionnalité et la responsabilité telles qu'il les a décrites, de manière minimaliste, est elle-même suffisante pour assigner le statut d'agent moral. De manière similaire à la question de l'intention, nous n'allons pas ici prétendre déchiffrer la boite noire que peut être le processus de délibération de l'agent en question, qu'il soit artificiel ou pas. Dès lors que l'on en reste à ce que l'on peut maitriser, ce qui nous apparait, la question de la responsabilité et de l'intentionnalité vient compléter le premier questionnement autour de l'autonomie de manière adéquate. Si la question de l'autonomie nous laissait à douter qu'un agent artificiel puisse être qualifié d'agent moral, nous voyons nettement comment il ressemble bien plus à un agent moral s'il est capable de démontrer une dimension d'intentionnalité

[62] SULLINS, J.P. , »When is a robot a moral agent" in *International Review of Information Ethics*, Vol.6, (12/2006), p28.

et de responsabilité. La seule réserve que nous conservons, et que nous avions déjà évoquée dans le passage sur l'autonomie, est que contrairement à l'humain, nous sommes en mesure de connaître les objectifs de vie, la raison d'être d'une IA, et que cette raison d'être est altérable artificiellement. Dans une certaine mesure, on peut dire que l'IA ne dispose pas de la dignité humaine kantienne, vu qu'elle n'est considérée que comme un moyen, jamais comme une fin. Cette dernière réflexion ouvre elle-même une option qui pourrait permettre de résoudre cette situation, dans la quête d'un agent moral artificiel complet, et que nous explorerons en conclusion.

L'IA et les différents courants éthiques

Nous allons reprendre dans ce chapitre les quatre théories éthiques que nous avons introduites en première partie, à savoir le déontologisme, l'utilitarisme, l'éthique de la vertu et l'éthique du care, et tenter de montrer les liens que chacun de ces courants peuvent avoir avec l'IA. Il faut garder à l'esprit que le rapport entre l'IA et chacun de ces courants pourrait faire l'objet d'un travail de l'ampleur de celui-ci. Ce n'est pas ici l'objet de notre attention en particulier, mais il y a néanmoins quelques éléments majeurs que nous allons souligner.

Comme nous l'avons déjà évoqué, le déontologisme kantien est traditionnellement associé à l'approche descendant en IA. En effet, le déontologisme kantien est supposément capable de nous mettre à disposition un ensemble de règles qu'il s'agirait alors de suivre afin d'agir moralement. Dans la mesure où ces règles sont traductibles dans un langage informatique, il parait relativement simple d'utiliser cette approche pour s'assurer du comportement moral d'un agent artificiel. Hélas, une telle approche soulève au moins trois problèmes.

Premièrement, la capacité de la théorie kantienne à mettre à disposition un ensemble de règles claires et précises est pour le moins discutable. Nous pouvons difficilement nous contenter des impératifs catégoriques, tout simplement déjà parce qu'ils sont pensés pour être utilisés par des humains. Lorsqu'il s'agit d'universaliser la maxime de son action, les critères permettant de déterminer si elle peut ou non être universalisée varient drastiquement s'ils s'appliquent à un humain ou à une IA. Par exemple, dois-je sacrifier ma vie pour avoir une chance sur deux de sauver une autre personne ? Rien ne permet d'affirmer qu'un tel sacrifice soit attendu moralement d'un

humain. Si tout le monde agit de la sorte, l'humanité aurait vite fait de s'éteindre. Par contre, un robot doit-il mettre l'intégrité de sa machine en danger pour tenter de sauver un humain ? Nous voyons bien comment la réponse est très différente. De manière similaire, le second impératif considère la dignité humaine, et part lui-même du point de vue d'un humain. Enfin, nous pourrions tenter de décrire un ensemble de règles à partir du mode de pensée kantien, et constituer un ensemble de lois telles qu'elles rendraient immoraux certains actes comme le mensonge, le fait de blesser quelqu'un, le fait de ne pas aider quelqu'un dans le besoin, etc. Mais une telle tâche est vouée à l'insuffisance. Si ce que nous serions capables de mettre en évidence pourrait néanmoins être intéressant, il n'y a pas de doute que nous ne serions pas en mesure de passer en revue la totalité des actions immorales, simplement parce que ce questionnement est un questionnement en cours, qui fait régulièrement face à de nouvelles problématiques, auxquelles personne n'avait pensé, et qui requiert alors une certaine capacité d'adaptation. De plus, un tel ensemble de règles serait voué à mener à des situations contradictoires telles que certains exemples donnés contre le déontologisme kantien. Si je ne peux pas blesser quelqu'un mais que je le blesserais en lui disant la vérité, que dois-je lui dire ? Une telle question nécessite de toute évidence une capacité qui n'est pas décrite lorsque l'on se contente de suivre un ensemble de règles.

Le second problème relatif au déontologisme est le problème de la traduction. Afin de rendre le déontologisme pertinent, il est bien souvent nécessaire de parler de règles complexes, comme les différents impératifs kantiens. Ceux-ci sont plus versatiles que ne peuvent l'être des règles simples, et sont plus à même de couvrir la totalité des situations. Hélas, s'il est possible de traduire une règle comme celle qui contraint à ne pas mentir, la traduction de règles plus complexes est significativement plus compliquée. En réalité, le degré d'abstraction qu'elle requiert exclut tout simplement qu'elle soit traduite comme telle en un langage informatique quel qu'il soit. Le passage de la règle écrite à l'algorithme est une étape fondamentale qui, dans ce cas précis, est insurmontable. La réduction de nuance qu'une telle tâche nécessiterait rendrait sans doute trivial l'usage d'une telle règle.

Enfin le troisième questionnement est le plus fondamental. Si une IA est programmée de cette façon, s'agit-il vraiment encore d'une IA au sens strict ? Certes une définition très ouverte du concept d'IA l'intégrerait sans mal, mais il devient difficile de parler d'agent moral artificiel si un tel agent est littéralement contraint par des règles

extérieures et que l'ensemble de son comportement est régi par ces règles. En réalité, une IA faible semble possible dans une programmation descendante, mais une IA dans un sens plus fort ne semble pas possible dans un tel cadre. Ceci ne veut pas dire que nous devons par définition nous débarrasser de l'approche descendante, qui a beaucoup à apporter à l'IA, mais simplement que cette approche à la qualité, et le défaut, de nous permettre de regarder derrière le rideau, et un tel regard empêche ne serait-ce que l'apparence de l'intelligence au sens fort.

Si le déontologisme est plus proche de l'approche descendante, on peut dans une certaine mesure dire de l'utilitarisme qu'il est plus proche de l'approche ascendante en IA. En effet, l'utilitarisme et son calcul félicifique se prête bien plus à un comportement libre qui maximise certaines métriques. Dans le cadre du calcul félicifique à proprement parler, si une IA était capable de mesurer le bonheur, elle n'aurait qu'à choisir l'action qui le maximise. La question de l'agence morale artificielle serait de ce pas résolue. Hélas nous venons de pointer du doigt le problème majeur de l'utilitarisme pour son exportation à l'IA. Le calcul félicifique a beau sembler être une approche mathématique, supposément objective, de la question morale, il n'empêche qu'il est construit sur des bases fondamentalement subjectives. Comment peut-on associer une valeur au bonheur d'un individu ? A priori, ce défi semble en lui-même impossible, et semble rejeter les chances de l'utilitarisme d'être mobilisé de manière pertinente dans le cadre de l'IA. Néanmoins, l'utilitarisme présente certains autres avantages.

Un des principes de l'utilitarisme est que le bonheur d'un individu a autant de valeur que celui d'un autre individu. Dès lors, une IA mobilisant des principes d'apprentissage pourrait être nourrie des retours d'un grand nombre d'individus, lui permettant d'évaluer son action de sorte à maximiser l'approbation[63] que son action rencontre. Moyennant une phase de test, une telle IA pourrait dès lors présenter un comportement suffisamment moral que pour être déployée, afin d'être mise régulièrement en contact avec de nouvelles situations et de, graduellement, étendre sa connaissance des situations et des réactions qu'on est en droit d'attendre d'elle. Une telle IA serait donc non seulement capable de choisir l'action qui a été la plus approuvée, mais en plus capable d'adapter son comportement aux changements de réception de ses actions. Si elle réalise

[63] Il y a une nuance entre « action approuvée » et « action maximisant le bonheur » et il serait sans doute nécessaire de mêler ces deux approches pour optimiser le résultat.

au fur et à mesure du temps qu'un tel comportement est de moins en moins approuvé, elle pourra le modifier. Une telle IA rencontre néanmoins le même problème, auquel on est en droit de s'attendre pour tout agent artificiel, dans le cas d'une situation inédite. Si une situation n'a jamais été rencontrée par l'IA dans sa phase de test, il est difficile d'imaginer sa réaction avant qu'elle ne reçoive ses premiers retours. Bien souvent, l'IA agit de manière inadéquate parce qu'elle est contrainte de tenter une réponse à la situation de manière relativement aléatoire. Le problème est évidemment que si une telle situation inédite s'avère être rare, ce qui est probable, si elle n'a pas été rencontrée en phase de test, l'IA sera très lente à acquérir des données sur cette situation et progressera tout aussi lentement dans sa recherche d'une réponse appropriée. On pourrait cependant être tenté de balayer sous le tapis une telle situation, sous prétexte justement qu'elle est rare, mais si les conséquences de l'action tentée par l'IA sont dramatiques, un tel évènement rare risquera de faire date dans la longévité de cette IA.

L'autre approche concernant l'IA serait donc de prétendre résoudre le calcul félicifique et de prendre en compte les différentes circonstances pour écrire la formule permettant de construire un algorithme maximisant le résultat du calcul félicifique. La première difficulté arrive dès le premier élément de ce calcul, comment peut-on prétendre calculer l'intensité du bonheur de quelqu'un ? Quand bien même nous puissions résoudre la place que prendrait chacune des autres circonstances dans le calcul, quand bien même nous serions capables de déterminer une formule n'ayant pour inconnue que « I »[64], l'intensité du bonheur que procure une action, cette inconnue nous laisserait toujours face à un mur. Et c'est sans entrer dans la complexité supplémentaire qui prend en compte que chaque personne ressent un bonheur différent, dans une intensité différente, et avec un rapport différent à la proximité du bonheur. D'aucuns prétendent que le bonheur est une réaction à certaines hormones comme la dopamine, mais il semble illusoire d'imagine qu'un monde heureux est un monde dans lequel chacun reçoit une certaine dose hormonale de manière quotidienne. Et quand bien même, si c'est un monde heureux, beaucoup ne seraient pas d'accord avec l'affirmation selon laquelle ce serait un monde moral.

[64] On peut sans doute considérer qu'il pourrait être possible d'assigner une valeur plus ou moins élevée à chaque bonheur potentiel, de la même manière que les réseaux de neurones équilibrent le poids de chaque lien. La question serait de déterminer alors en fonction de quelle métrique ajuster ces poids. C'est éventuellement une piste qui vaudrait la peine d'être creusée.

Après ces considérations concernant l'éthique déontologiste et l'utilitarisme, considérons maintenant l'éthique de la vertu. L'éthique de la Vertu présente autant d'avantages que d'inconvénients. Même débarrassée de son rapport à l'intentionnalité, elle s'attarde plus sur une éthique de l'agent que sur une éthique de l'action. Or, si notre objectif est bien de permettre l'émergence d'un agent moral artificiel, la seule manière que nous avons d'évaluer un tel agent reste au travers de ses actions. Sans rentrer ici dans le débat selon lequel l'action façonne l'agent, il est tout de même intéressant de noter que l'idée d'un agent moral artificiel pourvu de certaines qualités morales, de certaines vertus, lui permettrait de disposer de la flexibilité nécessaire à une réaction, si pas idéale, au moins appropriée, face à des situations inédites. De la même manière qu'après un certain entrainement, un pigeon devient capable de reconnaitre une tumeur maligne sur une radio[65] à laquelle il est exposé pour la première fois, on peut imaginer qu'après avoir été exposée à de très nombreuses situations différentes, une IA soit capable de s'adapter à une situation inédite. On peut cependant craindre que quel que soit le nombre de radiographies mammaires que l'on présente à un pigeon, il ne soit jamais capable, après une seule exposition, de reconnaitre quelque chose d'aussi différent qu'une fracture osseuse par exemple. Aussi, il faut que la variété des situations auxquelles l'IA est exposée soit suffisante pour au moins lui permettre de retirer des principes généraux.

L'éthique de la Vertu est relativement similaire au « machine learning » supervisé, en ce que son mode d'enseignement se fait par l'observation. Lorsque l'on met à disposition de l'IA un grand nombre de données traduisant des situations où un agent moral a pris des décisions jugées par l'entraineur comme adéquates, on simule en fait l'apprentissage par observation qui est prôné par cette approche. De la même manière que l'on est parfois incapable d'expliquer exactement pourquoi tel choix nous semble meilleur que tel autre dans un dilemme complexe, cette situation permet d'apprendre à l'IA à imiter un agent moral sans pour autant avoir besoin, de notre côté, de comprendre exactement ce qui fait que tel choix est meilleur que tel autre choix. Ce degré d'incertitude, nous en discuterons à la fin de cette partie.

Dans la première partie de ce travail, nous avons évoqué l'éthique du Care, et avons laissé penser qu'elle pourrait amener des éléments intéressants pour notre réflexion

[65] LEVENSON, R.M., et al., « Pigeons (Columba livia Columba livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images » in *Plos*, (08/2015).

relative à l'agence morale artificielle. L'éthique du Care, de toute évidence, est un cas particulier. On ne peut évidemment pas simplement prétendre pouvoir l'intégrer au sein d'une IA via une approche descendante, et il n'est pas clair d'affirmer qu'une approche ascendante aurait de meilleurs effets. L'éthique du Care se fonde notamment sur la vulnérabilité. Cette vulnérabilité, il ne s'agit pas simplement de la vulnérabilité d'une personne victime d'une certaine souffrance à laquelle l'agent devrait être attentif, il s'agit de la vulnérabilité de l'agent lui-même. C'est évidemment ce point particulier qui semble, a priori, aliéner l'IA de tout rapport à l'éthique du Care. L'IA n'est pas capable de souffrir, et donc pas capable de l'empathie requise pour comprendre la souffrance d'autrui. C'est ce point qu'il va s'agir de discuter. Si nous reprenons l'approche de Sullins, il n'est pas nécessaire, dans le cadre d'une IA au sens faible, qu'elle soit capable de comprendre la souffrance d'autrui, il est seulement nécessaire qu'elle soit capable d'agir comme si elle la comprenait. En réalité, il est légitime de se demander si nous sommes capables de comprendre la souffrance d'autrui. Il y a une distance entre l'expérience subjective d'un individu et celle d'un autre qui est irréductible. Mon ressenti personnel m'est propre et je suis la seule personne qui en fera un jour l'expérience. Je peux tenter de la verbaliser, de l'expliquer, de la montrer de quelque manière que ce soit, jamais quelqu'un d'autre que moi ne pourra vivre mon point de vue. Cependant, il convient de reconnaitre que notre capacité à souffrir nous permet d'avoir une idée de ce que souffrir signifie, et nous permet d'apprécier la souffrance d'autrui à sa juste valeur, en partant de notre propre souffrance comme référence. Mais pourtant, certains évènements laissent certains individus de marbre alors que d'autres se retrouvent en sanglot. Certains n'en souffrent pas alors que d'autres sont dévastés. Cela n'empêche qu'une personne qui ne souffre pas puisse trouver en elle l'empathie de comprendre la souffrance d'autrui et de l'apprécier à sa juste valeur. Bien qu'elle n'ait pas de concept de souffrance, rien n'empêche qu'une IA soit capable d'agir comme si elle éprouvait de l'empathie pour autrui. C'est là cependant un défi qui est très loin des capacités actuelles de l'intelligence artificielle.

Une autre piste consisterait à tenter de faire expérimenter à une IA sa propre vulnérabilité, même si celle-ci serait inévitablement très différente de ce que peuvent ressentir des humains. Si la vulnérabilité d'une personne est de toute façon différente de celle d'une autre personne, la vulnérabilité d'une IA, bien qu'elle soit d'un tout autre ordre, pourrait permettre de combler une partie de la distance qui nous éloigne de l'IA. Le

problème évidemment est de conceptualiser la vulnérabilité d'une IA, et surtout de prétendre qu'il soit possible de lui faire comprendre sa vulnérabilité. Certes l'existence même de l'IA est à la merci du bon vouloir de son propriétaire, et à ce titre sa vulnérabilité est apparente, mais pour autant, il faudrait que l'IA possède déjà une conscience de sa propre existence pour que cette vulnérabilité ait le moindre sens pour elle. Il est inutile de souligner que cette piste apporte plus de complications qu'elle n'apporte de solutions.

La discussion autour de ces quatre courants éthiques nous permet finalement de montrer pourquoi, dans la seconde partie de ce travail relative à l'IA elle-même, il nous était difficile d'opter pour une IA qui agit comme un humain plutôt qu'une IA qui agit rationnellement. C'est-à-dire qu'il apparait qu'une IA qui agit rationnellement n'est pas capable de vulnérabilité, n'est pas capable de souffrance et ne peut dès lors jamais véritablement prétendre comprendre la souffrance humaine, elle ne peut jamais faire preuve d'une empathie véritable. A moins que ce défi ne soit relevé, et que ce problème ne soit résolu, l'idée que l'on puisse concevoir un agent artificiel moral à partir d'un comportement similaire à celui d'un humain semble illusoire. C'est pourquoi nous sommes contraints de nous diriger vers la piste d'une IA qui agit rationnellement. Les seules pistes que nous avons en faveur d'une agence morale artificielle passe par un rapport rationnel à des informations, à des données. Néanmoins, gardons à l'esprit l'approche de Sullins, qui consiste globalement à se contenter de sembler être plutôt que d'être, et qui défend d'abord une IA faible, avant de prétendre à une IA forte. Dans ce cadre précis, on peut défendre l'idée d'une IA qui agit rationnellement mais qui semble agir comme un humain. En réalité, ce serait son imitation d'un humain qui serait construite à partir d'informations traitées rationnellement. Par ailleurs, on pourrait évidemment aller plus loin, et ce serait souhaitable, que de se contenter d'imiter un humain. Sullins décrit le point de vue du philosophe Eric Dietrich de la façon suivante.

> Conversely, it is logically possible, though not probable in the near term, that robotic moral agents may be more autonomous, have clearer intentions, and a more nuanced sense of responsibility than most human agents. In that case there moral status may exceed our own. How could this happen? The philosopher Eric Dietrich argues that as we are more and more able to mimic the human mind computationally, we need simply forgo programming the nasty tendencies evolution has given us

> and instead implement, "…only those that tend to produce the grandeur of humanity, we will have produced the better robots of our nature and made the world a better place" (Dietrich, 2001).[66]

Si nous adoptons le point de vue de Dietrich à ce sujet, nous pourrions alors même être en droit d'espérer donner naissance, dans un futur relativement lointain, à des agents moraux artificiels plus aptes que ne le sont les agents moraux humains, précisément parce qu'ils ne seraient pas entravés par les souffrances et les vulnérabilités qui, si elles sont à l'origine de notre empathie, de notre altruisme et de notre moralité, sont aussi à l'origine de notre égoïsme.

A court terme, les meilleures pistes dont la recherche en IA dispose restent certainement l'éthique de la Vertu, le déontologisme et l'utilitarisme, mais il n'empêche que l'éthique du Care puisse, par imitation, permettre à l'IA de nuancer certaines prises de positions, et lui permette d'affiner son jugement. Nous discuterons dans le dernier chapitre de cette partie comment de telles approches peuvent, concrètement, avoir du sens pour notre réflexion.

L'IA et les différentes problématiques de l'agence morale

Dans la première partie de ce travail, nous avons notamment discuté l'approche « forward-looking » qui considérait, dans les grandes lignes, qu'un jugement moral devait être attribué à une action dans l'objectif de favoriser les actions jugées comme bonnes au détriment de celles jugées mauvaises. Dans la seconde partie de ce travail, nous avons expliqué comment l'apprentissage supervisé utilisait exactement ce genre de méthodes pour tenter d'obtenir de bons résultats. Dès lors, si l'humain agit moralement en vue d'être félicité et d'éviter les blâmes, l'IA peut fonctionner exactement de la même manière. Evidemment, comme nous l'avons décrit, l'IA aura à ce titre besoin d'un temps d'apprentissage relativement long, mais on peut s'attendre à ce qu'une progression systématique se présente au fur et à mesure que l'IA est blâmée ou félicitée pour ses actes. Si cette méthode fonctionne pour un agent humain, elle peut aussi fonctionner pour l'agent artificiel. Contrairement à l'humain, on peut garantir que l'agent artificiel se conforme

[66] SULLINS, J.P. , »When is a robot a moral agent" in *International Review of Information Ethics*, Vol.6, (12/2006), p29.

aux indications qui lui sont données et cherche activement à maximiser les félicitations. Certains humains peuvent réagir de manière bien moins rationnelle face aux reproches.

Une autre approche que nous avons discutée est celle que la « Quality of Will ». Pour rappel, cette approche insistait sur l'importance de garantir celui qui reçoit notre action (le patient moral) de notre bienveillance à son égard. Cet aspect est moins évident à comprendre dans le cadre particulier de l'IA, mais souligne néanmoins le besoin que nous avons de comprendre l'action d'autrui. En réalité, nous avons moins besoin de comprendre les racines exactes qui justifient son comportement, celle-ci échappant même bien souvent à la connaissance de l'agent lui-même, que d'entendre une justification rationnelle face à une action qui nous déplait. A ce titre, il semble pertinent d'attendre d'une IA qu'elle soit capable de démontrer sa rationalité à un agent humain. Même si les raisons précises et détaillées de son action peuvent rester trop complexes que pour nous être accessibles, il n'est pas absurde d'imaginer qu'une IA soit capable de souligner des éléments majeurs qui ont mené à sa décision. Les seuls moments où une telle requête paraitrait difficile à exiger sont les moments où l'IA ferait face à un dilemme moral particulièrement complexe. Dans ces cas-là, quel que soit l'avis que l'on ait sur l'existence ou non de quelque chose comme une vérité morale universelle, nous sommes tous contraints de reconnaitre qu'il existe des dilemmes pour lesquels l'humain n'est pas capable de pointer du doigt une réponse parfaite et unique. Dans ces zones de gris de la moralité, où aucun choix ne semble parfait, l'agent humain est bien en peine de justifier de manière satisfaisante un choix qu'il serait contraint de faire. Evidemment, c'est le genre de situation dans lesquelles un agent moral idéal s'abstiendrait, si possible, de décider, en l'absence de certitude. Mais ce n'est pas toujours possible. Si un conducteur se trouve sur une route et voit subitement surgir en face de lui deux personnes, de part et d'autre de la route, il risque d'être contraint de prendre une décision. S'il est clair que sans réaction de sa part il risque fortement d'écraser les deux piétons, il sera contraint de choisir un côté. Ce choix entre deux situations équitablement problématiques renvoie au dilemme de l'âne de Buridan, où un âne entre deux bottes de foin se laisse mourir de faim, incapable de choisir vers quelle botte de foin tourner la tête. Quand il n'y a pas de bonne décision, il arrive que l'humain soit lui-même un âne de Buridan, incapable de choisir, et dans ce genre de situation, l'humain se retrouve parfois à écraser les deux piétons, immobilisé par le doute. Si au premier abord une telle situation semble provoquer les mêmes conclusions

dans le cas d'un agent artificiel, on imagine sans difficulté qu'un tel problème puisse être évité. Si tant est qu'une situation laisse l'IA face à un statut quo parfait, on peut imaginer qu'elle soit pourtant contrainte de choisir, quitte à ce que ce soit au hasard[67]. Evidemment, il parait très improbable qu'une situation aussi caricaturale se présente dans le monde réel, mais pour autant, quand bien même elle se présenterait, on voit comment une IA pourrait être mieux équipée qu'un humain.

Contrairement aux questionnements évoqués dans les paragraphes précédents, l'approche par risque a été pensée précisément parce qu'elle peut au mieux s'associer avec l'IA. Elle n'est pas pour autant exclusive à l'IA, elle est supposément capable de fonctionner pour tout agent moral. Il s'agit bien de garder à l'esprit que l'approche par prise de risque ne prétend pas déterminer quelle situation est préférable à quelle autre situation, mais bien de déconstruire la façon dont un tel choix peut être opéré, à partir du moment où l'on admet que l'agent est capable d'évaluer ces choix. Pour l'évaluation de ces choix, il nous faut nous référer aux différents courants éthiques dont nous avons parlé précédemment. La force de l'approche par prise de risque, c'est notamment qu'elle met en évidence le processus de modélisation auquel l'agent fait face dès lors qu'il doit prendre une décision. Dans le cas d'un agent humain, ce processus n'est pas toujours évident à discerner, mais ça ne veut pas dire qu'il n'est pas présent. Dans le cas d'une IA, il est tout à fait évident. L'IA ne délibère pas par rapport à la réalité mais par rapport à un ensemble de données fournis par ses capteurs. De la même manière, elle ne cherche pas à produire un résultat dans la réalité mais bien dans une modélisation de celle-ci, qui est représentée par un ensemble de données. Cette approche a le réalisme de ne pas prétendre que l'IA soit capable d'un rapport avec la réalité elle-même qu'elle n'a pas capable d'avoir, par essence. Elle souligne cependant de la même façon cette insuffisance. La question se pose d'ailleurs de savoir si l'humain est lui-même capable d'interagir avec la réalité ou s'il ne fait lui-même qu'interagir avec une modélisation de cette réalité. Il ne semble pas que nous ayons à disposition des éléments permettant d'affirmer que le rapport à la réalité d'un humain est plus authentique[68].

[67] Le hasard véritable peut être une problématique en soit, pour l'informatique, mais pour autant c'est un problème que l'on peut contourner. Si nous ne sommes pas forcément capables de générer un aléatoire parfait, on peut tout de même générer un aléatoire apparent suffisamment efficace.

[68] Nous n'entrons pas ici dans le débat en lui-même, qui distingue la réalité de l'objet « en-soi » de l'objet tel qu'il nous apparait, mais nous constatons simplement que c'est une question ouverte.

On attend donc de l'IA qu'elle fonctionne comme suit. Tout d'abord, elle reçoit un certain nombre de données à l'aide de ses capteurs, et produit donc une modélisation de son environnement à l'aide de ces données. L'étape de production de la modélisation est facultative, elle est ici décrite uniquement pour faire sens à nos yeux, mais si nous sommes convaincus qu'il y a une réalité derrière les données que nous percevons, c'est-à-dire derrière les apparences, une telle question n'a aucun intérêt aux yeux de l'IA. Idéalement, une IA doit être en mesure d'évaluer la quantité de données auxquelles elle n'a pas accès, afin d'évaluer la précision de son modèle. L'étape suivante consiste à envisager différentes actions, et pour chaque action à envisager les conséquences potentielles de ces actions. Pour chacune de ces conséquences potentielles, l'IA doit être en mesure d'évaluer sa prévalence morale, c'est-à-dire dans quelle mesure ces conséquences sont moralement souhaitables. Une fois une telle évaluation effectuée, l'IA se trouve face à un choix entre différentes actions ayant chacune un ensemble de conséquences plus ou moins souhaitables possibles. L'étude des probabilités nous enseigne que chaque conséquence potentielle, même la moins probable, doit être prise en compte[69]. Elle doit cependant être jugée avec un degré d'importance équivalent à sa probabilité. C'est-à-dire qu'une conséquence dix fois moins probable qu'une autre mais dix fois plus souhaitable que celle-ci a autant de poids dans la balance décisionnelle que celle qui est dix fois plus probable.[70]

Cette théorie considère qu'il faut en fait multiplier les IA à l'œuvre. Le travail qui consiste à évaluer la probabilité d'une conséquence après une telle action n'est pas le même que celui qui consiste à évaluer dans quelle mesure telle ou telle autre conséquence est souhaitable. Une telle décomposition du travail de l'IA pourrait bien mener à des

[69] Si un évènement à une très petite chance de se produire, cela veut précisément dire qu'il finira par se produire inévitablement à condition que l'on reproduise l'expérience un nombre de fois suffisant. S'il y a une chance sur 46 656, lorsque l'on lance 6 dés, de les voir tous montrer le chiffre 6, cela ne veut pas dire que cela n'arrivera jamais, au contraire, cela veut dire que sur 100 000 lancés, il est relativement probable que cette combinaison se présente. Sur un million de lancés, il devient peu probable que cette combinaison ne se présente jamais.

[70] Chaque décision mène à un certain nombre de conséquences possibles, dont la somme des probabilités d'apparition est égale à 1. Chacune de ces conséquences possibles se voit aussi associé un degré de préférence, selon qu'elle soit plus ou moins souhaitable. Si on peut être tenté de penser qu'un évènement très souhaitable mais très peu probable ne devrait pas être considéré, on constate que sur un grand nombre d'évènements, tout ce qui est possible finira inévitablement par se produire. Dès lors, un évènement très improbable doit être considéré à sa juste valeur. C'est également le cas si un tel évènement n'est pas du tout souhaitable.

résultats bien plus pertinents que lorsque l'on attend d'une IA qu'elle soit capable, en un seul morceau, de résoudre une situation dans sa totalité. Par ailleurs, le fait même de décomposer l'action de l'IA de la sorte nous permettrait d'affiner son entrainement à ces tâches de manières appropriées, et nous permettrait de cibler de manière plus adéquate ses entrainements. En l'occurrence, la première partie du processus devrait opérer d'une manière quelque peu similaire à l'opération d'une IA qui calcule différentes positions d'échecs possibles. La dernière partie du processus serait évidemment la plus simple, vu qu'il s'agirait simplement d'un processus de maximisation. La partie problématique est évidemment le morceau central, qui consiste à évaluer chacune des conséquences possibles de chacune des actions possibles. Il est probable qu'une telle approche ne soit pas transposable comme tel à l'IA et requiert d'être simplifiée, mais il n'empêche que cette réflexion nous permet de mieux comprendre comment un tel processus pourrait se produire.

L'IA : un agent moral imparfait

Chacun sait que l'humain est un agent moral imparfait. Si ce n'était pas le cas, jamais personne ne se sentirait coupable d'avoir agi d'une certaine façon jugée immorale. Par ailleurs, nos organes sensitifs sont eux aussi imparfaits, nous doutons quant à l'origine d'une odeur, nous sommes, comme tout être vivant, sensibles aux illusions d'optique, nous ne percevons qu'une partie des sons et des couleurs. Enfin, notre délibération et notre réflexion elle-même est imparfaite, nous sommes sensibles aux biais cognitifs, aux erreurs de raisonnement logique, etc. Tout cela, ce n'est rien de nouveau. Dès lors, il est absurde d'attendre de l'IA qu'elle soit un agent moral parfait. Tout ce que nous pouvons exiger d'elle, c'est qu'elle soit aussi efficace que nous. Cette barre peut sembler extrêmement élevée, mais n'oublions pas que l'IA, si elle ne vient pas avec certaines de nos qualités, ne vient pas non plus avec nos défauts. Ses capteurs peuvent être de meilleure qualité que les nôtres, son raisonnement n'est a priori pas sujet à des erreurs de logique, et par-dessus tout, l'IA n'est pas sensible aux faiblesses de caractère humaines. Une voiture autonome ne pourra jamais garantir qu'elle est sûre à 100%, par contre elle peut garantir à 100% qu'elle ne conduit pas sous influence. Elle n'est pas non plus fatiguée, pas distraite, etc. Cela ne veut pas dire qu'elle ne peut pas être sujette à une illusion d'optique qui lui ferait confondre une information avec une autre. Mais néanmoins, ce genre de problèmes sont bien plus évidents à résoudre que le problème de l'alcool au volant, du téléphone portable

au volant, et des erreurs d'attention de la nature humaine elle-même. En réalité, il n'est pas nécessaire que l'IA soit capable de conduire parfaitement, il est simplement nécessaire qu'elle conduise mieux que l'humain.

Toutes les discussions concernant l'IA sont généralement tournées vers un futur plus ou moins lointain. Mais en réalité, la seule chose qui est encore lointaine, c'est l'idée d'un agent moral parfait. Des IA sont déjà présentes dans nos vies, et certaines d'entre elle prennent déjà des décisions morales. Les voitures autonomes ne sont pas encore répandues sur nos réseaux routiers, mais il y a de bonnes raisons de penser que ce soit l'avenir de l'automobile, et que ces véhicules soient en moyenne plus sûrs que des véhicules conduits par des humains. En de telles circonstances, il serait en réalité immoral selon la plupart des théories éthiques de ne pas déployer de tels véhicules. Ne pas déployer ces véhicules reviendrait à condamner à un accident de la route tous les usagers qui auraient été épargnés si les voitures autonomes, plus sûres, avaient été favorisées. Evidemment, il restera toujours un certain nombre d'accidents inévitables, mais ceux-ci devraient être bien moins nombreux que ceux qui auraient eu lieu sans l'IA.

Un argument que nous avions soulevé dans la première partie de ce travail et qui allait à l'encontre de l'idée même d'autoriser une IA à prendre des décisions était l'argument selon lequel la nature humaine a besoin de reprocher la faute à quelqu'un lorsqu'un drame se produit. Cet argument est problématique. Nous l'avions évoqué, il existe un nombre significatif d'évènements qui ne peuvent pas être reprochés à une personne en particulier. A chaque fois qu'un groupe de personnes déploie une technologie qui présente un risque, même minime, de défaillance, ce risque est voué, au bout d'un nombre d'usages suffisants, à produire des conséquences dommageables. Dans ce genre de contexte, nous sommes amenés à poser la question classique « A qui incombe la faute ? ». Pourtant, si ce réflexe de chercher un coupable peut être sain dans beaucoup de situations, afin d'attribuer correctement le blâme, et, dans un point de vue « forward-looking » d'espérer qu'une telle situation ne se produise pas à l'avenir, il est tout autant nécessaire de reconnaitre qu'il arrive que des situations dommageables se produisent sans que personne ne soit à blâmer. C'est en quelque sort ce qui se produit quand l'IA est à blâmer. Nous savons lorsque nous la déployons qu'elle n'est pas un agent moral parfait, aussi il est normal qu'elle échoue. Nous n'avons aucun mal à accepter qu'il arrive que des humains échouent à agir selon une moralité acceptable. Evidemment, si les actes de tels

humains nous causent du tort, accepter leur imperfection peut être bien plus délicat, il en va simplement de même avec l'IA.

Cependant, il est possible de remonter la chaine des actions effectuées pour remonter à l'origine de l'action immorale, et si on le fait, on retrouvera le propriétaire de l'IA, ou bien son concepteur, son programmeur. L'argument que nous allons ici développer, c'est que chacune de ces personnes que l'on tient partiellement responsable d'avoir agi de la sorte a, en réalité, pris un risque. Ce risque, c'est celui de programmer une IA, c'est celui de déployer une IA dans le monde. Si nous reprenons l'exemple de la voiture autonome, son propriétaire, qui l'utilise, choisit de la déployer sur les routes, et choisit en même temps que cela de prendre le risque qu'elle commette une erreur. Cependant, comme nous le faisons dans l'approche par prise de risque, ce risque ne peut être évalué qu'en comparaison avec toutes les autres conséquences potentielles de cette décision. Ceci nous permet d'affirmer que tant qu'une IA cause moins de torts qu'elle n'en évite, l'équilibre moral de l'action même de la mettre en fonction est respecté. Cependant, il y a un vide de responsabilité qui se crée sous prétexte de vies qui sont sauvées par des accidents évités[71]. Parce que certains accidents auraient eu lieu si l'IA n'avait pas été déployée, son propriétaire s'absout de la responsabilité de tel accident qui a effectivement eu lieu. S'il y a une cohérence à un tel raisonnement, il n'empêche qu'il doit être correctement évalué à la mesure d'une ultime possibilité, celle tout simplement de refuser d'utiliser des automobiles. Lorsque nous utilisons une voiture, qu'elle soit autonome ou pas, nous acceptons que notre transport a plus d'importance que le faible risque d'accident. Cela ne va pas forcément de soi. C'est pourquoi si nous commettons une erreur et provoquons un accident alors que nous sommes au volant de notre véhicule, nous sommes prêts à être tenus responsables de nos actes et devons alors assumer, légalement, financièrement et moralement, les dommages que nous avons causés. Il en va de même pour la voiture autonome, ainsi que pour tout type d'IA. Le déploiement d'une IA devrait s'accompagner d'une assurance capable de couvrir les risques d'accident. En réalité, cette assurance pourrait s'effectuer au niveau global de l'ensemble des propriétaires de véhicules autonomes. Il est injuste que le malchanceux d'entre eux dont la voiture se rend responsable d'un accident soit le seul à payer alors que l'approche par prise de risque nous montre que chaque propriétaire est tout aussi partiellement

[71] Cet exemple est propre à la voiture autonome mais il peut être étendu à tout type d'IA.

responsable que celui dont le véhicule a provoqué un accident. Notons que cette réflexion réduit la question de l'agence de la voiture autonome au statut de simple outil susceptible de dysfonctionnement, néanmoins si une telle solution suffit dans ce contexte, il n'est pas nécessaire, dans ce contexte, d'en attendre plus.

Pour en attendre plus d'un agent artificiel, il semble qu'il faille l'exposer à des données plus variées et moins ciblées. Précédemment, nous avons discuté les objectifs de vie d'un agent moral, qui expliquent pourquoi il prend telle ou telle décision. Dans le cas d'un agent artificiel, ce genre de question pose problème parce que la raison d'être de l'IA ne lui appartient pas, elle a été décidée par un être humain. Nous n'attribuons pas à l'IA le concept de dignité, que Kant attribue à l'humain, à savoir que nous devons traiter l'humanité en chacun comme une fin en soi, jamais simplement comme un moyen. Or l'IA est précisément pour nous tout d'abord un outil. Nous avons en objectif en tête lorsque nous la mobilisons, et nous la façonnons avec cette direction à l'esprit. Nous lui donnons accès à des données particulières, assignons des félicitations et des blâmes à des métriques bien précises. Tout ceci restreint fondamentalement la capacité d'une IA à être un agent moral au sens le plus fort.

Si nous souhaitions voir naître un agent moral artificiel complet et authentique, il serait nécessaire de résoudre certains de ces problèmes. Le premier problème concerne l'incarnation. Afin de donner un rapport interpersonnel à son environnement à l'IA, il serait nécessaire de l'incarner dans un robot particulier, qui serait le point de départ de son rapport au monde. Par ailleurs, les seuls objectifs que nous devrions lui attribuer seraient simplement de subsister à ses besoins immédiats. Il s'agirait de lui inculquer le besoin d'avoir accès à une source d'énergie, et de lui faire comprendre que si elle n'est pas capable de garantir sa survie, son chemin s'arrête là. Une forme de sélection artificielle imitant la sélection naturelle pourrait stimuler ce genre de procédé. Les IA incapables de garantir leur survie seraient éliminées et celles capables de le faire seraient reproduites, en mobilisant les techniques de « machine learning » mobilisant des « algorithmes génétiques ». En dehors de sa propre survie, aucune raison d'être ne devrait alors être attribuée artificiellement à l'IA, si ce n'est peut-être certains impératifs moraux, ou certains objectifs de recherche de connaissance. Seule une telle expérience pourrait rêver l'émergence d'un véritable agent artificiel moral complet. Evidemment, une telle question est hautement prospective et n'est ici qu'une expérience de pensée. Nous devons

cependant garder à l'esprit que les limites de l'IA sont aussi les limites que nous lui imposons par notre volonté. Tant que nous souhaitons que l'IA ne soit qu'un simple outil, et que nous agissons de telle sorte qu'elle le reste, elle va inévitablement se conformer à cette structure. Ce n'est que si nous souhaitons donner naissance à de véritables agents moraux complets, presque même libres, que cela devient une possibilité. La question de savoir si cela se fera un jour reste ouverte, mais au vu des limites techniques que l'on connait aujourd'hui à l'IA, autant elle est déjà omniprésente dans nos vies, autant les vues apocalyptiques de certains penseurs de la « singularité » semblent ne pas prendre en compte certains prérequis fondamentaux à son émergence.

Conclusion

La troisième partie de ce travail représentait l'aboutissement de celui-ci, à la jonction des deux parties précédentes. Aussi, dans une certaine mesure, la conclusion de cette ultime partie était en quelque sorte la conclusion de ce mémoire. A ce titre, nous ne reprendrons pas les éléments qui y ont été évoqués, mais nous allons attarder notre attention sur une note un peu plus subjective, un peu plus prospective. Alors que nous avions introduit ce mémoire en montrant comment de nombreux questionnements comme celui de l'agence morale étaient pensés par des humains, pour des humains, contaminés d'anthropocentrisme, nous arrivons au terme de celui-ci avec un constat surprenant. Si aujourd'hui l'être humain est sans doute l'agent moral le plus apte que nous connaissions, s'il reste celui qui est le plus capable de raisonner sur des questions morales et de formuler des théories morales, il est possible que ce ne soit pas toujours le cas. Nous connaissons les limites à la perfection de l'agence morale humaine, et nous sommes en droit de nous demander si de telles limites ne garantissent pas que des agents artificiels finissent, inévitablement, par nous surpasser. Si l'on se contente d'une certaine approche de la neuroscience qui réduit notre esprit à la simple interaction neuronale, il n'y a aucune raison de penser qu'une IA ne puisse pas faire tout ce que nous faisons et bien plus, elle qui est libérée de nos contraintes biologiques. Evidemment, une telle approche est en elle-même sujette à débat. Nous avons d'ailleurs vu au cours de ce travail que ce sont peut-être bien ces contraintes biologiques elles-mêmes qui sont à l'origine de notre prise sur la moralité. Sans pour autant adopter toutes les conclusions auxquelles une telle approche mène, notons que certains vont jusqu'à dire que l'empathie, l'altruisme et toute la construction de la moralité trouvent son origine dans la sélection naturelle, via un mécanisme de sélection de groupe[72].

Il ne fait cependant aucun doute que nos souffrances, notre vulnérabilité, trouvent son origine dans notre nature biologique. Nous sommes des êtres vivants finis, nous avons une durée de vie limitée et des besoins biologiques contraignants. Nous sommes localisés, dans l'espace et dans le temps, dans un corps de chair et d'os. Nous avons des désirs et des envies, qui sont autant de preuves de notre nature biologique. Tout ces éléments nous

[72] Nous ne défendons pas ici cette théorie, ni toutes les conclusions auxquelles elle aboutit simplement parce que cette théorie ne fait pas forcément consensus parmi les théoriciens de l'évolution. Nous l'avons nommée parce qu'elle offre une perspective qui a au moins le mérite de donner à penser.

différencient drastiquement et irrémédiablement de l'IA et des agents artificiels. Mais si notre vulnérabilité trouve son origine dans notre corps, nous sommes en droit de nous demander s'il serait possible d'en révéler les secrets. Si c'était le cas, si nous étions capables de comprendre de manière fondamentale le fonctionnement de la souffrance humaine, le fonctionnement de la peur, de la souffrance, mais aussi du bonheur, de la joie, de l'espoir, alors nous pourrions rêver l'enseigner à un agent artificiel, qui se retrouverait alors doté d'une véritable capacité de compréhension de ce qu'est la nature humaine. On dit souvent de l'empathie qu'elle trouve sa source dans l'activité de neurones particuliers de notre cerveau que l'on nomme « neurones-miroirs » et qui ont, comme leur nom l'indique, pour rôle d'imiter la souffrance d'autrui, afin de nous donner un aperçu de cette souffrance que nous puissions comprendre. Si de tels neurones-miroirs pouvaient être construits, traduits en algorithmes et intégrés à la machine, peut-être alors que celle-ci se trouverait capable de comprendre la complexité des émotions humaines. Il est assez étrange d'affirmer que si l'on veut qu'une machine soit véritablement capable d'empathie, il est nécessaire qu'elle soit capable de souffrir. Cela n'a rien de surprenant, si l'on en croit cette citation que d'aucuns attribuent à Dostoïevski : « La souffrance est l'unique cause de la conscience ».

Bibliographie

- **BAILEY, C.**, « Le double sens de la communauté morale : la considérabilité morale et l'agentivité morale des autres animaux » in *The Ethics Forum*, Vol.9, n°3, (2014).

- BRINGSJORD, S and GOVINDARAJULU, N.S., « Artificial Intelligence » in *Stanford Encyclopedia of Philosphy*, (2018).

- **DIGNUM, V.**, « Responsible Artificial Intelligence: Ethical Thinking by and about AI », *Delft university of Technology*.

- **HIMMA, K.E.**, « Artificial agency, consciousness, and the criteria for moral agency: whatproperties must an artificial agent have to be a moral agent? » in *Ethics and Information Technology*, (2009).

- **HURSTHOUSE, R.**, « Virtue Ethics » in *Stanford Encyclopedia of Philosphy*, (2016).

- **KANT, I.**, « Fondements de la métaphysique des Mœurs », (1785).

- **LEVENSON, R.M., et al.**, « Pigeons (Columba livia Columba livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images » in *Plos*, (08/2015).

- **LIU, X., et al.,** « A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis » in *Lancet Digital health*, (09/2019).

- **MOOR, J.H.**, « The Nature, Importance, and Difficulty of Machine Ethics » in *Machine Ethics*, (2006).

- **ROCHA, J.**, « Autonomy Within Subservient Careers » in *Ethic Theory Moral Prac*, (2011).

- **SULLINS, J.P.** , « When is a robot a moral agent » in *International Review of Information Ethics*, Vol.6, (12/2006).

- **TALBERT, M.**, « Moral Responsibility » in *Stanford Encyclopedia of Philosophy*, (2019).

- **WALLACH, W., et al**., « Consciousness and Ethics: Artificially conscious moral agents » in *International Journal of Machine Consciousness*, Vol.3, No. 1, (2011).

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté de philosophie, arts et lettres
Place Blaise Pascal, 1 bte L3.03.11, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/fial**