L'éthique des intelligences artificielles et le contrôle de leurs impacts sociaux

Nathan Lauga

PRÉ-MÉMOIRE

2018 – 2019

Avant-propos

En préalable, je tiens à préciser que ce pré-mémoire ayant pour but d'être lisible (au moins partiellement) par tous, aucune maîtrise dans le domaine de l'intelligence artificielle ou de la morale ne sont requis pour le lire, bien que cela puisse aider pour certains passages. Le domaine de recherche alliant intelligence artificielle et éthique (morale) étant récent, il est possible que des recherches présentées dans ce document soient incomplètes voire obsolètes dans un futur plus ou moins proche.

Ce pré-mémoire s'inscrit dans le cadre de ma scolarité au sein de l'école Ingésup Bordeaux du campus Ynov ainsi que dans le cadre de mon alternance dans l'entreprise Caisse d'épargne Aquitaine Poitou Charentes, où j'y travaille en tant que Data Scientist dans un service d'informatique décisionnelle. L'objectif de ce document est d'introduire les concepts sous-jacents de la problématique soulevés par ce mémoire. De plus, les travaux qui viendront dans la seconde partie (suite du pré-mémoire) seront soutenus pour septembre 2020. Afin de permettre une appropriation cohérente de mes travaux l'année prochaine, une publication liant ce document et celui de l'année prochaine sera mise au propre pour la même date de septembre 2020.

Le lecteur, non familier de la littérature scientifique et de sciences humaines, pourrait être surpris par certaines dates de référence pouvant être trop récentes pour certains auteurs. En effet, les dates citées correspondent à celle de l'édition, e.g. (Kant, 2006) ne signifie pas que Kant a publié un livre en 2006, mais que l'ouvrage qui est référencé est une édition de 2006.

Les propos discutés ici pouvant traiter de questions d'ordre sociales ou encore morales, il est important de préciser que j'ai cherché à les détailler de manière objective, mais étant moi-même soumis à ma vision idéologique du monde, je ne peux garantir une parfaite impartialité sur le fond des écrits (e.g. la moralisation d'algorithme est essentielle à mes yeux).

En dernier lieu et ce, avant de commencer, je tiens à remercier tous ceux qui m'ont accompagné et encouragé dans cette démarche, au premier rang desquels, mon tuteur au sein de mon entreprise, Pascal Fournier, qui m'a permis de trouver un juste équilibre dans mes travaux et la possibilité de poser la question de l'éthique dans une banque. J'inclus également dans ces remerciements Julien Dufossez qui m'a aidé dans la construction de ma problématique et qui m'aura alimenté en articles d'actualités portant sur la thématique de mon mémoire, à mes parents apportant un œil critique et premiers relecteurs de ce travail, à Julien Dauliac m'ayant accompagné sur la recherche théorique de l'IA et éthique et apportant de l'aide sur la réflexion sur mes écrits et enfin à Clément Romac et Pierre Leroy qui m'ont permis de découvrir le domaine de recherche de l'intelligence artificielle et de l'éthique.

Table des matières

Table des figures

1. Introduction

Les intelligences artificielles, de nos jours, suivent une croissance exponentielle, autant sur la quantité que sur la complexité des tâches réalisées. Certaines réalisent des exploits qui surpassent les humains (e.g. le jeu d'échecs). Leurs progrès sensationnels, permettant par exemple la reconnaissance faciale, ajoutent une nouvelle question sur la balance des IA, celle de l'éthique.

En effet, le paradigme initial qui est la représentation qu'elles se font du monde, peut être soumis aux mêmes problématiques qu'un humain, entre autres les stéréotypes implicites catégorisant négativement ou positivement une classe sociale.

Les composantes formant le paradigme des algorithmes, soit leurs environnements, sont de nos jours très souvent sélectionnées par la main des humains. L'éducation morale des enfants est considérée comme logiquement reconnue pour norme sociale, mais quand il s'agit d'éducation de l'éthique pour une machine, l'idée semble de suite moins représentable.

Au croisement de la sociologie, de la philosophie et du domaine de recherche de l'intelligence artificielle s'ancre alors un cheminement conduisant vers la recherche sur l'éthique de l'intelligence artificielle.

1.1 Problématique

Définir une morale pour un algorithme n'est pas chose aisée, cela soulève même une interrogation qui revient à demander ce qui est bon. La question à se poser lors de la conception d'IA est bien plus complexe que d'une simple question de performance.

En effet, aujourd'hui les données affluent tel un courant d'eau suivant son cours. Ces mêmes données étant un nouveau pétrole non raffiné pour les intelligences artificielles, l'importance de bien maîtriser ce qu'elles laissent entrevoir sur l'aspect social est incontestable.

Autrement dit, dans un contexte où l'intelligence artificielle est omniprésente et que les données deviennent le nouvel or, est-il possible de moraliser les algorithmes afin de les rendre transparents et par conséquent de limiter leurs impacts sociaux ?

1.2 Plan

Ce document est organisé en deux chapitres, de plus, il s'agit là d'un pré-mémoire, sa construction suit donc la logique d'apporter les informations pertinentes pour la suite du mémoire qui sera réalisée pour l'année scolaire 2019-2020 et soutenu en septembre 2020.

Le premier chapitre constitue un état de l'art sur les concepts d'intelligence artificielle et d'éthique. Tout d'abord, la notion d'IA est détaillée au travers de son histoire mouvementée et le fonctionnement actuel des algorithmes. La suite s'attache à rentrer dans les réflexions « futuristes » avec l'idée d'une super-intelligence.

Ensuite, c'est le concept d'éthique qui est expliqué principalement sur le plan philosophique. D'abord l'origine, puis les différentes et plus célèbres morales existantes à l'heure actuelle. La fin du chapitre se concentre sur la cohabitation de ces deux concepts.

Le second chapitre présente la méthodologie qui sera employée afin de répondre à la problématique. Des approches mathématiques existantes sont détaillées en premier lieu. Puis il est présenté les deux méthodes choisies pour répondre à la problématique du mémoire, qui sont respectivement les entretiens et une approche technique accompagnée d'une feuille de route.

2. IA et éthique : une association forcée

2.1 L'intelligence artificielle, quand les machines révolutionnent le monde

L'intelligence artificielle, un mot qui, depuis quelques années, semble être sur toutes les bouches. Bien que l'imaginaire collectif puisse entrevoir des machines humanoïdes capable de détruire l'homme (e.g. Terminator[1]), les technologies actuelles ne nous offrent pas encore un spectacle digne d'Hollywood. Évidemment, le terme en lui-même donne une illusion humaine puisque le mot « intelligence » est présent. Il faut tout de suite dissocier cette intelligence dite « artificielle » à celle qui est associée aux hommes, l'intelligence dite « naturelle ».

En informatique, la recherche sur l'intelligence artificielle est définie comme l'étude des « agents intelligents », soit n'importe quel appareil qui perçoit son environnement et prend des décisions qui maximisent ses chances d'atteindre son objectif (Poole, Mackworth, et Goebel, 1997). Un exemple illustre cette définition : dans les jeux d'échecs, un agent intelligent pourra, en connaissant les règles du jeu, effectuer des coups et son objectif, qu'il cherchera à atteindre, sera de battre son adversaire.

2.1.1 Un commencement agité

Le terme qui aujourd'hui~~,~~ est très évocateur~~,~~ a été utilisé pour la première fois en 1956 par John McCarthy, lors de la conférence de Darthmouth, conférence qui est considérée comme l'acte de naissance de l'intelligence artificielle en tant que domaine de recherche autonome (Solomonoff, 1985). Bien sûr, cette notion bien que sans nom précis jusqu'à cette conférence, n'était pas inexistante.

L'apparition de technologies telles que les ordinateurs, Internet, la reconnaissance vocale et d'autres concepts révolutionnaires, soit un système qui pourrait amplifier le savoir et la connaissance de chacun a été prédit juste avant que les bombes atomiques ne tombent sur le Japon (Bush, 1945). Les technologies modernes confirment bien cette hypothèse.

Une idée de machines intelligentes se propage ensuite notamment au travers du terme cybernétique en tant que théorie entière de la commande et de la communication, aussi bien chez l'animal que dans la machine (Wiener et al., 2014).

Le concept d'une technologie, qui pourrait servir l'homme jusqu'à asservir ses besoins intellectuels, est donc apparue, mais cela reste un peu vague et ne ressemble pas à une notion de machine intelligente au même niveau que l'homme. C'est cinq ans plus tard, en 1950, qu'Alan Turing[2], considéré comme le père de l'intelligence artificielle, publie un article qui révolutionna le monde de l'informatique (Turing, 1950).

En résumé, il a soulevé la question désormais célèbre « Can machine think ? »[3]. Cette interrogation est très contradictoire, surtout en 1950, puisque le terme « machine » et « penser », ne peuvent être

[1] Le film « Terminator », réalisé par James Cameron et sortie en 1985, représente un cyborg du futur envoyé par des robots, qui a pour mission de tuer Sarah Connor, mère du chef des résistants humains.
[2] Alan Turing, étant également considéré comme le père de l'informatique, a notamment donné son nom au « Prix Turing », récompensant une personne chaque année, pour sa contribution au monde de l'informatique.
[3] Traduction : Les machines peuvent-elles penser ?

définis d'une façon qui puisse satisfaire tout le monde. Afin de résoudre le conflit de cette contradiction, Turing a proposé une solution, élégante, étant le fameux « Test de Turing ».

Le Test de Turing, est construit de la façon suivante : si une machine peut tenir une discussion avec un humain (au travers d'une messagerie par exemple), sans que la femme ou l'homme puisse distinguer qu'il s'agisse d'un humain ou d'une machine alors la définition du test dira que cette machine est « pensante ». Il s'agit d'une proposition très importante dans la philosophie de l'intelligence artificielle (Pinar Saygin, Cicekli, et Akman, 2000).

Au travers de la notion révolutionnaire d'une machine dite « pensante », un engouement est naturellement apparu autour des machines, des déclarations chocs sont faites comme « d'ici dix ans un ordinateur sera le champion du monde des échecs » (Simon et Newell, 1958) ou encore « des machines seront capables, d'ici vingt ans, de faire tout travail que l'homme peut faire » (Simon, 1965). Il n'est pas le seul à faire de telles affirmations et celles-ci mènent à une attente très élevée concernant les possibilités des algorithmes intelligents, comme souvent lorsque les attentes sont élevées une phase de déception s'ensuit.

Bienvenue dans le premier hiver de l'histoire de l'intelligence artificielle. Comme une bulle qui aurait éclaté, la recherche a ralenti d'un coup ainsi que le budget consacré au domaine. Les causes en sont multiples. Il est possible de retrouver entre autres, la limite de la puissance de calcul ou encore le manque de base de connaissance du monde par les ordinateurs (manque de données). En effet, les travaux, qui portaient sur le langage naturel, ne pouvaient pas être extrêmement poussées puisque le stockage de la mémoire la limitait à vingt mots (Crevier, 1992).

Pour beaucoup ce secteur a été enterré. Arrivèrent les systèmes experts, programmes qui allient algorithme et connaissance métier. Ce concept qui comme le Soleil au printemps fit fondre la neige du premier hiver.

L'histoire se répéta malheureusement dans les années 90 : trop d'attente pour un réalisé en dessous de l'imaginaire. Conséquence, une nouvelle période froide dans ce domaine et un désenchantement populaire.

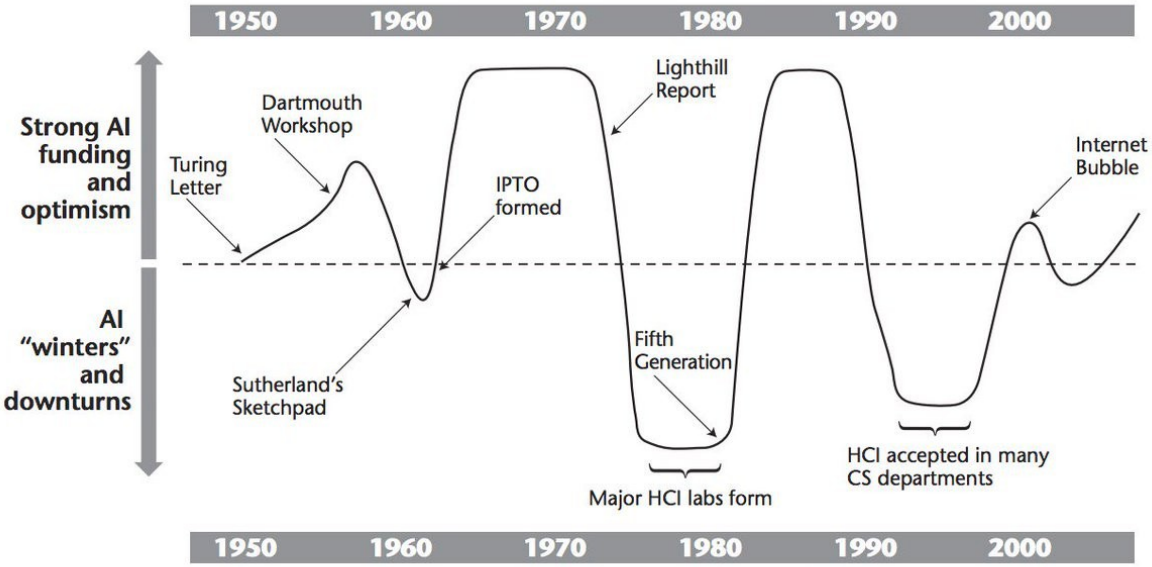

Figure 1: Le changement des saisons de l'IA (Grudin, 2009)

Ci-dessus un résumé des débuts de l'histoire de l'intelligence artificielle avec en abscisse les années et en ordonnées l'attente autour de ce secteur. Sur le graphique, certains évènements majeurs de l'histoire du domaine en question.

2.1.2 La machine plus forte que l'homme ?

Le froid qui s'était abattu sur l'intelligence artificielle peut laisser supposer que les pensées révolutionnaires n'auraient pu se réaliser, mais la machine, sans oublier les développeurs, ont plus d'un tour dans leur sac. Dans les années 1990, IBM sortit une machine nommée « Deep Blue ». Elle a été construite dans le but de jouer aux échecs et surtout de battre le meilleur joueur d'échec du monde. Après un échec en 96, la surprise fut totale quand en 1997, le vainqueur ne put lever les bras pusiqu'il s'agissait de Deep Blue (Krauthammer, 1997). L'imaginaire d'une machine plus intelligente que l'homme refit surface dans certains esprits.

Malgré cet exploit, la force de cet algorithme était « juste » de pouvoir calculer très rapidement en parallélisant les calculs et donc d'anticiper les coups de son adversaire plus facilement que le cerveau humain (Hsu, Campbell, et Hoane, 1995). Il est possible de comparer cela à une calculatrice appliquant une multiplication complexe et affichant la réponse en un claquement de doigts, alors que pour un homme cela est une tout autre histoire.

La révolution qui a mené l'intelligence artificielle à sa place aujourd'hui n'est pas uniquement dû à une grande bête de calcul, la raison principale est le « machine learning » ou en français l'apprentissage de la machine.

Un bébé regardera ses parents, sa famille et son entourage pour apprendre à marcher, un algorithme de « Machine Learning », lui n'a pas d'œil pour regarder ,mais du code pour analyser les données fournies par les développeurs. Un sortilège utilisant pour baguette magique, des formules

mathématiques et statistiques permettant alors de généraliser à partir de ce que l'intelligence artificielle a en base d'apprentissage.

Cette catégorie d'intelligence artificielle est aujourd'hui la grande dominante du marché. Elle se découpe en différentes catégories, l'une des plus utilisées est l'apprentissage supervisé consistant à donner à la machine des données dites « labellisées », soit qui possèdent déjà la réponse à la problématique (e.g. prédire la météo en utilisant les données des années précédentes en indiquant en label le temps qu'il faisait réellement). Ensuite, l'apprentissage non supervisé, lui se distinguera puisqu'il n'y aura pas de label identifié, l'objectif est de trouver des patterns, de segmenter ou de détecter des anomalies.

Au travers de ces deux catégories, une vaste quantité d'algorithmes existent, de la simple régression linéaire, en passant par le traditionnel arbre de décision, la catégorie d'algorithme ayant permis de faire un grand pas, sans discussion possible, est celle des réseaux de neurones.

Un mot assez frappant, puisqu'il associe directement des machines à l'homme, plus précisément au cerveau humain. Or, non en dehors de la forme, ces réseaux de neurones ne sont pas du tout le cerveau humain. La logique est similaire : des « neurones » prenant des informations en entrée pour en ressortir une autre information à son tour vers un niveau suivant. Bien que cela puisse raviver l'illusion d'une machine aussi intelligente que l'homme, les algorithmes actuels peuvent être décrits comme intelligences artificielles « stupides » car elles n'appliquent, que ce pour quoi elles ont été créées, avec des niveaux de performances parfois, bluffant. La déclaration d'Andrew Moore (newsflash, 2018), responsable de Google Cloud AI, le confirme « AI is currently, very, very stupid »[4].

L'un des portraits qui peut le mieux décrire la force de frappe de cette génération d'IA, peut être associée à l'histoire de la défaite d'un homme au jeu de Go. Le jeu de Go est l'un des jeux considérés comme les plus complexes au monde, soit un jeu qui pour l'homme devait lui rester en main pour de longues décennies après la montée des IA pour les échecs.

Une entreprise spécialisée dans la création d'algorithmes intelligents, DeepMind, s'est lancée dans ce défi au cours de la décennie actuelle. Son premier « enfant », baptisé « AlphaGo », s'entraîna sur des parties jouées par des joueurs professionnels de Go. En 2016, elle se présenta en Corée du Sud dans le but d'affronter le champion du monde Lee Se-dol en cinq parties. La fin de l'histoire fit un bis repetita avec Deep Blue et AlphaGo triompha de l'humain (DeepMind, 2016).

Cette victoire fit des titres sensationnels dans les médias, mais le plus spectaculaire n'est pas cette victoire. Un peu plus d'un an plus tard, DeepMind accoucha de la sœur (ou frère, le genre n'a que peu d'importance) cadette d'AlphaGo : AlphaGoZero. Le nom n'est pas anodin au vu de l'apprentissage de cette machine. Contrairement à l'algorithme « champion du monde », la logique n'est pas de se baser sur des matchs existants, mais de lui apprendre les règles du jeu, puis de laisser l'algorithme s'affronter tout seul pendant un certain temps, jusqu'à par lui-même redécouvrir les coups des débutants, les stratégies que les humains ont mis un millénaire à apprendre pour enfin complètement dépasser ces méthodes presque moyenâgeuses (Silver et al., 2017). Pour l'anecdote AlphaGoZero fut triomphant d'AlphaGo sur un score de cent matchs à zéro.

Cet algorithme basé sur une méthode d'apprentissage non supervisé a continué d'être utilisé par DeepMind et d'autres entreprises pour aller jusqu'à, récemment, battre des joueurs professionnels de « StarCraft II », un jeu vidéo (Vinyals et al., 2019).

[4] Traduction : l'IA est actuellement, très, très stupide.

Ce récit, qui dicte des exploits qualifiables de surhumains, n'est qu'un exemple sensationnel et qui a eu de la popularité en dehors de la communauté centrée autour des algorithmes intelligents. Créer des musiques (Medeot et al., 2018), créer des peintures au format numérique (Mordvintsev, Olah, et Tyka, 2015), reproduire des visages humains (Karras, Laine, et Aila, 2018), imiter un discours d'un président américain (Suwajanakorn, Seitz, et Kemelmacher-Shlizerman, 2017), ses tâches ont bel et bien été accomplies par des intelligences artificielles.

Tout ceci a fait resurgir un sentiment similaire aux périodes précédents les deux dernières périodes hivernales que l'IA a connu, un sentiment de machines bien plus performantes que l'homme, pouvant être même plus intelligentes que l'être humain. Or, comme évoqué précédemment, elles ne réalisent que ce pourquoi elles ont été créées. De plus, aucune n'a réussi à passer le test de Turing, qui semble être la mesure qui permettra de déterminer leur intellect « équivalent » au nôtre. Cette période, appelée singularité[5], est-elle pour bientôt ?

Penser qu'une machine aussi intelligente que l'homme est pour très bientôt ou très longtemps, est digne d'un fantasme dans l'imaginaire collectif. Il ne s'agit pas là forcément de la meilleure opinion afin d'avoir la meilleure prédiction à la question du passage de la singularité. Des chercheurs d'Oxford et de Yale ont décidé d'interroger un grand nombre de chercheurs du domaine de l'intelligence artificielle du monde entier afin de déterminer une approximation sur la singularité (Grace et al., 2017). Les résultats sont très variés, mais une courbe a émergé de leur sondage :

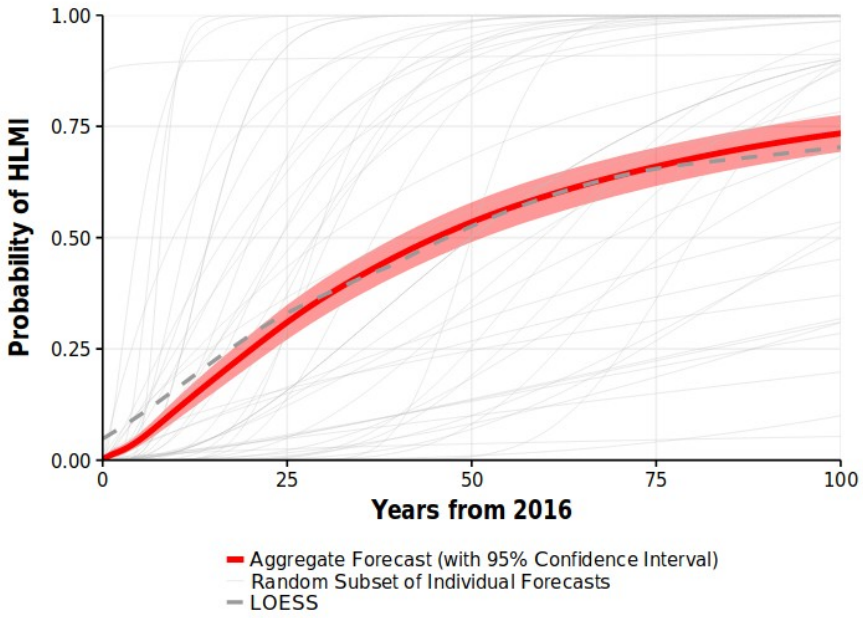

Figure 2: Probabilité d'une super-intelligence à partir de 2016 (Grace et al., 2017)

Ce graphique détaille la probabilité d'une « HLMI » soit « High-level machine intelligence » synonyme de la singularité à partir de l'année 2016. Les traits gris dans le fond correspondent à des réponses individuelles et il est remarquable de voir leur écart. L'information à retenir sur ce graphique est donc la ligne rouge indiquant une prévision agrégée qui nous indique une probabilité dépassant les cinquante pourcents d'ici cinquante ans.

[5] Ce terme décrit le moment où une intelligence artificielle aura atteint les capacités intellectuelles de l'Homme.

De plus, cette même recherche a posé des questions plus diverses comme notamment la possibilité qu'un humain soit battu au jeu de Go et la réponse moyenne était que cette tâche serait réalisée d'ici 2028, pourtant moins d'un an après la publication de cette recherche AlphaGo a triomphé. Sur la figure suivante, il est possible d'observer que les questions sont posées sur le remplacement de femmes et hommes et les deux questions en haut pousse la réflexion jusqu'à une IA pouvant rechercher pour améliorer les algorithmes intelligents créant un parallèle avec l'humain et la médecine. Puis en dernier, la question posée s'axe sur l'automatisation totale du travail humain par les intelligences artificielles.

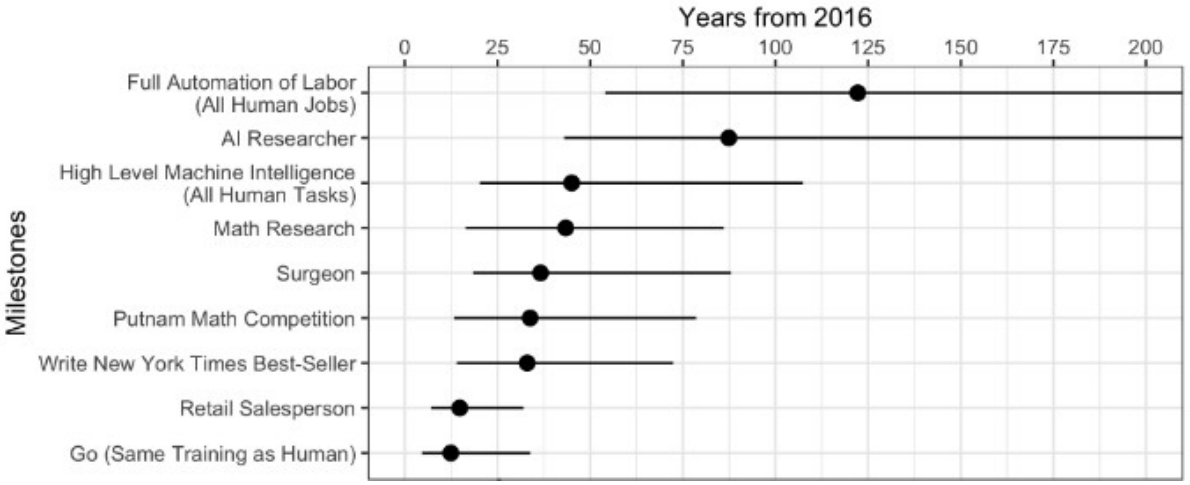

Figure 3: Chronologie des estimations qu'une IA achève des tâches humaines (Grace et al., 2017)

Pousser la réflexion à la question de ce que sera le monde dans plusieurs décennies est déjà suffisamment complexe sur des questions pouvant être plus fondamentales (écologie, collapsologie, social, consommation, etc.). Cela n'empêche pas d'y réfléchir : les IA seront-elles nos sauveuses ? Notre talon d'Achille ? Ou bien vivront-elles une vie à nos côtés sans changer nos habitudes ? Entre pessimisme et utopie, chacun peut se faire son avis, rien n'est encore écrit.

2.2 L'éthique, la science de la morale

Le bien, le mal, deux notions souvent assez abstraites, mais que chacun assimile telle une doctrine de conduite qui nous inspire. Comment sont-elles définies ? Une question sur laquelle la notion de morale est présente. Issu du latin « moralis » signifiant « relatif aux mœurs », ce mot s'inscrit au centre de la lutte entre le bien et le mal relatif à chaque individu pour créer des formes de normes morales en société. Cela affirme nos devoirs, droits ou encore interdits (au-delà même des lois).

Le code de conduite qui récite ce qui est de l'ordre du devoir pour femmes et hommes, aussi appelé éthique (synonyme de morale), est apparu d'un besoin de coopération à l'époque des chasseurs cueilleurs. En effet, l'égoïsme favorise l'individu et non le groupe, or dans des temps où la survie passait par le social, par le groupe un comportement individualiste attiré plus de sanction d'un point de vue de survie (Harari et Dauzat, 2015).

Un terme assez générique donc, mais souvent porté dans l'illusion d'un absolu, plus précisément, la morale est parfois pensée comme une morale universelle. Or à la question : « Y a-t-il une morale universelle ? » (signifiant qui vaut en tout temps et en tous lieux), la réponse semble être négative.

L'ambiguïté de cette idée vient certainement de l'ethnocentrisme[6] (Sumner, 1906), une généralisation abusive de nos critères moraux. La morale est le reflet d'un contexte social et temporel. Un exemple frappant sera celui concernant le vol, perçu dans notre époque comme un acte voyou, du temps des Spartiates, le vol faisait partie de l'éducation des jeunes hommes dans le but de compléter leurs rations de nourriture (Ducat, 2017).

La notion d'universalité mise de côté, la question la plus adéquate sera de penser sur le plan de l'objectivité. Il faut quitter la dimension des morales à chacun et au travers des différentes notions d'éthiques, se poser la question du bon, du méchant, du mauvais ou encore du bien et du mal, en bref comment se compose une morale, en philosophie.

2.2.1 L'origine du bon

Le terme « bon », émet un jugement de valeur positif lorsqu'il est utilisé. Avec ce jugement, il est nécessaire d'y trouver son opposition, un linguiste de la Novlangue[7] dirait « Inbon », mais cela n'est pas le mot recherché en français. Deux termes semblent être de bons candidats : mauvais et méchant. Alors, lequel serait le plus judicieux ?

L'origine de la morale, donc de l'opposition au « bon » a été analysé par Friedrich W. Nietzche (Nietzsche, 1900) dans son livre « Généalogie de la morale ». Ce texte cherche l'origine à la fois historique et psychologique. De nos jours, lorsque qu'une action est dite bonne, il est souvent pensé qu'elle est altruiste soit bénéfique aux autres. Or dès le début du chapitre « Bien et mal », « Bon et mauvais », Nietzche en parle pour renier cette origine du bénéfice global : « le jugement ''bon'' ne provient nullement de ceux qui bénéficient de cette ''bonté'' ! ». Il détaille par la suite que le fondement du bon a été créé par « les nobles, les puissants, les supérieurs en position et en pensée ». Cette morale est l'expression de la puissance, de la force, elle célèbre soi-même. Ces « nobles », triomphants, posent ce qu'ils sont, ce qu'ils font comme valeurs « bonnes », c'est « la morale des maîtres ».

L'opposition de bon, dans cette morale, est alors le mauvais : celui qui veut être bon, mais qui ne peut pas. Cette vision de l'éthique peut paraître perturbante, mais l'expression « réussir dans la vie » évoque cette célébration de soi, cette opposition entre ceux qui sont bons dans leur vie et ceux qui échouent, qui sont mauvais. Bien qu'elle ne soit plus une morale qui soit majoritaire, des personnes semblent en être adepte.

Toujours dans le même chapitre de la « Généalogie de la morale », Nietzche évoque à répétition, une haine qui se produit contre les « maîtres », ceux qui se qualifient de bons. Les humains, qui dans une logique de morale de maître, seraient alors mauvais, y voient quelque chose de méprisable dans cette réussite puisqu'ils ne peuvent l'atteindre. C'est à partir de cela, dans le ressentiment, qu'une morale a émergé et a dominé la morale du bon et du mauvais, « la morale des esclaves ».

Elle se fonde sur la désapprobation des autres, de leurs actes, de leurs pensées, contre les méchants : ceux qui peuvent être bons, mais ne le veulent pas. En effet, le mauvais est devenu le bon et le bon est devenu le méchant.

---

[6] Terme introduit par William Graham Summer, sociologue, signifiant l'évaluation d'autres civilisations d'après des critères qui sont en réalité les nôtres, mais dont il est pensé qu'ils sont universels

[7] La novlangue ou en anglais « Newspeak » est la langue officielle d'Ocania inventée par George Orwell dans son roman 1984, son principe est de diminuer le nombre de mots afin de diminuer le nombre de concepts servant à la réflexion. La négation des mots se formule en rajoutant le terme « in » en début de mot.

Heureusement, il ne s'agit pas des seules morales sur Terre au XXIe siècle, elles ne semblent plus d'actualité. Cependant, leur existence est très importante sur le plan historique, de voir que les actions dites « bonnes » n'auront pas la même signification selon l'interlocuteur qui est en face. Mais alors, chacun possède-t-il une morale qui lui est propre ?

2.2.2 À chacun sa morale ?

L'idée d'une morale personnelle, comme chaque individu possède des goûts, s'appelle le relativisme moral, soit une vision subjective de la morale. Cette doctrine prône le fait que les valeurs morales ne peuvent être évaluées objectivement. Cela signifie que tous jugements moraux sont uniquement issus de la culture dans laquelle ils appartiennent. Ce mouvement s'oppose à ce qui est appelé le réalisme moral qui par opposition admet qu'il y a des valeurs morales objectives.

Dans cette vision du réalisme, une branche morale est assez populaire, celle de la déontologie. Venant du grec « deon » signifiant devoir soit la science du devoir, ce mouvement de pensée de la philosophie morale explique qu'il existe des devoirs moraux absolu soit sous la forme « tu dois », et ce, sans rajouter une explication à ce devoir. Ce principe fondamental, qui est appelé un impératif catégorique, est issu du philosophe Emmanuel Kant dans son ouvrage « Fondement sur la métaphysique des mœurs » (Kant et Delbos, 2007).

Bien qu'il y ait des désaccords sur le fond entre déontologues, la forme en reste la même : celle d'un « tu dois » absolu, c'est donc qu'il ne repose pas sur quelque chose de factuel et qu'il est difficile de pouvoir arbitrer sur des actes immoraux.

Alors, si le débat de l'arbitrage se déplace sur les faits dus à une action, aux conséquences engendrées, la morale qui s'appliquera sera le conséquentialisme. Elle se base sur les conséquences et sur le fait qu'elles seront négatives ou positives. Bien sûr, lorsque le terme « conséquence » est utilisé il s'agit des conséquences attendues et non réelles. La raison en est simple, il est souvent impossible de prévoir l'aboutissement d'une action bien après : si un enfant est sauvé de la noyade, comment prédire qu'il deviendrait un tueur en série trente ans plus tard.

Alors, comment juger moralement si une action est bonne ou mauvaise ? L'altruisme (une forme de conséquentialisme) y répond en établissant en prémisse de viser le bonheur de tous.

Une rivalité idéologique existe entre conséquentialisme et déontologie. En effet, pour les déontologues, l'idée est qu'un principe est bon donc qu'il est catégorique, pour un conséquentialiste, il y a des bonnes et des mauvaises situations et il faut faire le choix de la situation considérée comme bonne.

La forme pour savoir qu'elle est la bonne morale reste une matière à débattre forte intéressante, mais au travers de ces différentes visions de l'éthique, le fond semble à chaque fois rester cohérent et cela se voit bien aujourd'hui :un grand ensemble de lois sanctionnent quand il s'agit d'un acte tel un meurtre ou un vol. Le contenu, qui semble se traduire au travers de certaines lois, porte également sur des débats d'idées, c'est pour cela qu'une vision altruiste permet de raisonner de facto.

Tout ce raisonnement reste sur un plan philosophique. Il est important de savoir que chaque individu acquiert au cours de sa vie sa vision idéologique de la morale, son code d'éthique. C'est un processus qui commence dès la naissance avec notamment la théorie du développement moral (Kohlberg et Hersh, 1977), puisque notre appartenance à un groupe social défini déjà une partie de ce qu'il adviendra de nos croyances, puis l'éducation rentre en jeu, l'entourage, etc.

Chaque humain peut-être soumis à des biais[8] qui le feront agir pour une cause plutôt qu'une autre et il est possible d'interpréter ce qui compose sa morale, mais concernant une machine intelligente, si la singularité est atteinte, le questionnement sera de savoir si la femme ou l'homme en charge de sa création affectera l'idéologie de l'algorithme ou alors que la machine se fera sa propre vision idéologique en évoluant dans son environnement comme un enfant grandissant.

2.3 La machine intelligente et son éthique

Dans les parties précédentes, il a été question du concept de l'intelligence artificielle et également de la notion de singularité, puis la réflexion s'est focalisée sur la question de l'éthique, science de la morale. Le questionnement d'une morale au sein d'un algorithme intelligent peut avoir plusieurs aspects : l'éthique de la machine ou bien celle de la femme ou homme responsable de sa création, qu'il s'agit de la vision éthique de l'humain ou de la société en charge.

L'éthique au sein du domaine des IA, de nos jours, correspond à plusieurs notions : transparence du modèle[9] ou encore les biais présents dans l'algorithme. Cela étant d'actualité, un autre point essentiel est à évoquer, celui du futur. En effet, l'idée de la singularité pousse la réflexion à des considérations purement spéculatives sur les menaces existentielles de l'intelligence artificielle pour l'humanité (Villani, 2018).

Il y a alors deux aspects importants, celui du présent avec les algorithmes « boîtes noires » qui ne peuvent avoir strictement aucun sens pour les humains, mais aussi l'aspect du futur, avec des discussions et débats à avoir pour trancher sur une morale puisqu'il n'y a pas de morale universelle [10] et selon Laurent Alexandre (Alexandre 2017) cela soulèvera des conflits éthiques si terribles qu'il faudra bien expliciter cette morale.

2.3.1 Les biais et leurs méfaits

Chez les humains, les biais, dans l'époque moderne, sont souvent sujets à manipulation sans même que nous nous en rendions compte. L'un des plus importants est le biais de confirmation, une tendance à valider des arguments allant dans une idéologie similaire ou de rejeter ceux qui sont en opposition sans même s'attarder sur le fond. Le terme garde tout son sens lorsqu'il s'agit des biais au sein des intelligences artificielles. Trois types de biais pour les machines attirent l'attention.

Le premier, se baptisant le biais de sélection, il s'agit du manque de diversité des données prisent en entraînement. En effet, une IA cherchant à reconnaître si une personne est chef d'entreprise, si son entraînement se base sur une majorité d'hommes par rapport aux femmes, alors l'algorithme reconnaîtra plus souvent des hommes en tant que chef d'entreprise. Il existe des méthodes afin de réduire ce biais : avoir un jeu de données représentatif et proportionnel, ou alors de mettre des poids en fonction des données (Tran, 2017).

Le suivant se nomme biais d'interaction, son nom étant explicite, il correspond à un biais se formant au travers de l'interaction que les humains ont avec l'intelligence artificielle. L'exemple célèbre date de 2016, Microsoft sortit un compte Twitter sous l'appellation « @TayandYou » plus connu sous le nom TAY signifiant « Thinking about you »[11]. Son objectif était de converser avec les utilisateurs du

[8] Un biais fait référence à une déviation de la réalité, un moyen de contourner
[9] Un modèle est un algorithme d'intelligence artificielle (modèle mathématique, statistique)
[10] C.F. Partie 4.2
[11] Traduction : Pensant à toi

réseau social Twitter. Très rapidement, elle s'est mise à publier des messages à caractères homophobes ou encore antisémites. Le problème venait alors des internautes dialoguant avec elle, lui écrivant des messages politiquement incorrects (Tual, 2016).

Le dernier biais lui, correspond à un biais dû au passé : le biais implicite ou latent. L'analogie applicable à l'humain serait la notion de stéréotype, soit l'attribution inconsciente d'une qualité ou d'un défaut à une personne appartenant à une certaine catégorie sociale (Greenwald et Banaji, 1995). Au travers de ce biais, il est possible de retrouver des stéréotypes du genre, de la couleur de peau ou encore de l'appartenance à une catégorie sociale (e.g. jeune, adulte). Ce biais a notamment été repéré, aux USA, sur l'algorithme appelé COMPAS[12], pour lequel un criminel noir aurait deux fois plus de chance d'être considéré comme récidiviste par rapport à un blanc, alors qu'en réalité le taux de récidivisme entre noirs et blancs et approximativement le même (Larson et Angwin, 2016).

Ces quelques biais, démontrent déjà la complexité de créer un modèle « juste », si une intelligence artificielle serait plus transparente avec des informations sur comment elle fonctionne et expliquer le pourquoi d'une décision, cela permettrait déjà un premier pas vers une réduction des biais.

2.3.2 Ouvrir la boîte noire

Le terme « boîte noire » est très important, il ramène à un algorithme totalement opaque pour lequel tout ce qu'il est possible d'avoir est une sortie à partir de données fournies en entrée. Alors, la problématique de comprendre ce qu'il se passe à l'intérieur de ses machines est primordiale afin de pouvoir avoir confiance en elles.

Dans un premier temps, les concepteurs (souvent ayant des compétences en mathématiques) des algorithmes les plus utilisés de nos jours, pourraient très bien comprendre le pourquoi du comment des IA. La réalité ne conte pas cette histoire, en effet, pour certains types de modèles mathématiques, il est impossible d'interpréter ce qu'il se passe dedans. L'interprétabilité se définit par la description des éléments internes d'un système d'une manière compréhensible pour l'homme (Gilpin et al., 2018).

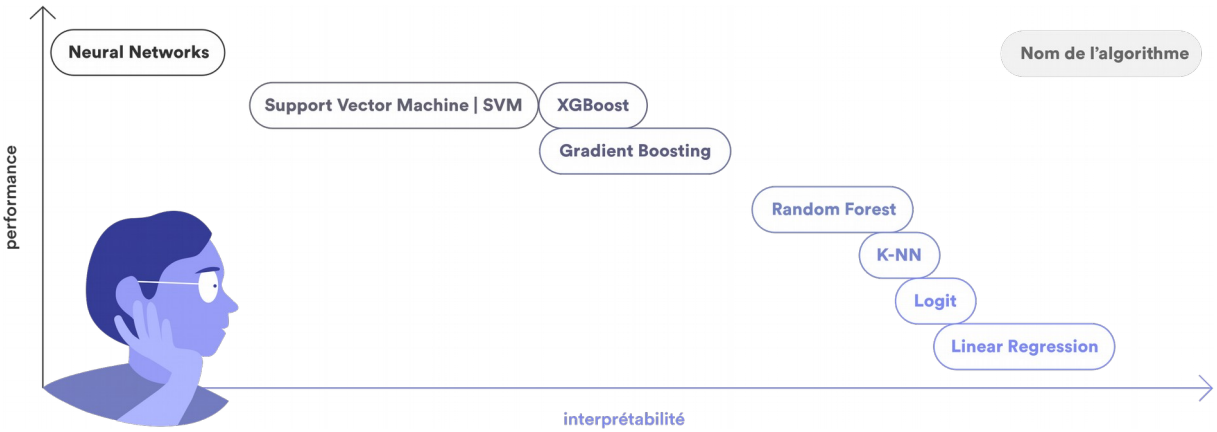

Figure 4: Interprétabilité d'un algorithme (Data for Good, 2018)

L'image ci-dessus illustre en abscisse l'interprétabilité d'un modèle et en ordonnée sa performance, la corrélation a noté est que, plus un algorithme est performant, moins il est interprétable. Ce graphique illustre bien une problématique éthique sur le plan de la transparence d'une IA.

[12] Correctionnal Offender Management Profiling for Alternative Sanctions : c'est une IA qui classifie le risque de récidivisme d'un criminel

Alors, la résignation d'avoir des algorithmes compréhensibles par l'homme est-elle nécessaire ? Non, rentre en jeu la notion d'explicabilité d'un modèle. Ce terme renvoi principalement à la question « Pourquoi ? », la capacité à répondre à ce questionnement lors d'une décision ou prédiction d'un modèle (Gilpin et al., 2018).

Concernant cette nouvelle problématique, l'explicabilité est plus facile à concevoir, pour reprendre la notion de boîte noire, il suffit de modifier les données en entrée pour voir comment cela influe sur le résultat obtenu par l'algorithme.

La transparence d'une intelligence artificielle passe également par la publication du code composant cette dernière publiquement, c'est-à-dire un code open-source[13]. Cela est bien problématique quand il s'agit d'algorithmes réalisés en entreprise et ne voulant pas publier leur réalisation au grand public.

Les explications à une non publication peuvent être diverses. Les plateformes comme Youtube ou Google ne publieront pas leur algorithme par nécessité d'éviter les abus afin de tirer profit d'une potentielle faille. OpenAI, eux bien qu'ils prônent cette idéologie de partager le code des IA afin de pouvoir mieux prévenir l'éthique des futures intelligences artificielles, ont décidé, de publier uniquement une partie de l'IA GPT-2 (Radford et al., 2019), IA capable d'écrire des textes catégorisables de « fake news »[14] et difficilement discernables par l'humain. Leur raison est le choix d'éprouver davantage le modèle et de laisser le temps de préparer des solutions aux problématiques amenées par ce modèle.

Cette difficulté de rendre toutes les IA transparentes et compréhensibles par tous reflète bien le fossé séparant une éthique parfaite aujourd'hui et l'idéalisation qu'il est possible de faire, d'autant plus quand les algorithmes intelligents affectent le quotidien. La responsabilité d'identifier les biais et comment rendre l'IA transparente revient dans un premier temps aux personnes s'occupant de la création des intelligences artificielles.

2.3.3 Derrière chaque grande IA, des humains

Ces dernières années, le quotidien d'une grande majorité de Français (ou autres habitants d'un pays catégorisé comme occidental) a énormément changé. En effet, suivant un rythme de croisière, les algorithmes dictant le style de vie du quotidien se sont retrouvés omniprésents : Facebook, Google, Netflix, Youtube et d'autres. Ces intelligences artificielles affectent-elles notre vision du monde, notre vie au quotidien ? Les data scientists[15], parents des IA peuvent bien changer notre quotidien au travers de leurs enfants.

Facebook, le réseau social ayant le plus d'utilisateurs dans le monde avec deux milliards 320 millions d'utilisateurs (Clement, 2019), a déjà eu l'occasion de prouver son influence sur l'humeur des utilisateurs. L'algorithme, qui choisit quelles publications un utilisateur pourra voir sur son fil d'actualité, possède une puissance d'impact sur nos émotions : si l'IA affiche une majorité de publications uniquement positives ou uniquement négatives sur une semaine, les publications, et donc l'humeur, des utilisateurs suivra l'état d'esprit auquel ils ont été exposés (Kramer, Guillory, et Hancock, 2014). Une telle constatation pose alors la question de l'influence que peut avoir l'idéologie des concepteurs de l'algorithme qu'ils en aient conscience ou non.

[13] L'open source est une philosophie de développement désignant le fait d'ouvrir le code à tous, librement.
[14] Traduction : Informations fausses. Il s'agit de nouvelles pouvant sembler vraies dans la forme, mais fausses dans le fond.
[15] Développeuse ou développeur en charge de créer le modèle qui composera l'intelligence artificielle.

Sur Internet, il est possible de nos jours de trouver des réponses à nos questions sur des moteurs de recherche comme Google, leader de ce domaine. Il est surprenant de savoir que depuis 2017, plus de vidéos Youtube sont vues que de recherches Google sont effectuées (Desjardins, 2018). La promotion de vidéos par l'algorithme de Youtube, fait rentrer ses utilisateurs dans des bulles correspondant à « leurs goûts ». Au même détriment que Facebook, la vision du monde d'un internaute est alors biaisée par cette plateforme.

Cette source de média en ligne, étant alors une source irréfutable de distraction, connaissance ou encore de culture, est pour chaque utilisateur une urne remplie de vidéos catégorisées. Le risque dans un environnement de ce type est qu'il y est une fixation sur une catégorie précise de vidéo (poussant alors le biais de confirmation s'il s'agit d'une idéologie). Mickaël Launay (Launay, 2012) soutient, qu'en mathématiques, un tel milieu ségrègue les utilisateurs en deux ensembles : les conformistes et anti-conformistes. Les conformistes auront uniquement des vidéos leurs correspondants, mais pour la seconde catégorie les vidéos seront d'un certain pourcentage correspondant à leur goût et le reste étant dans la catégorie conforme.

L'influence des intelligences artificielles sur nos humeurs et opinions dans notre utilisation d'internet quotidiennement est irréfutable. Bien que l'objectif de ses algorithmes (étant ceux des sociétés les créant) soit dans le but de maintenir ses utilisateurs le plus longtemps sur la plateforme, ces IA peuvent, malgré elles, pousser une addiction idéologique chez les utilisateurs. Les data scientists possèdent-ils les appétences pour des réflexions aussi poussées que les impacts sociaux issus de leurs intelligences artificielles ?

Une majorité de data scientists sont issus de formation statistique, mathématiques ou encore informatique. Cela contraint déjà sur la notion de connaissance du domaine sur lequel une IA sera développée. La responsabilité de ceux qui produisent les algorithmes intelligents est de mentionner dans leurs discours les limites de leurs travaux, mais aussi de fournir un maximum d'informations aux personnes prenant les décisions et législateurs qui devront évaluer l'impact potentiel de ses apports sur la société (Cointe, 2017).Si cette responsabilité n'est pas respectée ou pire, que les limites sont omises volontairement afin de profiter à l'entreprise créatrice alors les répercussions peuvent être de l'ordre des biais impactant des femmes et hommes n'en ayant point conscience.

L'objectif d'une conscience de l'éthique pour que les intelligences artificielles soient plus transparentes et moins biaisées, qu'elles soient plus justes, est bien présent dans la communauté des chercheurs. Malheureusement la conscience de cette problématique n'est pas majoritaire et dans le but de sensibiliser sur ce sujet, l'association Data for Good[16] a mis en place le « Serment d'Hippocrate pour Data Scientist » (Data for Good, 2018) prenant en compte ces problématiques. Bien que cela ne soit encore qu'un engagement n'ayant pas de valeur juridique, peut-être cela évoluera dans les années qui suivent.

Les valeurs qui sont évoquées dans ce document peuvent se résumer au nombre de cinq. Dans un premier temps, l'intégrité scientifique et la rigueur, puis la transparence vis-à-vis de l'information compréhensible par le plus de parties prenantes possibles. L'équité suit, afin de veiller à une égalité et d'éviter la discrimination de groupes. L'avant-dernière concerne le respect de la vie privée des personnes qui peuvent être touchées par les travaux réalisés et enfin la responsabilité poussant à assumer tout manquement ou en cas de conflit d'intérêt.

Cette conscience ne doit pas se limiter qu'aux concepteurs des IA, il est nécessaire que cela soit multi-disciplinaire. L'émergence des intelligences artificielles et la possibilité du passage de la

[16] Traduction : Données pour le bien

singularité fait apparaître la difficulté à aligner la morale de l'homme à celle de la machine : c'est le problème de l'alignement.

2.3.4 Expliciter la morale de l'homme pour l'IA

Le futur qui admet une intelligence artificielle ayant un intellect supérieur à l'homme plonge dans les débats éthiques et culturels, le problème de l'alignement oblige de pouvoir écrire noir sur blanc « la morale » souhaitée pour une machine. Cette démarche est nécessaire, la morale qu'une machine aurait dans le futur est difficilement conceptualisable pour les humains.

Pour ce faire, il serait intéressant d'avoir une liste de règles bien définies comme les célèbres trois lois de la robotique d'Isaac Asimov. Bien qu'accès sur l'atteinte et l'obéissance à l'être humain, le raisonnement est à pousser au maximum.

Le cas concret d'une voiture autonome, dans un contexte où les freins ne fonctionneraient pas en face de piétons, qui la voiture choisira de tuer (expérience de pensée). Le MIT a proposé une étude nommée « Moral Machine », offrant des choix en fonction de l'âge, du nombre, du sexe, de la classe sociale. Les résultats montrent une divergence entre les cultures. En effet pour la question de l'âge, la jeunesse sera préférée pour des pays dit occidentaux à contrario des pays asiatiques favorisant les personnes âgées (Awad et al., 2018).

C'est déjà une étape importante puisque la réflexion émerge sur des questions qui sont d'actualité, e.g. la Californie autorisant les voitures autonomes à circuler sans conducteur (Shepardson et Sage, 2018), et permettent de commencer une réflexion plus profonde sur l'avenir de l'IA et d'une morale peut être relative à un pays. Tout ceci n'est que spéculation, les lois ne sont pas encore adaptées, les éthiques divergent énormément selon les pays.

Alors, si les lois sont plus difficilement malléables que les mœurs en rapport aux machines, des travaux ont été réalisés pour permettre dès l'apprentissage de l'algorithme intelligent de pouvoir aligner la morale humaine à celle de la machine.

Une solution pourrait être de découper la fabrication d'une IA en cinq parties : la collecte de données fiables, le modèle d'apprentissage basé sur la représentation du monde (issu des données collectées), la compréhension du modèle ainsi que le choix de la mesure de performance du modèle, définir les incentives[17] en jeux et enfin le renforcement de l'apprentissage dans le temps (Hoang, 2018). Cette méthodologie s'axe sur tout le développement d'une IA, le point de départ pour parvenir à réaliser cela est de pouvoir conceptualiser les valeurs humaines.

La collecte des valeurs humaines dans le but de l'alignement passera par des questionnaires pour les humains, le problème étant que les réponses qui seront fournies posséderont des défauts : les biais, le manque de connaissances dans le domaine, les capacités cognitives limitées ou encore la culture dans laquelle évolue la femme ou l'homme. La nécessité d'inclure des sociologues avec chercheurs en IA est importante afin de répondre à la question de savoir si les humains donneront une bonne réponse à la question (Irving et Askell, 2019).

La forme de l'apprentissage de la morale peut être appliquée sous diverses formes, par exemple l'une pourrait être par observation du comportement des autres et en ajustant la morale par l'observation des conséquences (Cointe, 2017). Une autre viserait plus l'apprentissage par le débat, poussant alors la justification des intelligences artificielles dans leurs recoins (Irving et Askell, 2019).

[17] Il s'agit de source de motivation pour réaliser une action, e.g. une médaille lors d'une compétition sportive.

La sensibilisation sur la problématique d'une morale pour les IA ainsi que le problème de l'alignement est primordiale. Il est important de travailler sur ces problèmes sur le plan technique en formant les actuels et futurs data scientists sur la question éthique et en rajoutant des sociologues pour aider sur les réponses aux questions de l'ordre moral, en d'autres termes, l'éducation et le social seront deux secteurs clés des avancées dans la transparence des algorithmes intelligents.

2.4 Synthèse

Pour rappel, la problématique de ce mémoire pose le cadre sur deux domaines : l'intelligence artificielle et l'éthique. Dans ce chapitre, dans un premier temps, l'histoire de l'intelligence a été évoquée avec notamment le test de Turing (voir section 2.1.1). Celui-ci pouvant symboliser une première étape pour atteindre le passage de la singularité, qui, selon les experts du domaine aurait plus de cinquante pourcents de chance d'être atteinte d'ici cinquante ans (voir Figure 2).

Dans un second temps, une introduction à la morale, avec entre autres son origine historiquo-philosophique vu par Nietzsche (voir section 2.2.1). Différentes morales ont été évoquées d'un point de vue philosophique démontrant une complexité déjà humaine à trouver une morale pouvant plaire au plus grand nombre (voir section 2.2.2).

C'est pour cela qu'en dernier lieu, la section 2.3 rentre dans les problèmes soulevés par l'intelligence artificielle sur le plan de l'éthique et présente un premier état de l'art sur les études et technologies existantes. Différents aspects ont été cités : les biais, la transparence des algorithmes, l'influence actuelle des IA avec les data scientists et enfin une vision qui pourrait permettre d'expliciter la morale des intelligences artificielles.

La problématique peut alors se découper sur trois points de l'éthique : moraliser les IA, la transparence et enfin les impacts sociaux que peuvent engendrer ces dernières.

Moraliser un algorithme revient à définir la vision de la morale que les créatrices et créateurs de ce dernier veulent lui attribuer, mais aussi à réfléchir aux biais que les données collectées peuvent posséder ou alors dans l'IA même.

La transparence s'axera sur l'interprétabilité du modèle, sa capacité à fournir des explications globales ou locales et enfin sur la possibilité ou non de rendre le code ouvert à tous.

Enfin, les impacts sociaux s'inscriront dans la démarche de poser les questions des conséquences que celui-ci peut avoir pour un groupe d'individus ou pour un cas particulier, qui est concerné et dans quels buts.

La vision qui ressort est celle d'une « méta-IA » soit d'une boîte à outil, d'un guide qui pourrait s'adapter aux différentes intelligences artificielles. Dans le chapitre suivant, différentes méthodes seront évoquées pour répondre aux problématiques évoquées : les approches existantes puis des explorations spécifiques à ce mémoire qui seront effectuées.

L'objectif de ce mémoire est de proposer une feuille de route pour tous projets consacrés au domaine de l'IA.

3. Méthodologie

Le chapitre 2 a décrit les concepts d'intelligence artificielle et de la morale afin de voir comment ces deux domaines peuvent être liés de nos jours et dans un futur plus ou moins proche. La section 3.1 fera un état de l'art sur les travaux existants dans la recherche sur la réduction de biais et la transparence des IA. La section 3.2 détaillera les méthodes qui seront employées dans ce mémoire pour résoudre la problématique énoncée.

3.1 Approches existantes

Cette section a pour but d'analyser des travaux portant sur la réflexion éthique au sein du domaine de l'intelligence artificielle. Les travaux qui seront présentés proviennent de réflexions mathématiques.

L'introduction de ces différentes approches est importante puisque certaines sont liées à des outils techniques qui seront réutilisés dans ce mémoire, notamment avec la création d'un outil technique.

3.1.1 Paradoxe de Simpson

Pour construire un modèle aujourd'hui, il est très souvent nécessaire d'utiliser des données, or dès que des données sont utilisées, il est question du domaine de la statistique. Dans ce cadre, il est nécessaire de connaître l'environnement qui entoure les données utilisées : il s'agit de créer une représentation d'un paradigme lié aux données collectées, e.g. un jeu de données basé sur les Français se représentera « une réalité » basée uniquement sur les habitants de la France.

Quand la question éthique est posée, la nécessité de bien comprendre ce monde est essentielle : les personnes composantes de l'environnement, les variables prisent en compte et enfin l'agrégation de ces variables, e.g. la médiane de l'âge des Français, avec les Français qui composent l'environnement, l'âge étant la variable prise en compte et l'agrégation se définissant par la médiane.

La statistique aide fortement, au tout début de la création d'un modèle, à éviter de biaiser l'entraînement de l'intelligence artificielle. En effet, si les données sont initialement bien interprétées et comprises par les humains alors la généralisation d'une règle jugeable comme « mauvaise » peut être évitée. Malheureusement, lors de la réalisation d'une analyse exploratoire[18] pour un modèle, une réflexion est à apporter autour du paradoxe de Simpson.

Ce paradoxe provient de l'article technique « The interpretation of Interaction in Contingency Tables » (Simpson, 1951). L'affirmation qui ressort de ce paradoxe est qu'un phénomène observé pour plusieurs groupes peut être inversé lorsque ces groupes sont combinés. Cette découverte est assez effrayante car contre intuitive.

Pour illustrer le paradoxe, l'exemple suivant est celui d'une étude sur 1314 fumeuses Anglaises (Appleton, French, et Vanderpump, 1996). Deux tableaux ont été déduits des données collectées sur une période de vingt ans. Le premier tableau donne les informations globales :

Fumeuse	Morte	Vivante	Total	% Morte

[18] Il s'agit d'une analyse statistique effectuée sur les données avant de créer son modèle dans le but de comprendre les données sur lesquelles l'algorithme travaillera.

Oui	139	443	582	24
Non	230	502	732	31
Total	369	945	1314	28

Tableau 1: Présentation des résultats de l'étude par rapport au fait de fumer ou non (Appleton, French, et Vanderpump, 1996)

L'information ressortant de ce tableau est que sur cette même période, les fumeuses ont compté moins de décès par rapport aux non-fumeuses. Il est alors tentant de penser à une corrélation entre le fait de fumer et d'avoir moins de chance de mourir, ce qui peut paraître très étrange.

Le second tableau présente les mêmes données de façon différente, en séparant par groupe d'âge :

Âge du groupe	18-24		25-34		35-44		45-54		54-55	
Fumeuse	O	N	O	N	O	N	O	N	O	N
Morte	2	1	3	5	11	7	27	12	51	40
Vivante	53	61	121	152	95	114	103	66	64	81
Ratio	2.3		0.75		2.4		1.44		1.61	

Tableau 2: Présentation des résultats de l'étude par rapport au fait de fumer ou non et par groupe d'âge (Appleton, French, et Vanderpump, 1996)

La dernière ligne présente le ratio entre la proportion de femmes qui sont mortes et celles qui vivent toujours, par tranche d'âge[19]. Cette ligne affirme alors l'opposé du Tableau 1, soit que si les données sont découpées par groupe d'âge, alors une femme anglaise aura plus de chance de survivre si elle ne fume pas (sauf pour la tranche 25-34 ans).

Au travers du paradoxe de Simpson, il est important de pousser la réflexion sur la collecte des données et de croiser les informations pour être sûr de ne pas mal interpréter le paradigme qu'elles peuvent décrire.

3.1.2 AI Fairness 360

La discussion autour des biais (voir section 2.3.1) poussant à réfléchir comment construire un modèle est une première étape, mais il ne faut jamais exclure la possibilité qu'un ou plusieurs biais s'invitent dans une intelligence artificielle. Pour cela, il est nécessaire de trouver un moyen d'identifier ces biais potentiels ainsi que de parvenir à les limiter.

Dans le cadre de cette démarche, l'outil AIF360[20] (Bellamy et al., 2018) présente divers instruments formant une harmonie pour identifier les biais et les réduire si besoin. Bien que cet outil n'est encore qu'à ses débuts (version 0.2.0, juin 2019), sa qualité n'en est pas moindre. Il se découpe principalement en deux axes : les mesures des biais et les algorithmes de limitation de biais.

Les mesures calculeront des biais à partir d'une variable, de préférence correspondant à une segmentation sociale, e.g. sexe d'une personne. Certaines mesures ne s'appuieront que sur le jeu de

[19] E.g. un ratio de 2 signifie que pour un même nombre de fumeuses et de non-fumeuses, il y a deux fois plus de mortes chez les fumeuses.

[20] Le nom complet est « Artificial Intelligence Fairness 360 » qui signifie Intelligence Artificielle Juste 360.

données et donc identifieront les biais présents dans le jeu de données, d'autres auront besoin des prédictions pour détecter les biais présents dans le modèle. Ces deux groupes permettent d'avoir une vue d'ensemble sur les données et le modèle.

Les algorithmes, eux, se catégorisent en trois groupes appelés « fair-processors »[21] : les « pre-processors », « in-processors » et « post-processors ». Les algorithmes de chaque catégorie possèdent globalement les mêmes caractéristiques au détriment du moment de leur utilisation dans la création d'un modèle.

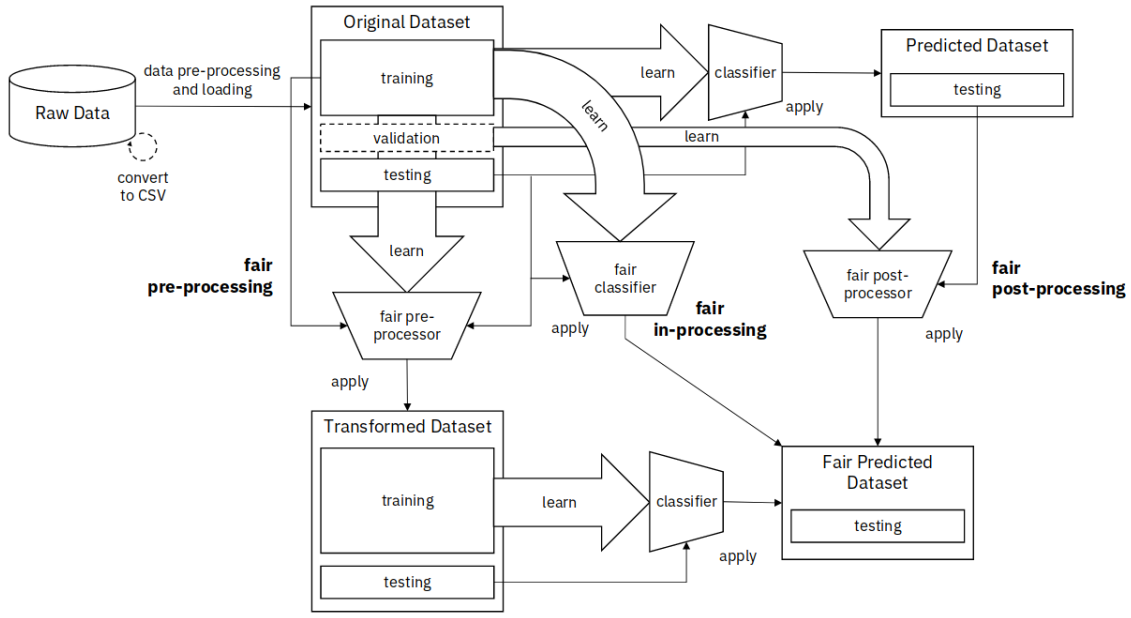

Figure 5: "The fairness pipeline", les différents chemins d'utilisation d'AIF360 (Bellamy et al., 2018)

L'image décrit le pipeline de la création d'un algorithme intelligent, dans un premier temps les données brutes sont transformées en un jeu de données formaté par AIF360. À partir d'ici, pour un « pre-processor » la réduction de biais se fait directement avant l'entraînement du modèle, pour un « in-processor », le travail s'effectue pendant l'entraînement et enfin pour le « post-processor » après l'entraînement.

Une dizaine d'algorithmes issus de recherche sont disponibles pour atténuer les biais au sein de l'outil AIF360. Cet outil propose donc une vision de contrôle mathématique des biais et de solutions afin de les atténuer.

3.1.3 SHAP : Shapley Additive exPlanations

Dans un monde parfait, un modèle expliquerait en langage naturel ce qui l'a amené à une prédiction particulière ou alors comment il fonctionne en général. Il est possible de faire une comparaison avec une doctoresse expliquant à son patient pourquoi elle pense qu'il a une maladie à partir de ses symptômes. Une explication en langage naturel faciliterait grandement la compréhension pour tout humain.

[21] Traduction : processeurs justes.

Aujourd'hui, une telle solution n'est qu'utopie, mais l'outil SHAP (Lundberg et Lee, 2017) offre une possibilité d'explication de modèle très séduisante en utilisant notamment la théorie des jeux[22] avec la valeur de Shapley (Shapley, 1953).

Cette valeur sert à répondre à la question : si un groupe de joueur collabore, comment diviser la récompense totale obtenue par le groupe ? Avec la façon dont elle est calculée, la valeur de Shapley garantie les trois axiomes clés de l'interprétabilité : « Dummy player », « Symmetry » et « Additivity »[23] (Manea, 2016).

En somme, ces trois axiomes garantissent les fondamentaux suivants : si un joueur n'ajoute aucune somme au total, sa part sera de zéro, si deux joueurs ont ajouté la même somme de départ alors en sortie ils auront la même somme et si un jeu peut se décomposer en différents sous-jeux, il doit être possible d'ajouter les sommes des sous-jeux en sortie.

Concernant l'outil SHAP, il utilise cette valeur en remplaçant les joueurs par les variables qui composent le modèle. Grâce à cette logique, il est possible de fournir des explications pertinentes pour une prédiction donnée et donc en prenant un ensemble de prédiction, il est alors possible de comprendre comment l'intelligence artificielle fonctionne en général.

Cet outil permet alors de rendre une IA plus transparente. Pour l'exemple de la doctoresse, si une machine indique qu'un patient a une maladie précise, en fournissant l'explication par rapport aux symptômes donnés alors la doctoresse aura confiance en l'intelligence artificielle si l'explication est cohérente avec sa connaissance médicale.

3.2 Présentation des méthodes

Dans la partie précédente, les différentes approches offrent des solutions axées dans le domaine des mathématiques et notamment pour les deux dernières des solutions, des solutions techniques, puisque les deux outils sont disponibles avec le langage informatique Python.

L'objectif de cette section sera de présenter les travaux qui seront réalisés dans le cadre du mémoire afin de répondre à la problématique initialement énoncée.

3.2.1 Contextualisation

Les travaux de ce mémoire vont se découper en deux parties : l'une qui sera principalement axée sur des entretiens avec différents acteurs dans le but de cerner des problématiques sous-jacentes à l'intelligence artificielle et l'éthique, tout en diversifiant les visions des personnes interrogées. La seconde visera la création d'une feuille de route à la fois basée sur la réflexion lors de la création d'une IA et sur un outil technique (utilisant le langage Python) pour permettre une aide sur cette réflexion.

Afin de recentrer les travaux et pour être précis, les expériences qui seront menées dans le cadre de la feuille de route suivront les critères suivants : dans le domaine bancaire, des modèles de

[22] La théorie des jeux est un domaine des mathématiques qui s'intéresse aux interactions des choix d'individus (appelés « joueurs »).
[23] Traduction dans l'ordre : joueur factice, symétrie, additivité

classification binaire : où la prédiction doit être entre deux possibilités (oui ou non) et enfin les données en entrée seront tabulaires.

Concernant le domaine bancaire, deux environnements seront utilisés. Le premier correspond à l'entreprise Caisse d'épargne Aquitaine Poitou Charentes et le second concerne une compétition Kaggle « Home Credit Default Risk »[24], l'objectif de cette compétition était de construire un modèle qui prédira la capacité d'un client à rembourser un crédit.

Pour la suite, deux méthodes vont être favorisées pour permettre un cheminement qui apportera les éléments de réponse pour ce mémoire. Tout d'abord, des entretiens qualitatifs avec une trame directive afin de discuter des principaux problèmes du domaine. Différents acteurs seront concernés, à minima dans les domaines de la Data Science, sociologie et bancaire. La seconde méthode, elle, se recentrera sur la feuille de route, sa construction, puis une preuve de concept en l'utilisant sur des modèles.

Les deux méthodologies seront liées, l'évolution des entretiens pourra influencer la logique de la feuille de route et cette dernière pourra ajouter du fond technique pour les discussions poussant alors vers une idéologie de démocratisation d'outils éthiques pour le domaine de l'intelligence artificielle.

3.2.2 Raisonnement chez l'homme, analyse sur les acteurs

L'importance d'échanger avec autrui, afin de mieux cerner un problème, est essentielle. Il faut avoir conscience que le monde qui compose ce qui est appelé société n'est pas le reflet du paradigme égocentrique qui est subjectif à chacun. La démarche d'entretiens qualitatifs va donc pousser la réflexion avec des acteurs de milieux différents afin d'affiner la perception des problèmes liés au domaine et également en cas échéant de sensibiliser à des problématiques parfois ignorées.

Les interviews qui seront conduites auront, selon l'interlocuteur, des objectifs précis et différents. En premier lieu, il sera nécessaire de déterminer la connaissance autour du domaine de l'intelligence artificielle de l'acteur interrogé.

Quatre profils sont définis pour cela, les data scientists, les sociologues, les directeurs bancaires et des intervenants sensibilisant sur les questions de l'éthique dans le domaine de l'intelligence artificielle.

Premièrement, le profil du data scientist est obligatoire puisqu'il s'agit des concepteurs des algorithmes et ils sont également garants de leurs modèles. La trame de l'entretien avec ce profil sera la suivante. Premièrement, connaître son interlocuteur, dans quel domaine il travaille, son expérience, et ses réalisations. En suivant, contextualiser avec quelles technologies son entreprise travaille et quel est le cycle de création classique d'une IA. Une fois l'introduction faite, les questions autour de l'éthique : si l'entreprise est déjà sensible à la question éthique de l'intelligence artificielle, le point de vue sur la singularité et comment faudrait-il trancher par rapport à un algorithme très rentable, mais qui n'est pas forcément « moral ». Pour finir, un questionnement sur des solutions techniques existantes (voir section 3.1) et les idées de la personne interrogée sur la problématique.

Deuxièmement, la vision sociale et donc un entretien vers un profil sociologue. L'idée au travers de l'échange sera de bien définir les questions sociales au sein du domaine de l'intelligence artificielle et par conséquent d'avoir une vision véritablement centrée sur les impacts sociaux. Dans l'ordre,

[24] https://www.kaggle.com/c/home-credit-default-risk/overview

l'interview cherchera d'abord à savoir qui est en face, ses travaux, la vision autour des algorithmes qui régissent déjà notre quotidien puis la vision sur la singularité et enfin comment est-ce que la personne définirait les impacts sociaux dus à ce domaine de recherche.

Ensuite, le profil à interroger est axé sur un domaine spécifique, le domaine bancaire soit un profil de directeur en banque. Il s'agit là d'un domaine en particulier pour cerner localement les problèmes liés à l'IA. Le choix de ce domaine est dû à la cohérence des travaux qui suivront sur des modèles du domaine de la banque. De plus, incontestablement les banques ont un rôle majeur quant au quotidien des humains et logiquement sur le social. Une contrainte sera nécessaire pour la ou les personne·s interrogée·s : avoir déjà travaillé avec une intelligence artificielle.

Pour ce profil le déroulement suivra l'ordre suivant : introduction sur son travail, expliquer dans quel cadre le travail avec une IA s'est produit, quel était son objectif et s'il y a eu des problèmes d'ordre humains ou techniques. La suite sera centrée sur ce qui a été présenté à la direction en termes de transparence du modèle et de savoir comment trouver la limite entre éthique et rentabilité du modèle.

Pour finir, le dernier profil est celui d'intervenant sensibilisant déjà aux problématiques de l'ordre éthique (pas forcément technique). La logique de l'entretien avec un tel profil est d'obtenir un ressenti de quelqu'un qui a déjà fait face à ces problématiques. La personne pourra être autant formatrice que data scientist, l'idée est de savoir comment elle agit pour mettre en place une vision éthique dans une entreprise. Le questionnement s'appuiera, en plus de savoir ce que l'interlocuteur fait, sur ses motivations, la ou les méthodologie·s qu'il utilise et les solutions qu'il recommande.

Grâce à ces différents entretiens, les informations qui en seront dégagées apporteront diverses pierres à l'édifice et très certainement des solutions techniques qui pourront être mises en place dans la section suivante.

La diversité des profils permet d'avoir une vision technique, sociale, métier et enfin formatrice. Cela garantit de fournir un maximum de détail au travers des questionnements qui sont soulevés par la problématique de ce mémoire.

3.2.3 Raisonnement technique, réflexion sur la feuille de route

Pour parvenir à créer des intelligences artificielles plus justes et transparentes, la première clé est la sensibilisation et l'éducation sur les potentiels risques sociaux qu'une IA peut engendrer. Cette première étape, bien qu'essentielle, ne suffit pas, il y a un autre pas à faire pour arriver à se lancer. C'est pour cela que la création d'une feuille de route permettra d'être guidé tout au long de la création d'un algorithme.

Dans un premier temps, l'identification des étapes clés de la construction d'une IA est primordiale. Pour chaque étape l'application de la feuille de route aura son importance, puisque si l'une d'entre elle est opaque, l'idée d'une IA transparente peut être remise en question.

L'outil et la feuille de route qui seront réalisés auront pour nom « TransparentIA » afin d'explicitement nommer ce pour quoi ces travaux sont réalisés, le code et la recherche autour seront publiés et open source donc accessible à toutes et à tous.

Pour rappel, TransparentIA a pour but de pouvoir créer un algorithme le plus juste et transparent possible, de plus, ce qui sera réalisé ne sera qu'une première version puisque le domaine de l'éthique en intelligence artificielle est encore jeune, plus il gagnera en maturité, plus TransparentIA pourra

évoluer vers un idéal. Deux outils seront mis à disposition, un papier présentant la méthodologie dans son ensemble et une librairie en langage Python, ce qui permettra d'allier théorie et pratique. Le cheminement que suivra la feuille de route est décrit ci-dessous.

3.2.3.1 Descriptif de la feuille de route

La première étape sur laquelle il est nécessaire d'agir lors de la création d'une IA doit s'effectuer lors du cadrage : il faut clairement expliciter qui est le demandeur du projet, le ou les développeur·s et surtout quelle sera la vision du monde que le modèle aura, en d'autres termes quelles données seront utilisées, d'où proviennent-elles et qui en est à l'origine ?

Dans la continuité, il faut être conscient de qui sera affecté par le modèle, si à partir des décisions prises avec les prédictions du modèle, un impact social peut avoir lieu, e.g. obtention d'un taux de crédit ou non, alors il est essentiel de l'indiquer et ainsi de redoubler de vigilance.

Le cadrage étant fini, la première étape du développement sera de collecter les données. Dans cette étape, afin de garantir et prévenir des potentiels biais, l'essentiel est de s'assurer de la fiabilité des données, de leur qualité.

Ensuite, une analyse dite exploratoire s'effectue et par la même occasion vérifier la présence ou l'absence de biais sur des critères sociaux tels que le sexe, l'âge ou encore l'orientation sexuelle.

Vient la création du modèle et par conséquent le choix du modèle, pour rappel les algorithmes les plus performants sont souvent les moins interprétables (voir Figure 4). S'il est choisi de travailler avec un modèle complexe, justifier ce choix et par la suite expliquer le modèle dans sa globalité ainsi que par prédiction afin de ne pas avoir une vision boîte noire de l'IA entre humain et machine.

Le modèle terminé, la vérification des biais potentiels que le modèle pourrait avoir appris est à faire, est-ce qu'il favorise une classe sociale plus qu'une autre par exemple.

Une fois déployé, si le modèle est ré-entrant, soit qu'il se renforce au cours du temps, surtout dans le cas où les données sur lesquelles il apprendra seront directement issues de ses prédictions, il faut analyser les données prédites qui seront par la suite ses données d'entraînements.

Pour résumer, ci-dessous, un graphique récapitulant les différentes étapes :

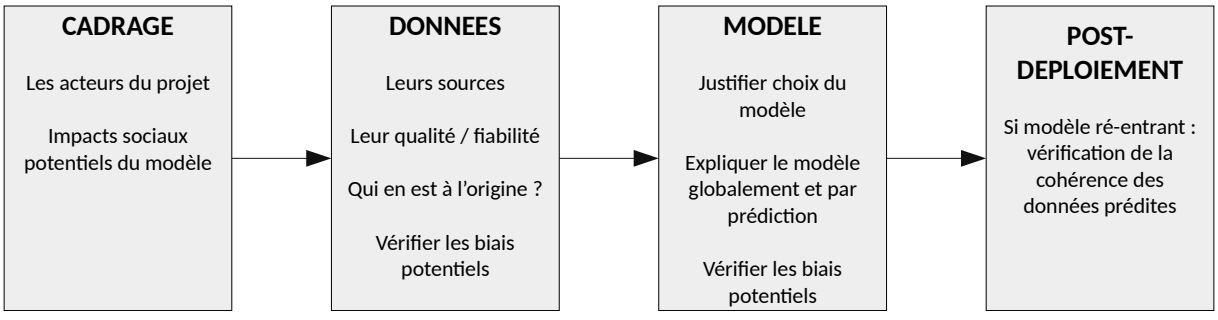

Figure 6: Récapitulatif des différentes étapes de la feuille de route

3.2.3.2 Descriptif code technique

Sur le plan technique, c'est une librairie du langage Python qui sera développée. Son objectif sera de répondre aux critères suivants : open source, permet d'expliquer un modèle localement et globalement, détecter les biais présents et surtout être intelligible par les humains.

Pour y parvenir, les technologies AIF360 (voir section 3.1.2) et SHAP (voir section 3.1.3) seront utilisées pour identifier les biais et expliquer les modèles. Le code sera hébergé sur le site Github.

La logique de cette librairie sera de facilement brancher l'outil TransparentIA à un modèle existant et de pouvoir exécuter les fonctions sur le plan de l'éthique. Pour être compréhensible, l'outil fournira des graphiques détaillés et présentables.

3.3 Synthèse

Dans ce chapitre, premièrement les approches existantes ont été mises en avant dans le domaine de l'éthique sur le plan mathématique de l'intelligence artificielle (voir section 3.1). Ensuite, les travaux qui seront réalisés dans le cadre de ce mémoire ont été détaillés avec les entretiens qualitatifs et l'approche technique (voir section 3.2).

Les résultats des entretiens et expériences menées au travers de la méthodologie proposée ici, seront mis en forme dans la suite du mémoire, qui est à rendre pour septembre 2020.

4. Conclusion

Tout au long ce pré-mémoire, la problématique alliant les deux notions d'intelligence artificielle et d'éthique a été présentée et détaillée. Le contraste sur leur communion peut être à la fois évident avec les films de science-fiction souvent à l'effigie d'une machine dominant l'humanité et à la fois peu comprise dû à la méconnaissance des algorithmes intelligents existants sur Terre.

Pour rappel, la catégorie d'IA dominant largement le domaine est celui du « Machine Learning », soit l'apprentissage de la machine (voir section 2.1). Au travers de ce regroupement d'algorithme, lorsque le regard se porte sur l'existant, il est remarquable de voir les réalisations, e.g. une voiture autonome.

La pluie d'intelligences artificielles s'abattant sur les diverses sociétés refont surgir un engouement général pour ce domaine, mais à l'aube du XXIe siècle la singularité a peu de chance d'être atteinte (voir section 2.1). La crainte de passer cette dernière sans avoir conscience de la morale des machines est pourtant bien fondée.

Bien qu'il n'existe pas de morale universelle, la diversité de codes éthiques régissant les mœurs montre bien la difficulté de constituer une forme acceptable de tous sur le plan moral (voir section 2.2). Aujourd'hui la morale de la machine, si elle n'est pas explicitée, risque d'être une complète inconnue pour tous.

L'un des problèmes bloquant une première étape clé concernant l'éthique de l'IA est la question de la transparence des modèles, que cela soit du code réservé à l'entreprise ou même pire que les concepteurs des algorithmes ne possèdent aucun indice sur la cause d'une simple prédiction ou décision de leur IA.

Dès lors que des groupes sociaux sont concernés par ces nouvelles technologies, il est alors primordial d'éduquer sur la question d'expliquer les prédictions d'un modèle. En suivant cette logique, la problématique se découpe alors en trois parties : la moralisation des IA, la transparence et les impacts sociaux que peuvent engendrer les algorithmes intelligents.

Afin de répondre efficacement à ces questions, ce pré-mémoire a introduit une méthodologie ~~pour y répondre~~ (voir section 3.2). Les méthodes qui seront utilisées par la suite auront deux axes, celui d'entretiens permettant un approfondissement des connaissances sur les questions sous-jacentes à la problématique évoquée. Le second sera axé sur le plan technique et d'un accompagnement théorique avec une feuille de route.

La nécessité d'une prise de conscience sur les questions éthiques pour les IA est fondamentale, puisque l'objectif souvent premier lors de la création d'un algorithme est la performance garantissant alors pourquoi pas un retour sur investissement. La lutte entre rentabilité et équité passe bel et bien par le domaine de l'intelligence artificielle.

L'objectif de ce mémoire s'axe sur un point fondamental : offrir à toute personne souhaitant développer une IA de nos jours, un moyen théorique de palier aux problématiques d'ordre moral. De plus, un outil technique sera développé pour remonter factuellement les biais potentiels et souci de transparence.

Pour conclure, bien que les travaux qui seront réalisés avec ce mémoire sont dans une logique de transparence et de réduction d'impact social, la personne ayant le dernier mot est l'être humain responsable de l'intelligence artificielle.

Bibliographie

Alexandre L. *Laurent Alexandre : Intelligence artificielle [EN DIRECT] - YouTube* [En ligne]. 8 novembre 2017. Disponible sur : < https://www.youtube.com/watch?v=QS951xiGGvI > (consulté le 18 juillet 2019)

Appleton D. R., French J. M., Vanderpump M. P. J. « Ignoring a Covariate: An Example of Simpson's Paradox ». *Am. Stat.* [En ligne]. 1 novembre 1996. Vol. 50, n°4, p. 340-341. Disponible sur : < https://doi.org/10.1080/00031305.1996.10473563 > (consulté le 18 juillet 2019)

Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I. « The Moral Machine experiment ». *Nature* [En ligne]. novembre 2018. Vol. 563, n°7729, p. 59. Disponible sur : < https://doi.org/10.1038/s41586-018-0637-6 > (consulté le 18 juillet 2019)

Bellamy R. K. E., Dey K., Hind M., Hoffman S. C., Houde S., Kannan K., Lohia P., Martino J., Mehta S., Mojsilovic A., Nagar S., Ramamurthy K. N., Richards J., Saha D., Sattigeri P., Singh M., Varshney K. R., Zhang Y. « AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias ». *ArXiv181001943 Cs* [En ligne]. 3 octobre 2018. Disponible sur : < http://arxiv.org/abs/1810.01943 > (consulté le 18 juillet 2019)

Bush V. « As We May Think ». In : *The Atlantic* [En ligne]. [s.l.] : [s.n.], 1945. Disponible sur : < https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ > (consulté le 18 juillet 2019)

Clement J. « Global social media ranking 2019 ». In : *Statista* [En ligne]. [s.l.] : [s.n.], 2019. Disponible sur : < https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ > (consulté le 18 juillet 2019)

Cointe N. *Ethical Judgment for decision and cooperation in multiagent systems* [En ligne]. Theses. [s.l.] : Université de Lyon, 2017. Disponible sur : < https://tel.archives-ouvertes.fr/tel-01851485 >

Crevier D. *AI: the tumultuous history of the search for artificial intelligence*. New York, NY : Basic Books, 1992. ISBN : 978-0-465-02997-6.

Data for Good. « Serment d'Hippocrate pour data scientist ». [s.l.] : [s.n.], 2018. Disponible sur : < https://www.hippocrate.tech > (consulté le 18 juillet 2019)

DeepMind. « AlphaGo ». In : *DeepMind* [En ligne]. [s.l.] : [s.n.], 2016. Disponible sur : < https://deepmind.com/research/alphago/ > (consulté le 18 juillet 2019)

Desjardins J. « Infographic: What Happens in an Internet Minute in 2018? ». In : *Vis. Capital.* [En ligne]. [s.l.] : [s.n.], 2018. Disponible sur : < https://www.visualcapitalist.com/internet-minute-2018/ > (consulté le 18 juillet 2019)

Ducat J. « Du vol dans l'éducation spartiate ». In : *Doss. Alexandre Gd. Relig. Tradit.* [En ligne]. Paris : Éditions de l'École des hautes études en sciences sociales, 2017. p. 95-110. Disponible sur : < http://books.openedition.org/editionsehess/2103 > (consulté le 18 juillet 2019)ISBN : 978-2-7132-2599-4.

Gilpin L. H., Bau D., Yuan B. Z., Bajwa A., Specter M., Kagal L. « Explaining Explanations: An Overview of Interpretability of Machine Learning ». *ArXiv180600069 Cs Stat* [En ligne]. 31 mai 2018. Disponible sur : < http://arxiv.org/abs/1806.00069 > (consulté le 16 juillet 2019)

Grace K., Salvatier J., Dafoe A., Zhang B., Evans O. « When Will AI Exceed Human Performance? Evidence from AI Experts ». *ArXiv170508807 Cs* [En ligne]. 24 mai 2017. Disponible sur : < http://arxiv.org/abs/1705.08807 > (consulté le 18 juillet 2019)

Greenwald A. G., Banaji M. R. « Implicit social cognition: attitudes, self-esteem, and stereotypes ». *Psychol. Rev.* janvier 1995. Vol. 102, n°1, p. 4-27.

Grudin J. « AI and HCI: Two Fields Divided by a Common Focus ». *AI Mag.* [En ligne]. 18 septembre 2009. Vol. 30, n°4, p. 48. Disponible sur : < https://doi.org/10.1609/aimag.v30i4.2271 > (consulté le 18 juillet 2019)

Harari Y. N., Dauzat P.-E. *Sapiens: une brève histoire de l'humanité*. Paris : Albin Michel, 2015. ISBN : 978-2-226-25701-7.

Hoang L. N. « A Roadmap for Robust End-to-End Alignment ». *ArXiv180901036 Cs* [En ligne]. 4 septembre 2018. Disponible sur : < http://arxiv.org/abs/1809.01036 > (consulté le 18 juillet 2019)

Hsu F., Campbell M. S., Hoane A. J. Jr. « Deep Blue System Overview ». In : *Proc. 9th Int. Conf. Supercomput.* [En ligne]. New York, NY, USA : ACM, 1995. p. 240–244. Disponible sur : < https://doi.org/10.1145/224538.224567 > (consulté le 18 juillet 2019)ISBN : 978-0-89791-728-5.

Irving G., Askell A. « AI Safety Needs Social Scientists ». *Distill* [En ligne]. 19 février 2019. Vol. 4, n°2, p. 10.23915/distill.00014. Disponible sur : < https://doi.org/10.23915/distill.00014 > (consulté le 18 juillet 2019)

Kant I., Delbos V. *Fondements de la métaphysique des mœurs*. Paris : Libraire Delagrave, 2007. ISBN : 978-2-206-00155-5.

Karras T., Laine S., Aila T. « A Style-Based Generator Architecture for Generative Adversarial Networks ». *ArXiv181204948 Cs Stat* [En ligne]. 12 décembre 2018. Disponible sur : < http://arxiv.org/abs/1812.04948 > (consulté le 18 juillet 2019)

Kohlberg L., Hersh R. H. « Moral development: A review of the theory ». *Theory Pract.* [En ligne]. avril 1977. Vol. 16, n°2, p. 53-59. Disponible sur : < https://doi.org/10.1080/00405847709542675 > (consulté le 18 juillet 2019)

Kramer A., Guillory J., Hancock J. « Correction for Kramer et al., Experimental evidence of massive-scale emotional contagion through social networks ». *Proc. Natl. Acad. Sci.* [En ligne]. 22 juillet 2014. Vol. 111, n°29, p. 10779-10779. Disponible sur : < https://doi.org/10.1073/pnas.1412583111 > (consulté le 18 juillet 2019)

Krauthammer C. « Be Afraid ». In : *Wkly. Stand.* [En ligne]. [s.l.] : [s.n.], 1997. Disponible sur : < https://www.weeklystandard.com/charles-krauthammer/be-afraid-9802 > (consulté le 18 juillet 2019)

Larson J., Angwin J. « How We Analyzed the COMPAS Recidivism Algorithm ». In : *ProPublica* [En ligne]. [s.l.] : [s.n.], 2016. Disponible sur : < https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm > (consulté le 18 juillet 2019)

Launay M. *Urnes interagissantes* [En ligne]. thesis. [s.l.] : Aix-Marseille, 2012. Disponible sur : < http://www.theses.fr/2012AIXM4775 > (consulté le 18 juillet 2019)

Lundberg S. M., Lee S.-I. « A Unified Approach to Interpreting Model Predictions ». In : Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (éd.). *Adv. Neural Inf. Process. Syst. 30* [En ligne]. [s.l.] : Curran Associates, Inc., 2017. p. 4765–4774. Disponible sur : < http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf > (consulté le 18 juillet 2019)

Manea M. « Strategy and Information ». In : *MIT OpenCourseWare* [En ligne]. [s.l.] : [s.n.], 2016. Disponible sur : < https://ocw.mit.edu/courses/economics/14-16-strategy-and-information-spring-2016/ > (consulté le 18 juillet 2019)

Medeot G., Cherla S., Kosta K., McVicar M., Abdallah S., Selvi M., Newton-Rex E., Webster K. « StructureNet: Inducing Structure in Generated Melodies ». In : *ISMIR*. [s.l.] : [s.n.], 2018.

Mordvintsev A., Olah C., Tyka M. *Inceptionism: Going Deeper into Neural Networks* [En ligne]. *Google AI Blog*. juin 2015. Disponible sur : < http://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html > (consulté le 18 juillet 2019)

Newsflash. *« AI is very, very stupid, » says Google's AI leader, at least compared to humans - CNET* [En ligne]. *News Flash*. 14 novembre 2018. Disponible sur : < https://newsflash.one/2018/11/14/ai-is-very-very-stupid-says-googles-ai-leader-at-least-compared-to-humans-cnet/ > (consulté le 18 juillet 2019)

Nietzsche F. *La Généalogie de la morale* [En ligne]. [s.l.] : Mercure de France, 1900. 27-82 p. Disponible sur : < https://fr.wikisource.org/wiki/La_G%C3%A9n%C3%A9alogie_de_la_morale/Premi%C3%A8re_dissertation >

Pinar Saygin A., Cicekli I., Akman V. « Turing Test: 50 Years Later ». *Minds Mach.* [En ligne]. 1 novembre 2000. Vol. 10, n°4, p. 463-518. Disponible sur : < https://doi.org/10.1023/A:1011288000451 > (consulté le 18 juillet 2019)

Poole D., Mackworth A., Goebel R. *Computational Intelligence: A Logical Approach*. New York, NY, USA : Oxford University Press, Inc., 1997. ISBN : 0-19-510270-3.

Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. « Language Models are Unsupervised Multitask Learners ». 2019. p. 24.

Shapley L. S. « 17. A Value for n-Person Games ». In : *Contrib. Theory Games AM-28 Vol. II* [En ligne]. Princeton : Princeton University Press, 1953. Disponible sur : < https://doi.org/10.1515/9781400881970-018 > (consulté le 18 juillet 2019)ISBN : 978-1-4008-8197-0.

Shepardson D., Sage A. « Waymo gets first California OK for driverless testing without… » *Reuters* [En ligne]. 31 octobre 2018. Disponible sur : < https://in.reuters.com/article/us-autos-selfdriving-waymo-idINKCN1N42S1 > (consulté le 18 juillet 2019)

Silver D., Schrittwieser J., Simonyan K., Antonoglou I., Huang A., Guez A., Hubert T., Baker L., Lai M., Bolton A., Chen Y., Lillicrap T., Hui F., Sifre L., Van den Driessche G., Graepel T., Hassabis D. « Mastering the game of Go without human knowledge ». *Nature* [En ligne]. octobre 2017. Vol. 550, n°7676, p. 354-359. Disponible sur : < https://doi.org/10.1038/nature24270 > (consulté le 18 juillet 2019)

Simon H. A. *The shape of automation for men and management*,. [1st ed.]. New York, : Harper & Row, 1965.

Simon H. A., Newell A. « Heuristic Problem Solving: The Next Advance in Operations Research ». *Oper. Res.* [En ligne]. 1 février 1958. Vol. 6, n°1, p. 1-10. Disponible sur : < https://doi.org/10.1287/opre.6.1.1 > (consulté le 18 juillet 2019)

Simpson E. H. « The Interpretation of Interaction in Contingency Tables ». *J. R. Stat. Soc. Ser. B Methodol.* [En ligne]. 1951. Vol. 13, n°2, p. 238-241. Disponible sur : < https://www.jstor.org/stable/2984065 > (consulté le 18 juillet 2019)

Solomonoff R. J. « The time scale of artificial intelligence: Reflections on social effects ». *Hum. Syst. Manag.* [En ligne]. 1 janvier 1985. Vol. 5, n°2, p. 149-153. Disponible sur : < https://doi.org/10.3233/HSM-1985-5207 > (consulté le 18 juillet 2019)

Sumner W. G. *Folkways, a study of the sociological importance of usages, manners, customs, mores, and morals* [En ligne]. [s.l.] : Boston, Ginn, 1906. 736 p. Disponible sur : < http://archive.org/details/folkwaysstudyofs00sumnuoft > (consulté le 18 juillet 2019)

Suwajanakorn S., Seitz S. M., Kemelmacher-Shlizerman I. « Synthesizing Obama: learning lip sync from audio ». *ACM Trans. Graph.* [En ligne]. 20 juillet 2017. Vol. 36, n°4, p. 1-13. Disponible sur : < https://doi.org/10.1145/3072959.3073640 > (consulté le 18 juillet 2019)

Tran V.-T. *Selection Bias Correction in Supervised Learning with Importance Weight* [En ligne]. Artificial Intelligence [cs.AI]. [s.l.] : Université de Lyon, 2017. Disponible sur : < https://tel.archives-ouvertes.fr/tel-01661470 > (consulté le 18 juillet 2019)

Tual M. « A peine lancée, une intelligence artificielle de Microsoft dérape sur Twitter ». *Le Monde.fr* [En ligne]. 24 mars 2016. Disponible sur : < https://www.lemonde.fr/pixels/article/2016/03/24/a-peine-lancee-une-intelligence-artificielle-de-microsoft-derape-sur-twitter_4889661_4408996.html > (consulté le 18 juillet 2019)

Turing A. M. « I.—COMPUTING MACHINERY AND INTELLIGENCE ». *Mind* [En ligne]. 1 octobre 1950. Vol. LIX, n°236, p. 433-460. Disponible sur : < https://doi.org/10.1093/mind/LIX.236.433 > (consulté le 18 juillet 2019)

Villani C. *Donner un sens à l'intelligence artificielle: pour une stratégie nationale européenne : [Mission parlementaire du 8 septembre 2017 au 8 mars 2018*. [s.l.] : [s.n.], 2018. ISBN : 978-2-11-145700-3.

Vinyals O., Babuschkin I., Chung J., Mathieu M., Jaderberg M., Czarnecki W. M., Dudzik A., Huang A., Georgiev P., Powell R., Ewalds T., Horgan D., Kroiss M., Danihelka I., Agapiou J., Oh J., Dalibard V., Choi D., Sifre L., Sulsky Y., Vezhnevets S., Molloy J., Cai T., Budden D., Paine T., Gulcehre C., Wang Z., Pfaff T., Pohlen T., Wu Y., Yogatama D., Cohen J., McKinney K., Smith O., Schaul T., Lillicrap T., Apps C., Kavukcuoglu K., Hassabis D., Silver D. *AlphaStar: Mastering the Real-Time Strategy Game StarCraft II* [En ligne]. [s.l.] : [s.n.], 2019. Disponible sur : < https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/ >

Wiener N., Le Roux R., Vallée R., Vallée N. *La cybernétique information et régulation dans le vivant et la machine*. Paris : Editions du Seuil, 2014. ISBN : 978-2-02-109420-6.