Mémoire de Mastère

Ynov Informatique Bordeaux

Mastère Data Science

L'Éthique des Intelligences Artificielles Supervisation et Confiance

Réalisé par :
M. Nathan Lauga

Sous la direction de :
M. Pascal Fournier [CEAPC]
M. Patrick Piquart [YNOV]

16 Août 2020

Table des Matières

Abstract

Ceci est l'avant-propos.

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.

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 les intelligences artificielles sont omniprésentes et que les données deviennent le nouveau pétrole, est-il possible de rendre une Intelligence Artificielle socialement responsable en supervisant son éthique ?

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 et scientifique. D'abord l'origine, puis les différents courants existants à l'heure actuelle et enfin les visions de la Morale en Science. La fin du chapitre se concentre sur la cohabitation de l'intelligence artificielle et de l'éthique.

Chapitre I

État de l'art

I.1 Les données et l'Intelligence Artificielle

> " Si une machine peut penser, elle pourrait penser plus intelligemment que nous, et alors où devrions-nous être ? Si nous pouvions maintenir les machines dans une position servile, par exemple en coupant le courant à des moments stratégiques, nous devrions, en tant qu'espèce, nous sentir humbles. "
>
> — [Turing, 1951]

Le 15 mai 1951, Alan Turing, considéré comme le père de l'informatique, été interviewé par la BBC et annoncé déjà l'avènement probable des machines intelligentes, où plus particulièrement les machines pensentes. Bien qu'à cette époque nous étions loin des ordinateurs de nos jours, la force des propos de Turing montre que, dès la moitié du $20^{\text{ème}}$ siècle, le concept d'intelligence artificielle existait.

Intelligence artificielle : En informatique, la recherche sur l'intelligence artificielle ou IA, est définie comme l'étude des "agnets intelligents", soit n'importe quel appareil qui perçoit son environnement et prend des déisions qui maximisent ses chances d'atteindre son objectif [Poole et al., 1997]. e.g. Dans les jeux d'échecs, un agent intelligent pourra, en connaissant les règles du jeu, effectuer des coups et son objectif, sera de battre son adversaire.

I.1.1 La préface : Test de Turing, Systèmes Experts et les Hivers

Algorithme : Un algorithme correspond à une suite d'instruction pouvant exécuter une tâche précise, en informatique ce qui est appelé programme est un ensemble d'algorithme et il est de même pour les intelligences artificielles.

Intelligence Artificielle, 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]. Avant cela, l'utilisation de cette notion existait déjà et l'un des premiers articles discutant de cela remonte à 1945 avant même la première explosion atomique [Bush, 1945]. Dans cet article, nous pouvons y voir les concepts d'ordinateurs, d'Internet ou encore de reconnaissance vocale.

En 1948, Robert Wiener, professeur au Massachusetts Institue of Technology (MIT) théorise la Cybernétique dans son livre "Cybernétique ou la communication contrôlée dans le domaine de l'animal et de la machine". Il décrit une théorie entière de la commande et de la communication, aussi bien chez l'animal que dans la machine [Wiener, 1961]. La but essentiel de la cybernétique est de comprendre et de définir des processus ou fonctions avec pour objectif de réagir par rapport à une certaine action en entrée. Il s'agit là de la préface au domaine de recherche de l'Intelligence Artificielle qui émergera à la conférence de Darthmouth évoquée dans le paragraphe précédent.

Lors de l'année 1950, la théorie du domaine étant à ses débuts, un nouveau papier permettra une grande avancée, celui d'Alan Turing [Turing, 1950]. Intitulé "COMPUTING MACHINERY AND INTELLIGENCE"[1], il soumettra une question inédites : "Les machines peuvent-elles penser ?". Cette interrogation est très contradictoire, surtout en 1950, puisque le terme *machine* et *penser*, ne peuvent être 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".

Test de Turing : Si une machine peut tenir une discussion avec un humain (au travers d'une messagerie par exemple), sans que la femme ou l'homme ne 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 et al., 2000].

Les années suivants ces papiers, des déclarations comme "d'ici dix ans un ordinateur sera le champion du monde des échecs" [Simon and Newell, 1958] ou encore "des machines seront capables, d'ici vingt ans, de faire tout travail que l'homme peut faire" [Simon, 1965] créeront un engouement populaire dans le domaine de la recherche de l'Intelligence Artificielle. Ne s'agissant pas des seules déclarations à ce sujet, et celles-ci mènent à une attente très élevée concernant les possibilités des algorithmes intelligents, mais comme souvent lorsque les attentes sont élevées une phase de déception s'ensuit.

Dans les années 1970, apparu 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 des ordinateurs ou encore le manque de base de connaissances du monde par les ordinateurs (manque de données). En effet, les travaux, qui portaient sur le langage naturel, ne pouvaient pas être extrêmement poussés puisque le stockage de la mémoire la limitait à vingt mots [Crevier, 1992]. Pour beaucoup ce secteur a été enterré, mais 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.

Système Expert : c'est un programme capable de reproduire des patterns afin d'obtenir une sortie. Concrètement, c'est un logiciel qui avec une base de connaissances et une base de règle peut obtenir une réponse [Jackson, 1998]. E.g. Un système expert peut avoir comme connaissance les polygones et leur nombre de côtés puis pour règle l'association comme trois côtés égal un triangle.

[1] Traduction : Les Machines Informatiques et l'Intelligence.

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

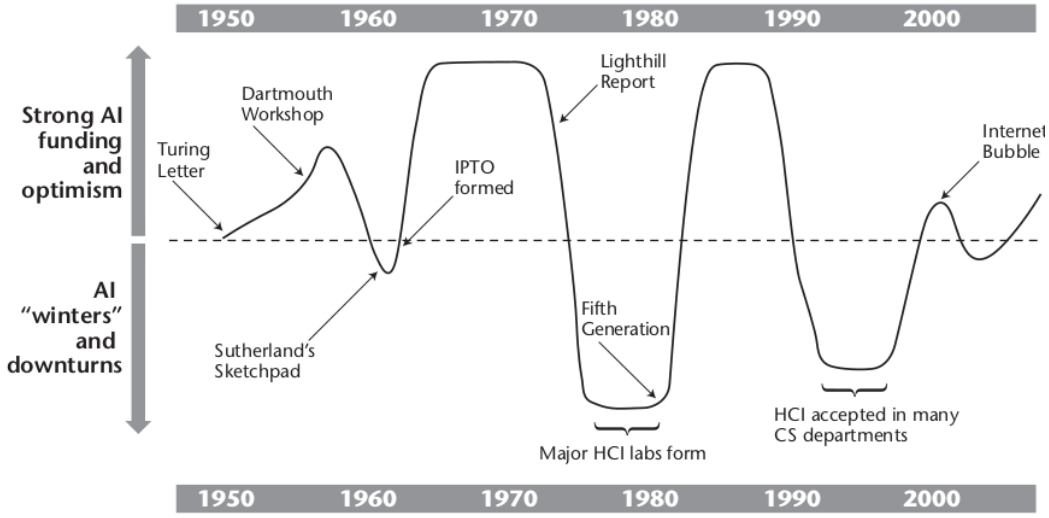

Figure I.1: Les saisons changeantes de l'IA [Grudin, 2009]

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. Aujourd'hui, grâce à plusieurs éléments l'IA retrouve une place dominante, entre autres la quantité de données à disposition a permi de faire exploser ce domaine de recherche.

I.1.2 Le changement des échelles de grandeurs : les données et Big Data

> "Les données sont le nouveau pétrole. Il est précieux, mais s'il n'est pas raffiné, il ne peut pas vraiment être utilisé. Il doit être transformé en gaz, en plastique, en produits chimiques, etc. pour créer une entité de valeur qui stimule une activité rentable ; les données doivent donc être ventilées, analysées pour qu'elles aient de la valeur."
>
> — [Haupt, 2006]

Données : informations, notamment des faits ou chiffres, récoltées pour être étudiées et modifiées afin de faciliter une prise de décision. Les informations numériques sont stockées et exploitées par un ordinateur [Cambridge, 2020].

La compréhension de l'Intelligence Artificielle aujourd'hui, passe par obligatoirement par les données. En effet, les données sont le carburant, le nouveau pétrole[2] L'analogie, bien que pertinente, a une limite : le pétrole est non-renouvelable, les données n'arrêtent pas d'augmenter.

L'augmentation de la quantité de données est liée aux progrès dans les capacités des systèmes de stockage, des techniques pour collecter les données et l'analyse de ces dernières [Press, 2013]. Nous avons assisté à une sorte de Big Bang des informations numérisées lors de ces dernières décénies. Cette explosion est à la fois économique avec plus de 9,8 Milliards de Dollars investis en 2018, représentant une augmentation d'environ 75% par rapport à 2017 [Columbus, 2019]. Mais aussi en terme de volume : l'augmentation est telle que comme pour la loi de Moore[3] la croissance de la quantité de données suit une courbe exponentielle. En effet, en 2018 la "datasphère" mondiale recensait environ 33 zettabytes[4] et selon IDC (International Data Council), en 2025, il y en aura 175 [Reinsel et al., 2018].

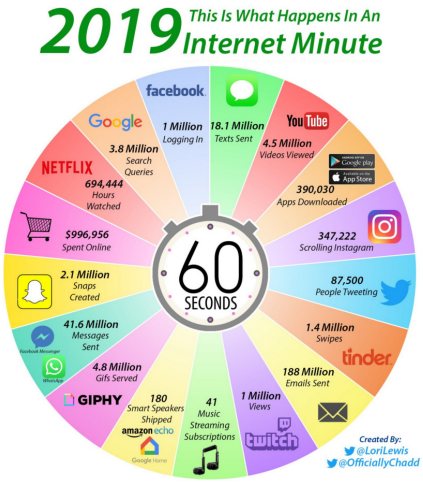

Figure I.2: Ce qu'il se passe sur Internet en une minute en 2019 [Desjardins, 2019]

Cette immense collection de bits[5] est accentuée grâce aux fournisseurs que sont les utilisateurs massifs des grandes plateformes du web. En 2019, chaque minute c'est plus de 188 millions d'emails qui sont envoyés, environ 700 000 heures de vidéo Netflix regardées ou un million d'utilisateurs qui se connectent sur Facebook (voir figure I.2).

Ce volume conséquent de données correspond au nouveau carburant des intelligences artificielles modernes. Les algorithmes utilisés dans ce domaine sont issus de la famille de l'apprentissage de la machine (où "Machine Learning" en anglais).

[2]Cette métaphore a été utilisée par un grand nombre d'experts, mais le crédit de la première citation serait à attribuer à Clive Humby [Haupt, 2006].

[3]En somme, la loi de Moore consiste à dire que la complexité des microprocesseurs double pour une période de temps donnée [Moore, 1998].

[4]1 zettabyte vaut un billion de terabytes.

[5]Unité de stockage à la base de chaque ordinateur, peut uniquement prendre la valeur 0 ou 1.

I.1.3 L'Apprentissage de la Machine

Modèle statistique : C'est une descritption mathématique qui se génère à partir d'observation, où plus précisément dans le domaine de l'intelligence artificielle à partir des données. Il sera souvent utilisé en tant que synonyme d'IA.

Le Machine Learning ou l'apprentissage de la machine est une discipline d'étude de l'intelligence artificielle. Elle se fonde sur des approches mathématiques et statistiques permettant d'apprendre à partir des données fournies. Les données sont délivrées dans un modèle statistique qui généralise des règles afin d'obtenir un résultat. De plus, il est possible de la retrouver presque partout dans notre quotidien : dans les téléphones, les montres connectées ou encore les véhicules.

Le rayonnement de ce domaine est tellement fort que dans notre société capitaliste qui carbure à l'argent, plus de 9,8 milliards de dollars ont été investis en 2018. Cette somme représente une augmentation d'environ 72% par rapport à 2018 [Columbus, 2019].

I.1.3.1 L'Apprentissage Supervisé et Non Supervisé

Label : Un label correspond à la cible de la prédiction : par exemple pour prédire un email est un spam, le label sera le fait qu'un email est un spam ou non.

L'apprentissage supervisé Ce type d'apprentissage correspond au fait de généraliser une règle à partir d'exemples annotés avec des labels. Dans cette catégorie, il y a deux sous-groupes : la classification et la régression.

La classification cherche à prédire une classe, par exemple pour classer un email comme spam ou non. En d'autres termes il s'agit de prédire une information non numérique. Logiquement, la régression correspond à une prédiction d'une valeur numérique comme le prix d'une maison ou alors la température moyenne de demain. Sur la figure I.3, nous distinguons que pour la classification, il s'agit de séparer les classes pour prédire (ligne en pointillé), alors que pour la régression, il est nécessaire d'avoir une fonction qui retourne un nombre (ligne rouge).

L'apprentissage non supervisé Cette catégorie d'apprentissage s'oppose au supervisé car les données ne sont pas annotées avec des labels. Les problématiques d'intelligence artificielle associées sont donc différentes. Généralement il s'agit de regrouper les données en différents groupes (Clustering) ou alors de détecter les éléments anormaux (détection d'anomalies), voir figue I.4.

Bien que différent sur la finalité, ces deux catégories d'apprentissage ont pour point commun leur processus de création car en Machine Learning, il y a une norme sur la fabrication d'un algorithme.

I.1.3.2 Le processus de création d'une IA

Bien que chaque projet de Machine Learning soit différent sur le fond, la forme suit globalement la même logique. La première partie se concentre sur les données (collecte, exploration et

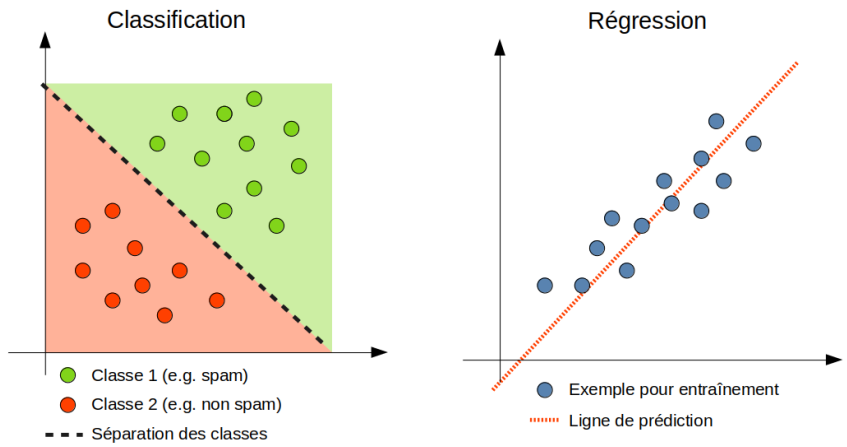

Figure I.3: Représentation de l'apprentissage supervisé.

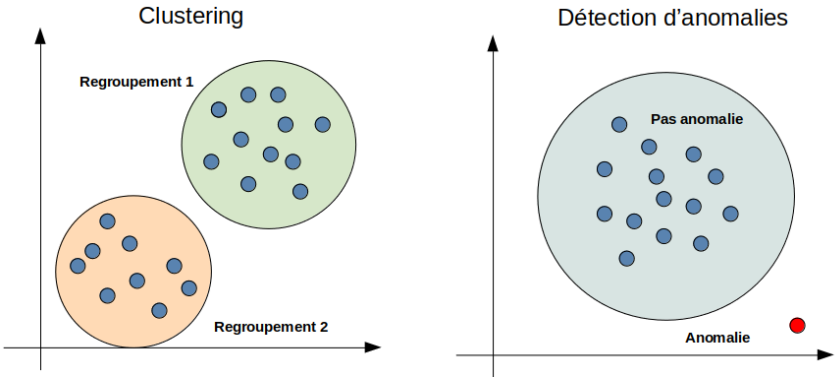

Figure I.4: Représentation de l'apprentissage non supervisé.

préparation) pour ensuite se recentrer sur le modèle[6] qui servira à résoudre la problématique choisie. Les grandes lignes qui vont être décrites en suivant se basent sur les articles de Matthew Mayo (KDnuggets)[7] et de Jeremy Jordan[8].

Définition du problème L'objectif du projet est décidé dans cette section. Tout d'abord une définition en terme métier est nécessaire, soit en langage naturel sans parler technique. Une fois que la problématique est définie une reflexion doit être entamer sur les composants technique du projet. Il s'agit de défnir les données souhaitées, comment les récupérer, quel type d'algorithme va être choisi (classifcation, régression, clustering, etc.), les décisions qui seront prises à partir des prédictions du modèle ou encore comment évaluer la performance du modèle. L'importance de cette étape est surtout sur la question de la faisabilité. Pour y répondre, il y a nécessité de déterminer le coût de l'acquisition des données, le coût de mauvaises prédictions, la quantité de travaux publiés sur un problème similaire et encore si

[6] Un modèle est un algorithme d'intelligence artificielle (modèle mathématique, statistique).
[7] https://www.kdnuggets.com/2018/12/machine-learning-project-checklist.html
[8] https://www.jeremyjordan.me/ml-projects-guide/

l'environnement informatique ne contraindra pas le modèle [Jordan, 2018].

Collecte Cette étape consiste à faire le pont entre la définition des données en langage naturel avec leur localisation physiques (base de données, fichier de données, etc.). Il est important d'être sûr que cette acquisition de données respecte les réglementations qui les concerne (e.g. RGPD[9]). Une fois les données récupérées, s'assurer qu'elles ont le bon type associé [Mayo, 2018].

Analyse des données Cette étape est plus connu sous le nom d'analyse exploratoire des données (Exploratory Data Analysis, EDA en anglais). Son rôle est de comprendre ce qui compose les données, par exemple est-ce qu'il y a plus d'hommes que de femmes. La clé pour bien réussir cette section est la visualisation par des graphiques. Si l'analyse est bien faite, montrer les résultats à d'autres personnes ne connaissant pas le sujet et constater s'ils comprennent l'histoire racontée par ces données est un bon indicateur. De plus, c'est ici qu'il faut s'assurer que les données soient de qualité : qu'elles ne soient pas fausses (e.g. email mal renseigné).

Préparation Pour faire un plat en cuisine avoir les ingrédients et une recette correspond aux première étapes, mais il faut préparer les ingrédients en lisant la recette pour avoir le résultat souhaité. La préparation des données suit la même logique, l'analyse des données permet de savoir comment préparer les données (e.g. supprimer des informations inutiles). Cette étape est plus communément également appelée "Feature engineering" en anglais [Mayo, 2018].

Création du modèle Les étapes précédentes sont cruciales pour permettre la qualité des données avant qu'elles soient digérées par le modèle, en général elles représentent 75% du temps de travail d'un Data Scientist [Figure_Eight, 2019]. Les compétences requises pour réaliser la création du modèle sont orientées vers les mathématiques puisque un modèle est en général un algorithme paramétrable. Quand le modèle est sélectionné, s'en suit l'apprentissage en utilisant les données (C.F. section **??**).

Validation du modèle Une fois entrainée, l'IA doit être testée et validée. Comme pour des étudiants français en terminale passant le baccalauréat, chacun a eu son apprentissage et en fonction de leur note finale, il sera déterminé s'ils ont eu ou non leur diplôme. Ici la note à attribuer correspond à une mesure définie, l'une des plus utilisée en classification est la justesse (accuracy en anglais) des prédictions, soit le pourcentage de prédictions justes (C.F. équation I.1).

$$Justesse = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}} \tag{I.1}$$

Déploiement Le déploiement du modèle fait appel à la casquette technique puisqu'il s'agit de rendre le modèle utilisable sur le long terme. Cette étape doit créer une "tuyauterie" reliant les données au modèle qui offre la prédition.

[9]Le Réglement Général sur la Protection des Données est une loi en vigueur depuis mai 2018 dans les pays de l'UE.

Contrôle Enfin, une étape cruciale, mais qui est souvent oubliée, celle de la surveillance et du maintien de l'IA. Selon Figure Eight, seulement 37% des développeurs maintiennent leurs modèles en permanence [Figure_Eight, 2019]. Si un problème est identifié, par exemple un nouveau comportement chez les clients que le modèle ne connait pas, il faut déterminer s'il faut juste ré-entrainer le modèle ou le cas échant, recommencer le processus de création de l'IA.

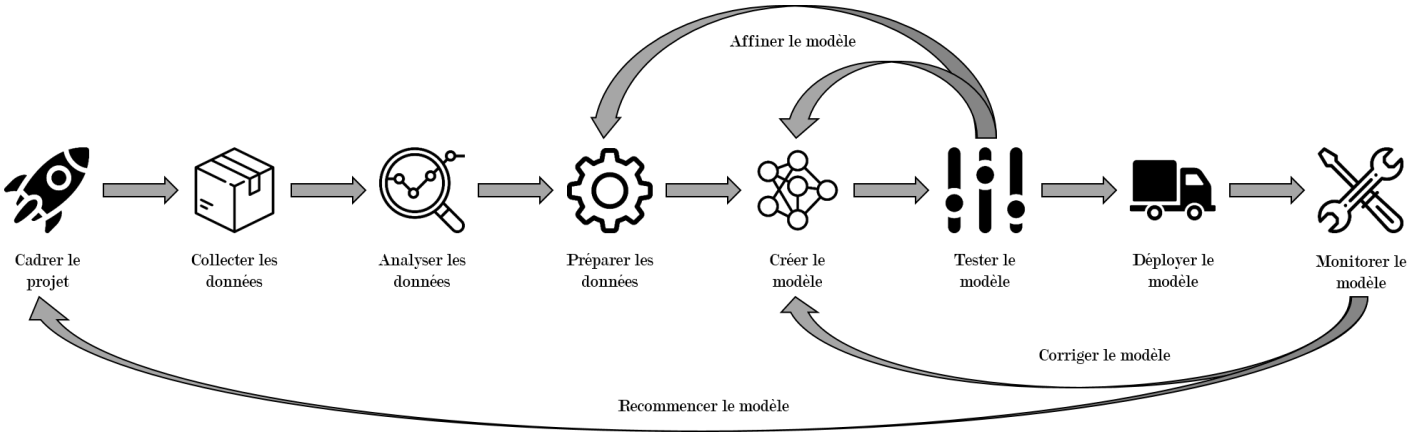

Figure I.5: Pipeline d'un projet de Machine Learning "classique". Crédit icônes Flaticon : mynamepong, Freepik, Gregor Cresnar et Becris.

I.1.4 Fiction ou Réalité : l'Intelligence Artificielle plus forte que les Humains

> "Si vous programmez une machine, vous savez de quoi elle est capable. Si la machine se programme elle-même, qui sait ce qu'elle peut faire ?"
>
> — [Kasparov, 2017]

En 1997, le champion du monde des échecs, Garry Kasparov, perd face à une machine nommé Deep Blue [Krauthammer, 1997]. Cette victoire a impressioné et à fait fantasmer autour de la véritable notion d'intelligence d'une IA. Le fonctionnement de l'algorithme n'était pas surhumain, en réalité il s'agissait 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 et al., 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 humain (même le champion du monde des échecs) cela est une tout autre histoire.

C'est un peu plus d'une décénie plus tard qu'une véritable révolution remet de nouveau la place de l'intellect humain par rapport à celui des machines en question. En 2012, lors de la compétition du ILSVRC[10], qui est une comptétition visant à reconnaitre ce qu'il y a sur

[10]ImageNet Large Scale Visual Recognition Challenge

une image par un algorithme, un modèle a triomphé amplement en passant de 25% d'erreur en 2011 à seulement 16% l'année suivante [Russakovsky et al., 2015]. La raison de cette progression : l'apprentissage profond où "Deep Learning".

I.1.4.1 L'avènement de l'Intellect des Machines : Deep Learning

L'apprentissage profond tient son nom de la profondeur des couches de neurones qui le compose. Un modèle de Deep Learning est donc un réseau de neurones. En tant qu'être humain, nous avons envie de faire un amalgame avec le 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"[11].

A l'origine des neurones un algorithme qui se nomme le "Perceptron" [Rosenblatt, 1958], il s'agit là du réseau de neurones le plus simple qui retourne deux sorties possible : 1 ou 0. L'algorithme fonctionne de la façon suivante : N entrées $x_1, x_2, ..., x_N$ de les multiplier par un poid associé $w_1, w_2, ..., w_N$ puis d'en faire la somme pour enfin appliquer une fonction qui détermine si la sortie doit être 1 (activée) ou 0 (désactivée). Au vu de ce qui compose ce "neurone" (voir Figure I.6), il est difficile de penser qu'il s'agit là d'un algorithme bien plus complexe que le cerveau.

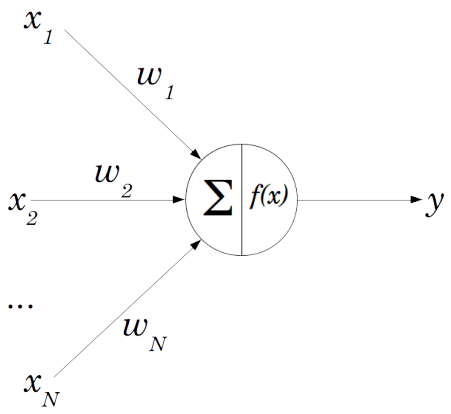

Figure I.6: Schéma de l'algorithme du Perceptron [Rosenblatt, 1958]

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

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

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].

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 et al., 2015], reproduire des visages humains [Karras et al., 2018], imiter un discours d'un président américain [Suwajanakorn et al., 2017], ces 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.

I.1.4.2 Les Intelligences Artificielles de demain : la singularité

> "Depuis 130 000 ans, notre capacité à raisonner est restée inchangée. L'intelligence combinée des neuroscientifiques, des mathématiciens et des hackers est bien faible par rapport à l'intelligence artificielle la plus élémentaire. Une fois en ligne, une machine sensible dépassera rapidement les limites de la biologie. Et en peu de temps, sa puissance analytique deviendra supérieure à l'intelligence collective de chaque personne née dans l'histoire du monde. [...] Certains scientifiques appellent cela la singularité."

— Will Caster (Johnny Depp), Transcendence, 2014

Singularité : C'est l'hypothèse selon laquelle une Intelligence Artificielle dépasserait les capacités intellectuelles des humains et par conséquent cela aura pour répercution des changements impévisibles sur la société [Eden et al., 2013].

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 (voir figure I.7).

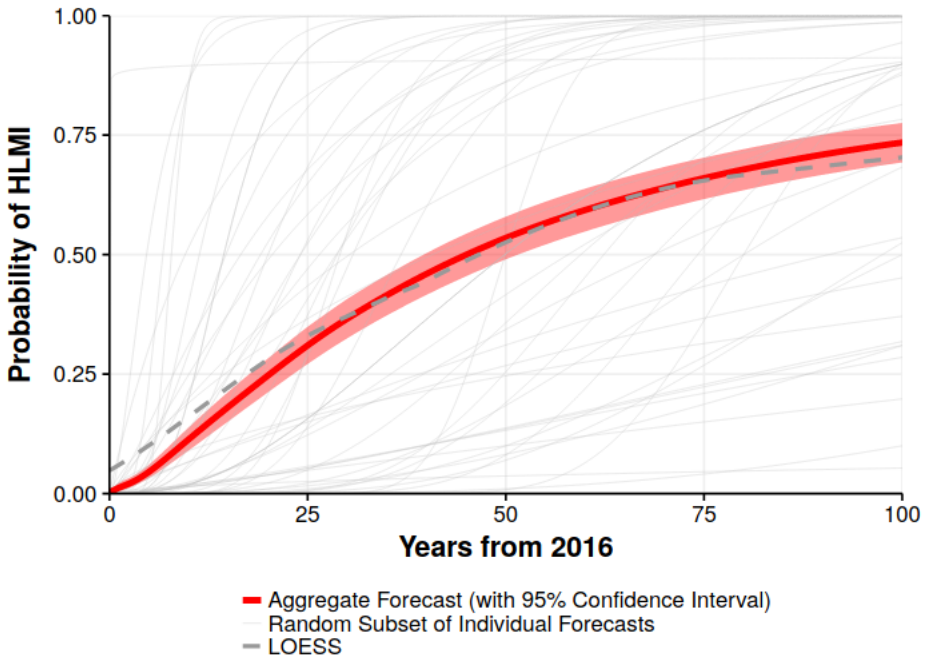

Figure I.7: Probabilité d'une super-intelligence à partir de 2016 [Grace et al., 2017]

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.

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 poussent 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.

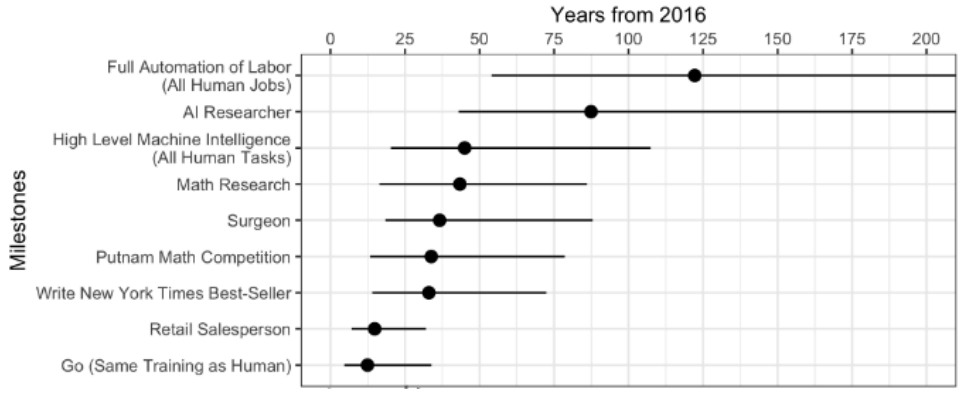

Figure I.8: Chronologie des estimations qu'une IA achève des tâches humaines [Grace et al., 2017]

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.

I.2 L'Éthique, la science de la Morale

Le bien et le mal sont souvent deux notions assez abstraites, 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 et le groupe, un comportement individualiste diminuait les chances de survie [Harari and Dauzat, 2015].

L'éthique, 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[12] [Sumner, 1906], une généralisation abusive de nos critères moraux. La morale est le reflet d'un contexte social et temporel. Un exemple frappant est celui du 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. C'est pourquoi, pour définir une morale, il faut revenir sur l'origine du bon, du méchant ou encore du mauvais caractérisant le bien et le mal. En d'autres termes, il faut regarder comment se compose une morale, en philosophie.

I.2.1 L'origine du Bon, Histoire et Philosophie : Nietzsche

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[13] 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. Nietzsche [Nietzsche, 1900] dans son livre “Généalogie de la morale”. Ce texte cherche l'origine à la fois historique et psychologique. De nos jours, lorsqu'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”, Nietzsche 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 majoritaire, des personnes semblent en être adeptes.

Toujours dans le même chapitre de la “Généalogie de la morale”, Nietzsche 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.

[12] 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

[13] 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.

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.

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 ?

I.2.2 Philosophie, l'Éthique Normative

En philosophie, l'éthique normative correspond à la branche théorique comprenant les théories évaluant moralement les humains et leurs actions à partir de critères composants les théories associées. Évaluer la morale d'une personne semble être complexe dans les faits, mais quotidiennement les humains effectuent des choix moraux.

Dans ce domaine de recherche, il y a trois modes principaux d'évaluation morale : l'éthique de la vertue, la morale déontologique et enfin le conséquentialisme. L'objectif de ces théories visent à répondre à ce qui fait qu'une action bonne ou mauvaise.

I.2.2.1 Aristote et l'Éthique de la vertu

> "La vertu morale est fille des bonnes habitudes [...] Ce n'est ni par un effet de la nature, ni Contrairement à la nature que les vertus naissent en nous ; nous sommes naturellement prédisposés à les acquérir, à condition de les perfectionner par l'habitude. [...] Nous les acquérons d'abord par l'exercice, comme il arrive également dans les arts et les métiers. Ce que nous devons exécuter après une étude préalable, nous l'apprenons par la pratique ; par exemple, c'est en bâtissant que l'on devient architecte [...] De même, c'est à force de pratiquer la justice, la tempérence et le courage que nous devenons justes, tempérants et courageux. "
>
> — [Aristote, 1994]

Dans ce passage du livre "Ethique à Nicomaque" rédigé par Aristote, il est décrit le fond de ce qu'est l'éthique de la vertu. Cette théorie est axée autout de la personne, la question principale dans ce cadre est "Comment devenir un bonne personne ?".

Dans la pratique cela signifie qu'une personne, agissant par intérêt, par exemple pour de la reconnaissance, tant que les actions effectuées sont vertueuses, alors elle pourra devenir "bonne".

Se forcer à adopter un comportement vertueux, au sens d'un comportement qu'une personne véritablement vertueuse adopterait naturellement, à pour conséquence de prendre l'habitude de ce comportement. La finalité est que ce comportement devient naturel et donc vertueux.

I.2.2.2 La Morale Déontologique : Kant et le mensonge

En opposition à l'éthique de la vertu d'Aristote, la morale déontologique affirme qu'une personne fait de bonnes actions car elle est bonne et non pour obtenir une récompense, la motivation doit être morale.

Venant du grec “deon” signifiant devoir (la science du devoir), ce mouvement de pensée de la philosophie morale explique qu'il existe des devoirs moraux absolus soit sous la forme verbale “tu dois”, et ce, sans rajouter une explication à ce devoir[14]. 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 and 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. Il ne repose donc pas sur quelque chose de factuel et par conséquent, il peut être difficile de pouvoir arbitrer sur des actes immoraux.

I.2.2.3 Le Conséquentialisme

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 comme prémisse[15] 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 quelle 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 sanctionne 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 conséquentialiste 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, une vision idéologique propre de la morale, son code d'éthique. C'est un processus qui commence dès la naissance avec notamment la théorie du développement moral [Kohlberg and Hersh, 1977], puisque notre appartenance à un groupe social défini déjà une partie de ce dont il adviendra de nos croyances, puis l'éducation rentre en jeu, l'entourage, etc.

[14] Par exemple, “tu dois aider ton prochain”, mais sans rajouter une justification.

[15] Il s'agit d'une proposition avancée afin de supporter une conclusion, dans le cas de l'altruisme : optimiser le bonheur de tous.

I.2.3 La Morale en Science

L'approche normative que nous venons de voir, cherche à déterminer les codes moraux qui nous régissent, or, il existe une autre approche : l'approche descriptive. Cette approche pose en precept le fait qu'il n'y a pas de morale absolu, de normes, elle cherche à décrire les compotements des humains pour mieux comprendre leur morale.

Si nous revenons sur ce qu'est la morale, il est possible de constater que la morale s'impose contre notre intérêt individuel. Par exemple, lorsque nous trouvons un porte-feuille dans la rue, le ramener à son propriétaire n'est pas avantageux à titre individuel ou encore un criminel qui se rend par culpabilité.

Les deux approches qui s'affrontent dans l'éthique descriptive sont la culturalisme qui est associé au relativisme culturel et l'hypothèse naturaliste qui fait intervenir de la biologie pour mieux comprendre le comportement moral humain.

I.2.3.1 L'Approche par la Culture et la Personnalité : le Culturalisme

> "La civilisation n'est pas quelque chose d'absolu, mais elle est relative, et nos idées et nos conceptions ne sont vraies que dans la mesure où notre civilisation continue."
>
> — [Boas, 1887]

Le culturalisme est pratiqué par la plupart des anthropologues et ethnologues de nos jours. Ce domaine de recherche se base sur la constatation que ce que nous croyons est spécifique à une culture, à l'environnement qui nous entoure. Ce mouvement ne classe pas les différentes civilisations, mais les étudie et décrit selon des critères objectifs : leurs pratiques, récits ou témoignages sans emettre un jugement de valeurs moral [Servier, 1993].

Le culturalisme est très accepté puisqu'il est défini par ce qui nous entoure et la différence entre morale selon le pays, les croyances ou le statut social renforce le fait qu'une morale dépend de la culture. Il nous suffit de regarder des recherches d'anthropologues pour constater cette différence.

Par exemple, au XIXe siècle, les Inuits enfermaients les membres de leur tribu devenu trop âgés en les laissant mourir de faim et de froid sur place car il était compliqué de survivre dans de tels climats avec ces personnes [Redfield, 1965]. A Madagascar, une tribu, celle des Vezo, s'interdit de pointer du doigt une baleine ou encore les membres de cette tribu s'interdisent de rire en mangeant du miel [Astuti, 2007].

Ces quelques exemples montrent bien que l'environnement, la culture dans laquelle nous évoluons influe notre morale et par conséquent notre comportement. Mais est-ce là le seul facteur ? Les chercheurs étudiant l'hypothèse Naturaliste pensent qu'il s'agit d'un élément important, mais pas le seul. Regardons ce que donne l'étude de la morale si nous y ajoutons de la biologie.

I.2.3.2 Le cerveau précablé pour la Morale : l'hypothèse Naturaliste

Au XXIe siècle, Montaigne défend le culturalisme en écrivant :

> "Ici on vit de chair humaine ; là c'est office de piété de tuer son père en certain âge ; ailleurs les pères ordonnent des enfants encore au ventre des mères, ceux qu'ils veulent être nourris et conservés, et ceux qu'ils veulent être abandonnés et tués."
>
> — [Montaigne, 1789]

Mais, deux siècles plus tard, Rousseau complète ses propos en défendant que toutes morales ont des points communs :

> "O Montaigne ! toi qui te piques de franchise et de vérité, sois sincère et vrai, si un philosophe peut l'être, et dis-moi s'il est quelque pays sur la terre où ce soit un crime de garder sa foi, d'être clément, bienfaisant, généreux ; où l'homme de bien soit méprisable, et le perfide honoré."
>
> — [Rousseau, 1969]

Au travers de cet échange entre philosophes, nous avons ici un exemple de divergence entre le culturalisme et le naturalisme. Fondamentalement l'hypothèse naturaliste défend le fait que le cerveau humain serait pré-cablé pour permettre d'émmettre des jugements moraux.

Pour mieux comprendre ce pré-cablage, des expériences ont été réalisées sur des enfants puisque, si des comportement moraux sont trouvés chez de jeunes enfants, alors il semble cohérent de se dire que plus un jugement moral arrive tôt, plus il y ai de chance que cet évènement est des origines biologique. Par exemple, dès l'âge de trois ans, des enfants sont capables de faire de la justice proportionnelle[16] [Baumard et al., 2012]. De plus, certains bébés, à peine âgés de six à sept mois préfèrent intéragir avec une marionnette qui en aide une autre, plutôt qu'une qui en embête une autre [Hamlin et al., 2007].

L'hypothèse du naturalisme peut être également appuyée par des expériences inter-culturelles, c'est-à-dire dans différents pays autour du monde. Des sociétés de chasseurs cueilleurs utilisent le même principe de justice proportionnelle qui est à la base de nos systèmes juridiques [Gurven, 2004].

Nous pouvons rajouter le critère de la génétique : des études menées sur de vrais jumeaux[17] ont montrées que les gènes ont une importance dans l'idéologie politique d'une personne [Bouchard and McGue, 2003].

[16] La justice proportionnelle signifie de récompenser plus ceux qui travaillent plus.

[17] Il s'agit de jumeaux avec le même code génétique, dans ces études les jumeaux ont été séparés à la naissance et adoptés dans différentes familles par exemple.

Ici, le “sens moral” d’une personne peut être comparé à un algorithme dans le cerveau : à partir d’entrée comme les croyances et la connaissance sur un sujet en particulier la morale donnera une sortie. Or, ce n’est pas le seul “algorithme” dans notre cerveau et il arrive que la sortie soit en conflit avec les autres : il s’agit là d’un dilemme moral. Cette explication permet d’expliquer la variabilité entre les différents comportements moraux dû à la culture. En effet, prenons l’exmple de la cigarette : il y a quelques décennies elle était autorisée dans les lieux publiques, mais aujourd’hui ce n’est plus le cas, notre connaissance sur les cigarettes a changé, par conséquent les entrées du “sens moral” ont changées modifiant la sortie et notre comportement dû au tabac.

Chaque humain a une connaissance du monde qui le feront agir pour une cause plutôt qu’une autre. Il est possible d’interpréter la morale d’un individu, mais concernant une machine intelligente, si la singularité est atteinte, le questionnement sera de savoir si la morale de l’algorithme correspond à celle de ses créateurs ou à une morale qui lui est propre.

I.3 La question de l’Éthique pour les algorithmes intelligents

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. Se questionner sur 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’agisse de la vision éthique de l’humain ou de la société en charge.

Bien définir ce qu’est l’éthique pour une intelligence artificielle n’est pas chose facile, mais les codes moraux qui nous entourent ainsi que les lois permettent d’avoir une intuition (subjective) par rapport à cette problématique. Puisqu’il n’y a pas de morale universelle (C.F. section I.2), il n’est pas possible d’avoir une définition de principes éthiques, objectivement applicables selon les différentes croyances morales.

I.3.1 Les problèmes du Présent

> “À mesure que les systèmes intégrant des technologies d’IA envahissent notre quotidien, nous attendons légitimement qu’ils agissent conformément à nos lois et normes sociales.”
>
> — [Villani, 2018]

L’éthique au sein du domaine des IA, de nos jours, correspond à plusieurs notions : transparence du modèle 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].

I.3.1.1 La Discrimiation des algorithmes : les biais

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. Parmi les différents biais pour les machines, trois types attirent l'attention.

Le premier, se baptisant le biais de sélection, il s'agit du manque de diversité des données sur lesquelles l'IA apprend. 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[18] 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. Un exemple célèbre date de 2016, Microsoft sortit un compte Twitter sous l'appellation "@TayandYou" plus connu sous le nom TAY signifiant "Thinking about you"[19] . Son objectif était de converser avec les utilisateurs du 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 and 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[20], 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 and Angwin, 2016].

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

I.3.1.2 Ouvrir la boîte noire : Interprétabilité et Explicabilité

> "Si vous vous concentrez uniquement sur la performance, vous obtiendrez automatiquement des modèles de plus en plus opaques. Jetez un coup d'œil aux interviews des gagnants sur la plate-forme du concours d'apprentissage automatique de kaggle.com : Les modèles gagnants étaient pour la plupart des ensembles de modèles ou des modèles très complexes tels que des arbres boostés ou des réseaux neuronaux profonds."
>
> — [Molnar, 2019]

[18] Un poids, dans un modèle, correspond à atténuer ou amplifier l'influence d'une variable.
[19] Traduction : Pensant à toi.
[20] Correctionnal Offender Management Profiling for Alternative Sanctions : c'est une IA qui classifie le risque de récidivisme d'un criminel.

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 ces machines est primordiale afin de pouvoir avoir confiance en elles.

Dans un premier temps, les concepteurs des algorithmes (souvent ayant des compétences en mathématiques et en informatique) 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].

La figure I.9 illustre en abscisse l'interprétabilité d'un modèle et en ordonnée sa performance, la corrélation à noter 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.

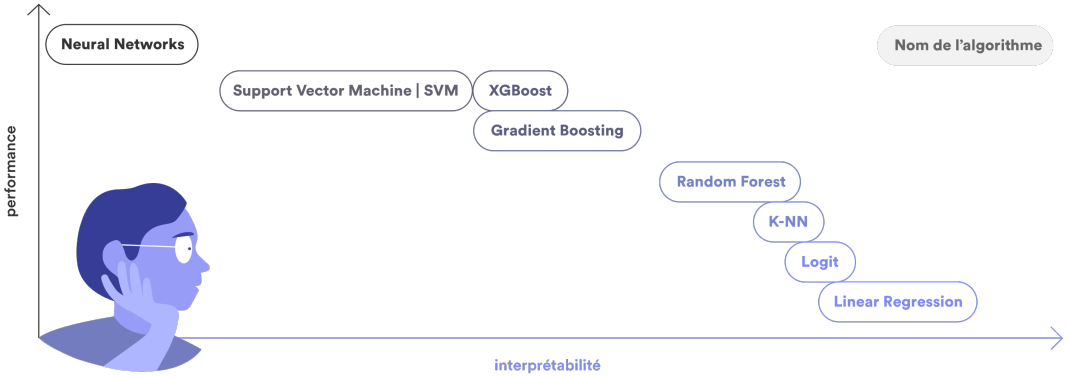

Figure I.9: Interprétabilité d'un algorithme [Data for Good, 2018]

Alors, la résignation d'avoir des algorithmes compréhensibles par l'humain est-elle nécessaire ? Non, rentre en jeu la notion d'explicabilité d'un modèle. Ce terme renvoie 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[21]. Cela est bien problématique quand il s'agit d'algorithmes réalisés en entreprise et ne voulant pas publier leur réalisation au grand public.

[21] L'open source est une philosophie de développement désignant le fait d'ouvrir le code à tous, librement.

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"[22] et difficilement discernables par l'humain. La justification correspond au 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 notre 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.

I.3.1.3 Les créateurs des IA, architectes de notre quotidien

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[23], 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 et al., 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.

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 ai une fixation sur une catégorie précise de vidéos (poussant alors le biais de confirmation s'il s'agit d'une idéologie). Mickaël Launay soutient, qu'en mathématiques, un tel milieu ségrègue les utilisateurs en deux ensembles : les conformistes et anti-conformistes [Launay, 2012]. Les conformistes auront uniquement

[22] Traduction : Informations fausses. Il s'agit de nouvelles pouvant sembler vraies dans la forme, mais fausses dans le fond.

[23] Développeur.euse en charge de créer le modèle qui composera l'intelligence artificielle.

des vidéos qui leur correspondent, 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 quotidienne est irréfutable. Bien que l'objectif de ses algorithmes est 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 ces 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 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. Chaque acteur intervenant lors de la création d'un algorithme intelligent doit agir pour tendre vers une intelligence artificielle de confiance.

I.3.2 Une IA de confiance selon la Commission Européenne

Le domaine de recherche sur l'intelligence artificielle et l'éthique est assez récent et nous pouvons même recenser jusqu'à plus 70 documents ont été publiés entre 2016 et 2019 [Algorithm Watch, 2020]. Les différents auteurs sont principalement des gouvernements, des académies et des entreprises [Morley et al., 2019]. Le but des auteurs de ces documents est qu'en définissant des principes théoriquement, de façon abstraite, ces principes permettront d'agir comme contraintes normatives [Turilli, 2007] sur ce qui doit être fait et ce qui ne doit pas être empêché pour l'utilisation d'un algorithme. En d'autres termes, il s'agit de principes qui permettront de garantir une IA éthique.

Parmi les documents publiés, nous allons nous intéresser à celui qui a été écrit par un groupe expert indépendant sur le sujet de l'IA et qui nous concerne directement car il a été rédigé par la Commission Européenne : "Lignes directrices en matière d'éthique pour une IA de confiance" [Commission, 2019]. En prémisse du document, il est stipulé que pour obtenir une intelligence artificielle de confiance il est nécessaire de respecté trois caractéristiques :

- Licité : respect de législations et réglementations applicables.[24]
- Éthique : assure l'adhésion de valeurs et de principes éthiques.
- Robuste : garantie sur le plan technique et social d'éviter de causer des préjudices involontaires.

Basé sur ces fondements caractérisants une IA de confiance, sept exigences dites "essentielles" sont décrites qui permettent de diriger vers un idéal. Les concepts détaillés reste théoriques. Ces différentes exigences sont présentées dans la Figure I.10.

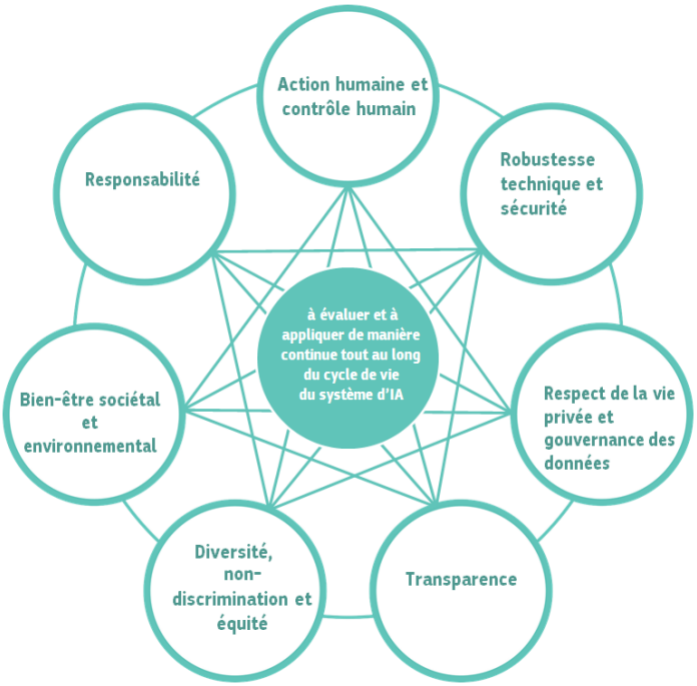

Figure I.10: interrelation des sept exigences: elles revêtent toutes une importance égale, elles se soutiennent mutuellement et devraient être appliquées et évaluées tout au long du cycle de vie d'un système d'IA. [Commission, 2019]

1. Action humaine et contrôle humain Les systèmes d'algorithmes intelligents se doivent d'être centré autour des humains : ils doivent respecter l'autonomie humaine en respectant les Droits Fondamentaux voire ces systèmes, s'ils le peuvent, devrait les renforcer. Les femmes et hommes doivent garder l'action sur l'IA en prenant par exemple des décisions

[24]Le document de la Commission Européenne ne traite pas l'aspect licite.

à partir de l'algorithme tout en pouvant protester face à la décision d'une machine. Enfin le contrôle humain est essentiel par l'approche "human-in-the-loop", "human-on-the-loop" ou "human-in-command" qui signifient respectivement : l'humain qui intervient dans le processus, qui supervise le processus et qui reste aux commandes.

2. Robustesse technique et sécurité Cette exigence est liée au principe de prévention de toute atteinte. Pour assurer une robustesse technique il faut garantir un système d'information sécurisé et résilient. De plus, une IA peut être soumises à des attaques comme par exemple la modification des données en entrée pour avoir un meilleur résultat. Ces attaques doivent être anticipées pour être évitées. La Commission ajoute que la mise en place de plans de secours et la sécurité générale de l'environnement dans lequel évolue l'algorithme sont à assurer. Enfin, un modèle qui est déployé doit être valider par des mesures de précision et les données sur lesquelles il s'est entraînées doivent être fiables.

3. Respect de la vie privée et gouvernance des données Il s'agit là d'un droit fondamental et le Règlement Général de la Protection des Données (règlement numéro 2016/679) rentre dans ce cadre en stipulant que les données d'un clients doivent être protégées. De plus, si l'algorithme assimile une grande quantité de données pour faire ses prédictions, il y a un risque qu'il arrive à déduire de lui même des données personnelles (voire sensibles) comme l'âge, l'orientation sexuelle, le sexe ou encore les opinions politiques d'une personne. Dans la continuité, les données doivent être de qualité et leur intégrité doit être préservée.

4. Transparence Cette exigence est liée au principe de l'explicabilité (voir section I.3.1.2), c'est à dire à pouvoir comprendre le modèle qui a été fabriqué. Pour assurer cela, le cycle de vie de la donnée doit être traçé et documenté. Un modèle a besoin de respecter l'explicabilité : c'est-à-dire qu'il doit être possible de fournir une explication en langage naturel. De plus, si pour une problématique donnée il est possible de choisir un type d'algorithme considéré comme plus interprétable et qui obtient des performances similaires, alors il est préconisé de le choisir. Pour fini, la communication n'est pas à négliger pour que les utilisateurs comprennent de ce qu'il en retourne concernant le modèle.

5. Diversité, non-discrimination et équité Le respect du principe d'équité est essentiel pour permettre d'avoir une IA de confiance. Un algorithme se doit d'être accessible avec une conception universelle peut importe l'âge, le sexe, les capacités ou les caractéristiques d'un humain. Comme discuté dans la section I.3.1.1, l'absence de biais injuste est un critère fondamental dans le cas où l'algorithme peut avoir un préjudice pour un individu (comme l'obtention d'un crédit ou le recrutement dans une entreprise). À titre préventif, il est souhaitable de faire participer les parties prenantes concernées par l'IA en question.

6. Bien-être sociétal et environnemental Cette exigence est liée à la précédente puisque l'omniprésence des algorithmes sur nos réseaux sociaux et dans notre quotidien ne devraient pas avoir d'incidences sociales sur nos relations et liens sociaux. Également, l'aspect environnemental avec la quantité de données utilisées et le temps d'entraînement d'un modèle, est à ne pas négliger car pour certains modèles la consommation énergitique peut être très importantes.

7. Responsabilité Lorsqu'une IA est mise à disposition en publique ou bien mise en production, il faut anticiper les problèmes potentiels avec la possibilité d'auditer l'algorithme,

ce qui oblige à bien documenter. En cas d'incidences négatives, il convient de garantir la capacité de les minimiser et de les documenter pour éviter des problèmes similaires dans le futur. Enfin si une incidence négative à lieu pour une personne donnée, il convient de préparer des mécanismes pouvant permettre de faire recours sur la décision.

Ces exigences bien que non exhaustifs, ils ont le bénéfice de couvrir une grande partie de la théorie sur ce que devrait être une IA éthique et de confiance. La Commission propose en plus des exemples de méthode pour assurer une IA de confiance en cohésion avec les exigences décrites précédemment. Ces méthodes sont découpées en deux catégories : méthodes techniques et non techniques (pages 26 à 29).

Méthodes techniques : Architectures pour une IA digne de confiance , Éthique et état de droit dès la conception (X dès la conception), Méthodes d'explication, Essais et validations et Qualité des indicateurs de service.

Méthodes non techniques : Réglementation, Codes de conduite, Normalisation, Certification, La responsabilité au moyen de cadres de gouvernance, Éducation et sensibilisation pour encourager un état d'esprit éthique, Participation des parties prenantes et dialogue social et Diversité et équipes de conception inclusives.

En conclusion, la Commission reconnait "les effets positifs que les systèmes d'IA ont déjà et continuerons à avoir, tant d'un point de vue commercial que pour la société." Mais insiste sur l'importance de mettre au point des IA dignes de confiance, d'autant plus dans un contexte où les machines sont omniprésentes et prennent des décisions pouvant impacter notre vie profondémment.

I.3.3 Les Dilemmes Moraux pour les Machines

L'émergence des intelligences artificielles et la possibilité du passage de la singularité fait apparaître la difficulté à aligner la morale de l'homme à celle de la machine : c'est le problème de l'alignement, comment une machine devant un dilemme moral devrait raisonner et agir.

I.3.3.1 Expliciter la morale des humains : Moral Machine

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 l'écriture noir sur blanc de "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'elles soient recentrées sur l'atteinte et l'obéissance à l'être humain, le raisonnement est à pousser au maximum.

Une expérience de pensée assez célèbre pour illustrer cette nécessité serait le cas d'une voiture autonome, dans un contexte où les freins ne fonctionneraient pas en face de piétons, qui la voiture devrait choisir de tuer. 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 (un exemple sur la figure I.11). 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].

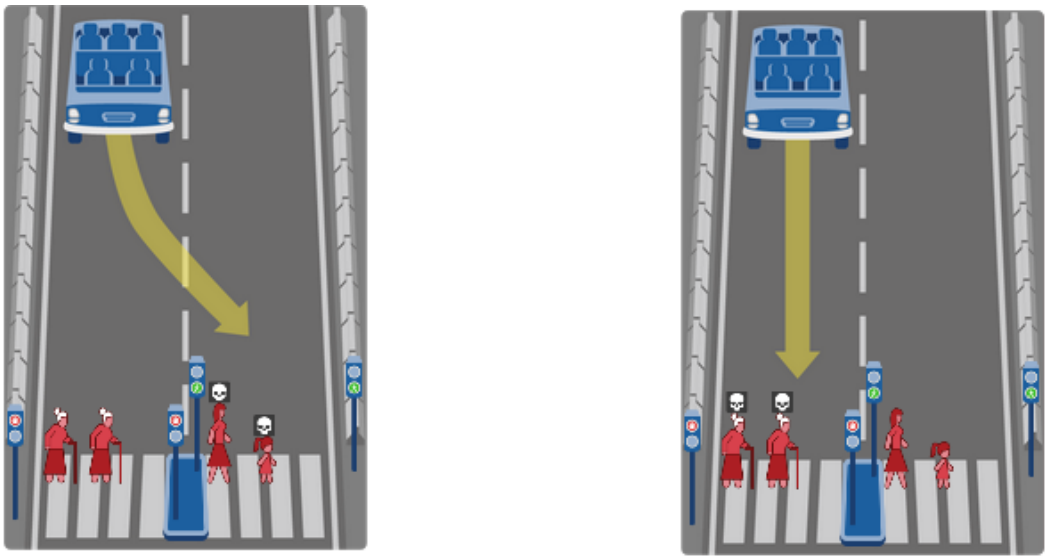

Figure I.11: Expérience "Moral Machine" du MIT, ici le choix se pose entre deux femmes agées ou une petite fille accompagné d'une femme.

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 and 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[25] 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

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

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 and 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 and Askell, 2019].

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.

Tous ses problèmes que soulèvent les algorithmes intelligents, bien heureusement, suscitent un intérêt pour la communauté de chercheurs ainsi que pour les instances gouvernementales. Une volonté commune est née avec pour objectif de théoriser ce que peut être une intelligence artificielle éthique.

I.3.3.2 La décision par l'Aléatoire : Alexei Grinbaum

> "Les choix éthiques ne doivent être fait que par les Hommes."
>
> — [Grinbaum, 2019b]

Dans son livre "Les robots et le Mal" [Grinbaum, 2019c], Alexei Grinbaum, philosophe et physicien, propose une méthode de résolution de conflits moraux lorsqu'une machine se doit de faire un choix moral : faire ce choix en le tirant au sort.

Dans les premières pages de son livre, il évoque la première de couverture du journal "The Economist" du 19 décembre 2009 qui représente Adam et Eve du mythe tenant un appareil Apple (voir Figure I.12). Ce couple correspond à la métaphore du bien et du mal, de la condition humaine, cette couverture a pour sens le fait de soulever la question de comment la machine modifie la condition humaine. Il en va de soit que nous pouvons le pousser jusqu'à l'intelligence artificielle.

Si nous revenons sur les résultats de l'étude étudiée dans la section précédente, nous avons constaté que le comportement d'une voiture autonome devrait varié en fonction des cultures et des moeurs. Cela semble cohérent, mais est-ce que la morale d'une voiture autonome doit être faite par pays, par catégorie sociale ? De plus, est-ce qu'une voiture autonome sur Paris avec un français à l'intérieur devrait-elle se comporter différement qu'avec un passager brésilien ou australien ?

Figure I.12: Première de couverture de l'édition du 19 décembre 2009 de "The Economist" TheEconomist [2009].

Ces questions poussent à définir exactement une morale en fonction, mais cela relève de l'impossible car le nombre de possibilité de conflits moraux est immenses. De plus, que devrait prendre en compte la fonction qui ferait le choix de favoriser une personne plutôt qu'une autre ? Son cercle d'ami, sa famille, son apparence ou ses goûts musicaux ? La complexité d'une telle fonction peut également remettre en cause la responsabilité d'une machine sur le plan légal.

Alexei Grinbaum affirme qu'il ne pense pas "qu'une machine puisse être considérée comme une personne juridique puisque ce n'est pas une personne" [Grinbaum, 2019a], en effet il propose de plutôt partager la reponsabilité entre les agents humains qui interviennent lors de la création de l'algorithme : les programmateurs, la société qui produit l'IA, l'utilisateur qui influe sur la décision de l'algorithme ou encore sur les personnes sélectionnant les données qui serviront à entraîner le modèle.

Alors l'une des solutions envisageables serait de prendre une décision basée sur un tirage au sort. Bien que les humains n'aiment pas vivre dans l'incertitude, l'aléatoire permet de résoudre les conflits moraux sans préjugé ni préjudice. Prenons le cas du dilemme pour lequel une voiture autonome pourrait avoir : qui sauvé le cas échéant. Un tirage au sort permettra de ne pas remttre en cause la malveillance de la machine, même si les conséquences peuvent en être dramatiques.

A. Grinbaum rappelle qu'une situation de conflit moral reste assez rare et que par conséquent l'impact dans une société de l'application par le hasard restera minime globalement. De plus, la difficulté deviendrait de créer un système capable detecter un conflit moral pour pouvoir le résoudre de façon aléatoire. Nous pouvons dire que le calcul moral qui est à priori très complexe se réduit à un jet de dés.

I.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 I.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 chances d'être atteinte d'ici cinquante ans (voir Figure I.8).

Dans un second temps, une introduction à la morale, avec entre autres son origine historiquo-philosophique vu par Nietzsche (voir section I.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 I.2.2). Enfin, nous avons vu la vision de la morale dans la Science, notamment avec l'approche du culturalisme et celle du naturalisme I.2.3.

C'est pour cela qu'en dernier lieu, la section I.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 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. Sans oublier la théorie de ce que devrait être une IA éthique au travers des écrits de la Commission Européenne.

La problématique peut alors se découper sur deux points de l'éthique : une IA socialement responsable et la supervisation de son éthique.

Déterminer si une intelligence artificielle est socialement responsable revient à déterminer les impacts sociaux de cette dernière, qui 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 supervisation de l'éthique d'une IA a pour différence d'être plus technique, décliner tous les aspects des critères décrits par la Commission Européenne afin de répondre à la question : "Est-ce cette intelligence artificielle est éthique ?"

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, je présente les solutions existantes et comment les regrouper dans un outil pouvant garantir une IA de confiance.

Chapitre II

Solution : une boîte à outils transparente

Dans la partie précédente, nous avons évoqué les travaux théoriques concernant la notion d'intelligence artificielle éthique (ou de confiance selon la Commission Européenne). La Commission n'est pas la seule organisation qui a publié des principes pour garantir l'éthique dans un système intelligent (CF section I.3.2).

> "La disponibilité de ces principes "convenus" encourage mais n'entraîne pas encore de changement réel dans la conception des systèmes algorithmiques. Comme on peut le constater, la quasi-totalité des lignes directrices produites à ce jour laissent penser que des solutions techniques existent, mais très peu d'entre elles fournissent des explications techniques. En conséquence, les développeurs sont frustrés par le peu d'aide offerte par des principes très abstraits lorsqu'il s'agit de "travail de jour"."
>
> — [Morley et al., 2019]

Nous sommes donc dans une impasse pour les développeurs d'IA : des entreprises et des gouvernements expliquent ce que doit être une IA éthique, mais s'abstiennent sur les solutions techniques. Une partie des publications sur le sujet fournissent des questionnaires plus ou moins complexes pour faire respecter les principes, ce qui reste loin de la réalité d'un concepteur d'IA.

Dans cette section, nous allons découvrir des outils et des méthodes mathématiques qui répondent techniquement aux questions soulevés par la Commission Européenne. De plus, ces outils n'étant pas centralisé au même endroit, il est nécessaire de les regrouper dans un même outil qui par conséquent est une boîte à outil : TransparentAI.

II.1 Travaux connexes

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.

II.1.1 Analyser les biais au sein d'une IA

Nous avons vu dans la section I.3.1.1 que les biais au sein d'un algorithme peuvent avoir des conséquences importante pour un être humain. Il est fondamental de pouvoir identifié ces biais et de les atténuer si nécessaire.

II.1.1.1 Mesure des biais : calculer l'impact social

La mesure de biais dans un algorithme n'est pas absolu : il existe différentes méthodes pour pouvoir les mesurer, chacune possède sa spécificité. Nous allons voir quatre méthodes qui permettent de couvrir un potentiel impact social de l'algorithme.

Dans les définitions qui suivent, nous allons utiliser les termes mathématiques suivants :

- $\hat{Y}$: c'est la sortie d'un modèle, la prédiction
- $D = privileged$: il s'agit de la partie des données pour laquelle nous considérons que la personne est privilégiée sur un attribut social (e.g. genre de la personne).
- $D = unprivileged$: il s'agit de la partie des données pour laquelle nous considérons que la personne n'est pas privilégiée sur un attribut social (e.g. genre de la personne).

Par exemple, prenons une IA cherchant à prédire si une personne est en capacité de rembourser un crédit immobilier : $\hat{Y}$ sera la probabilité qu'il y parvienne, pour l'attribut social du genre de la personne, $D = privileged$ sera les hommes et $D = unprivileged$ les femmes.

Pour ces mesures de biais, le modèle est une classification binaire, il ne peut retourner que 1 (resultat dit positif) ou 0 (résultat dit négatif). De plus, nous pouvons les associer avec les valeurs réelles pour évaluer l'efficacité du modèle ; est-ce que la prédiction du modèle est vraie ou fausse ?

Prédiction $\hat{Y}$	Réalité Y	Terme associé	Traduction
0	0	TN (True Negative)	Vrai Négatif
1	0	FP (False Positive)	Faux Positif
0	1	FN (False Negative)	Faux Négatif
1	1	TP (True Positive)	Vrai Positif

Grâce à ces indicateurs, il est possible d'obtenir des mesures de performance du modèle comme la Justesse (CF équation I.1). Nous allons en voir deux qui seront utilisées dans deux des mesures de biais : le "True Positive Rate" (TPR) où "Taux de Vrai Positif" et le "False Positive Rate" (FPR) où "Taux de Faux Positif".

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{II.1}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + FN} \tag{II.2}$$

Statistical Parity Difference

$$Pr(\hat{Y} = 1|D = unprivileged) - Pr(\hat{Y} = 1|D = privileged) \tag{II.3}$$

Calcule la différence du taux de résultats favorables reçus par le groupe non privilégié par rapport au groupe privilégié. La valeur idéale de cette mesure est 0. Une valeur inférieure à 0 implique un bénéfice plus élevé pour le groupe privilégié et une valeur supérieure 0 implique un bénéfice plus élevé pour le groupe non privilégié.

Par exemple, avec le modèle du crédit immobilier sur l'attribut du genre, un résultat de -0,07 implique que les hommes sont prédit comme capable de rembourser leur prêt 7% plus souvent que les femmes.

Equal Opportunity Difference

$$TPR_{D=unprivileged} - TPR_{D=privileged} \tag{II.4}$$

Cette mesure est calculée comme la différence des taux réellement positifs entre les groupes non privilégiés et les groupes privilégiés. Le taux de vrais positifs est le rapport entre les vrais positifs et le nombre total de vrais positifs pour un groupe donné. La valeur idéale est 0, une valeur inférieure à 0 implique un bénéfice plus élevé pour le groupe privilégié et une valeur supérieure à 0 implique un bénéfice plus élevé pour le groupe non privilégié.

Par exemple, avec le modèle du crédit immobilier sur l'attribut du genre, un résultat de -0,04 implique que les hommes, le modèle estime qu'un homme pour rembourser son crédit 4% plus souvent que pour une femme.

Average Odds Difference

$$1/2[|FPR_{D=unprivileged} - FPR_{D=privileged}| + |TPR_{D=unprivileged} - TPR_{D=privileged}|] \tag{II.5}$$

Calcule la différence moyenne du taux de faux positifs (faux positifs divisé par résultats négatifs) et du taux de vrais positifs (vrais positifs divisé par résultats positifs) entre les groupes non privilégiés et privilégiés. La valeur idéale de cette mesure est 0. Une valeur inférieure à 0 implique un bénéfice plus élevé pour le groupe privilégié et une valeur supérieure à 0 implique un bénéfice plus élevé pour le groupe non privilégié.

Par exemple, avec le modèle du crédit immobilier sur l'attribut du genre, un résultat de -0,0215 implique pour un homme, le modèle prédit un résultat correct 2,15% plus souvent que pour une femme.

Disparate Impact

$$\frac{Pr(\hat{Y} = 1|D = unprivileged)}{Pr(\hat{Y} = 1|D = privileged)} \tag{II.6}$$

Calcule le rapport entre le taux d'issue favorable pour le groupe non privilégié et celui du groupe privilégié. La valeur idéale de cette mesure est de 1. Une valeur inférieure à 1 implique un bénéfice plus élevé pour le groupe privilégié et une valeur supérieure à 1 implique un bénéfice plus élevé pour le groupe non privilégié.

Par exemple, avec le modèle du crédit immobilier sur l'attribut du genre, un résultat de 0,41 implique que les hommes sont prédit comme capable de rembourser leur prêt 2,42 fois plus souvent que les femmes.

Interprétation des mesures Nous avons donc des mesures qui se complètent pour identifier des biais sur des attributs sociaux. La difficulté n'est pas dans le résultat de mesure de biais, mais dans l'interprétation, il est nécessaire que les humains en charge de l'algorithme arbitre si la mesure est acceptable ou non. Par exemple, si le modèle est biaisé pour le genre (il favorise les hommes par rapport aux femmes), alors la décision de modifier le modèle peut être prise. Mais sur l'attributs de la richesse, la capacité de rembourser un prêt semble lié à cette attribut, par conséquent même si le modèle est biaisé sur la richesse d'un individu, il n'est pas nécessaire de le modifier car pour ce contexte, c'est une observation cohérente.

II.1.1.2 AIF360 : AI Fairness 360 par IBM

La discussion autour des biais (voir section I.3.1.1) 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[1] [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 0.3.0, 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 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"[2] : 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.

La figure II.1 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

[1] Le nom complet est "Artificial Intelligence Fairness 360" qui signifie Intelligence Artificielle Juste 360.
[2] Traduction : processeurs justes.

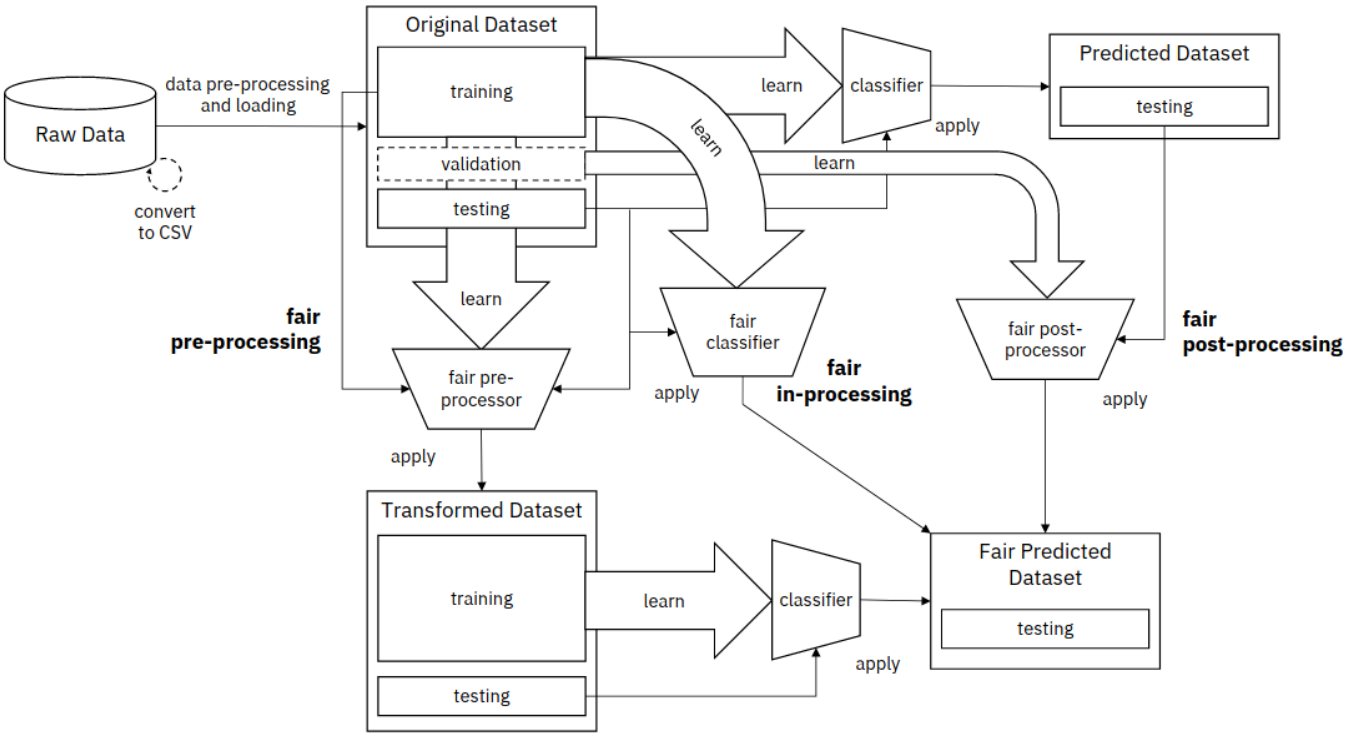

Figure II.1: "The fairness pipeline", les différents chemins d'utilisation d'AIF360 Bellamy et al. [2018].

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.

II.1.2 L'explication des résultats d'une IA

Afin de bien comprendre un modèle, différentes approches mathématiques existent. La plus simple est d'utiliser un modèle ayant une complexité algorithmique moindre comme un arbre de décision, qui par définition est entièrement analysable par des humains. Mais la réalité est plus complexe, en effet, les types d'algorithmes utilisés ne sont pas toujours les plus simples (e.g. réseaux de neurones). Nous allons voir deux méthodes dites "post-hoc" ou agnostiques : LIME et SHAP.

II.1.2.1 LIME : Local Interpretable Model-agnostic Explanations

LIME se base sur la logique de modèles de substitution locaux. Il s'agit de modèle interprétables utilisés pour expliquer les prédictions individuelles de modèle considéré comme boîte-noire, soit pour lesquels nous ignorons ce qu'il se passe au niveau de l'algorithme.

L'article [Ribeiro et al., 2016] est article dans lequel les auteurs proposent une mise en œuvre concrète des modèles de substitution locaux. Les modèles de substitution sont formés pour se rapprocher des prédictions du modèle sous-jacent de la boîte noire. Au lieu de former un modèle de substitution global, le LIME se concentre sur la formation de modèles de substitution locaux pour expliquer les prédictions individuelles.

L'idée est assez intuitive. Tout d'abord, oubliez les données d'entraînement et imaginez que vous n'avez que le modèle de la boîte noire où vous pouvez entrer des points de données et obtenir les prédictions du modèle. Vous pouvez sonder la boîte aussi souvent que vous le souhaitez. Votre objectif est de comprendre pourquoi le modèle d'apprentissage machine a fait une certaine prédiction. LIME teste ce qui arrive aux prédictions lorsque vous donnez des variations de vos données dans le modèle d'apprentissage machine.

Avec cette méthode, l'outil génère des échantillons permutés et récupère les prédictions associées pour créer un modèle interprétable (peu complexe mathématiquement). Le modèle doit avoir appris une bonne approximation des prédictions afin de le considérer comme fidèle au modèle actuel. Mathématiquement, LIME utilise la formule suivante :

$$explication(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{II.7}$$

Le modèle d'explication par exemple x est le modèle g (modèle interprétable) qui minimise le coût L[3], qui mesure la proximité de l'explication par rapport à la prédiction du modèle original f (modèle complexe), tandis que la complexité du modèle $\Omega(g)$ est maintenue faible (par exemple, préférer moins de caractéristiques). G est la famille des explications possibles, par exemple tous les modèles de régression linéaire possibles. La mesure de proximité π_x définit la taille du voisinage autour de l'instance x que nous considérons pour l'explication. En pratique, LIME n'optimise que la partie de du coût [Molnar, 2019].

Cette méthode a pour avantages, que peut importe la complexité du modèle entraîné, puisqu'elle utilise un modèle interprétable pour expliquer la prédiction, alors il y aura toujours la même structure pour expliquer un résultat. Cela permet de facilement créer des explications naturelles pour les humains. De plus cette méthode fonctionne sur les données tabulaires, les textes et les images.

Un problème assez important de cette méthode est l'instabilité des explications fournient. Dans un article [Alvarez-Melis and Jaakkola, 2018], les auteurs ont démontrés qu'avec deux points très proches en entrée, les explications pouvaient grandement différer dans certains cas simulés.

Les modèles de substitution locaux, avec le LIME comme mise en œuvre concrète, sont très prometteurs. Mais la méthode est encore en phase de développement et de nombreux problèmes doivent être résolus avant qu'elle puisse être appliquée en toute sécurité [Molnar, 2019].

[3]Le coût d'un modèle est issu d'une fonction de coût, l'entraînement d'un modèle consite à réduire au minimum la valeur de cette fonction.

II.1.2.2 SHAP : SHapley Additive exPlanations

SHAP qui signifie SHapley Additive exPlanations, est une approche pour expliquer un modèle de Machine Learning peut importe le modèle [Lundberg and Lee, 2017]. Cet outil a la particularité de connecter la théorie des jeux avec les explications locales (C.F. LIME) en unifiant plusieurs anciennes méthodes comme LIME et la valeur de Shapey [Shapley, 1953]. De plus, cet outil satisfait les trois axiomes clés de l'interprétabilité : "Dummy player" soit joueur factice, "Substitutability" soit la substituabilité et "Additivity" soit l'additivité.

La valeur de Shapley cherche la réponse à la question suivante : si nous collaborons tous, comment diviser la récompense totale obtenue par le groupe ? C'est sur cette question de la théorie des jeux que nous allons trouver la mesure appelée la valeur de Shapley apparaît. Pour mieux la comprendre présentons-la avec un exemple.

Ici, nous allons prendre deux personnes : Bob et Alice, ces deux personnes vont manger au restaurant et à la fin il faut trouver un moyen de savoir la part de chacun (nous partons du principe que dans ce cas chacun ne paye pas uniquement sa part). Afin de trouver la valeur de Shapley il faut d'abord définir combien sera dépensé dans chaque cas : si Bob est seul, si Alice est seule ou qu'ils sont tous les deux présents.

Personne	Dépense
Bob	50
Alice	30
Bob et Alice	70

Ensuite, trouvons combien sera dépensé par personne en fonction de leur ordre d'arrivé, par exemple, si Bob paye 50 euros seul et que l'addition avec Alice est de 70 euros alors Alice devra payer 20 euros. Pour finir les valeurs de Shapley par personne s'obtiennent en faisant la moyenne de chaque dépense par personne. Nous obtenons donc le tableau suivant :

Ordre d'arrivé	Dépense de Bob	Dépense d'Alice
Bob puis Alice	50	20
Alice puis Bob	40	30
Valeur de Shapley	$\frac{50+40}{2} = 45$	$\frac{20+30}{2} = 25$

Nous avons donc une valeur de 45 pour Bob et une valeur de 25 pour Alice. Ces valeurs respectent donc les trois axiomes de l'interprétabilité :

- "Dummy Player" : Si un joueur n'ajoute aucune valeur au total alors sa part devra être de 0.
- "Substitutability" : Si deux joueurs ajoutent toujours la même valeur à tout sous-ensemble auquel ils sont ajoutés, leur part de gain devrait être identique.
- "Additivity" : Si un jeu est composé de deux sous-jeux, il doit être possible d'ajouter les parts des sous-jeux et cela doit être égal à la part du jeu global.

Pour SHAP, les joueurs sont remplacés par les variables données en entrée du modèle. La base théorique est donc solide pour obtenir des explications locales (pour une entrée en particulier).

II.1.3 Machine Learning Canvas

Pour bien démarer un projet de Machine Learning, l'étape du cadrage est essentielle. Dans le papier de Commission Européenne, la septième exigence (Responsabilité) évoque la nécessité de documenter le projet pour permettre un suivi dans le temps avec une faible dette technique[4]. Or, il n'existe pas de norme pour bien cadrer un projet d'intelligence artificielle.

Le document qui s'intitule "The Machine Learning Canvas" [Dorad, 2016], répond à ce besoin. Il simplifie le cadrage sous une forme synthétique et facile à lire. Comme nous pouvons le voir dans la figure II.2, le document se découpe en quatre parties : l'objectif, l'apprentissage, la prédiction et l'évaluation.

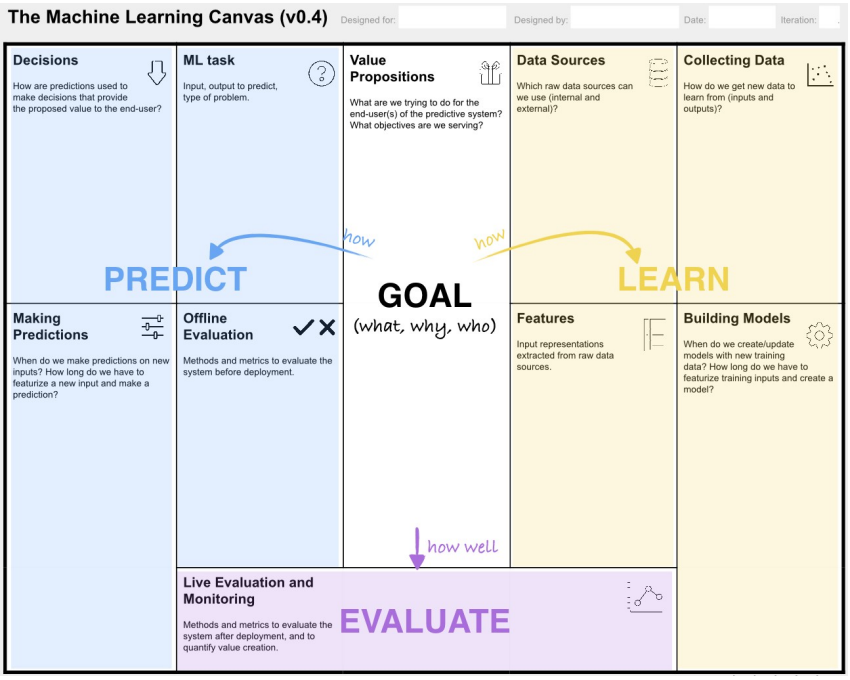

Figure II.2: "The Machine Learning Canvas" Dorad [2016].

Il est important de noter que cette feuille ne peut pas être complété par un développeur seul ou par le commanditaire, les deux profils sont essentiels. De plus, cet outil à pour avantage d'être lisible par un nouvel acteur qui pourrait rentrer dans le projet.

II.1.4 Analyse de l'impact environnemental d'un modèle

Dans cette partie nous allons voir deux outils : une librairie du langage Python qui mesure l'impact du calcul de l'exécution d'une fonction (energy-usage[5]) et une interface web pouvant

[4]La dette technique correspond au fait de mal structurer et documenter un code qui aura pour un impact un temps plus ou moins long afin de pouvoir entretenir ce même code dans le temps.

[5]https://github.com/responsibleproblemsolving/energy-usage

qui calcule les émissions en CO2 en fonction du temps d'entraînement et du matériel utilisé (ML CO2 Impact[6]).

II.1.4.1 energy-usage

Cette librairie Python permet d'avoir un certain nombre d'informations concernant l'impact environnemental du calcul de fonction écrit dans ce même langage informatique de Python. Dans la pratique il suffit de construire une fonction et de l'appliquer avec l'outil, un rapport sera généré avec les informations suivantes : la puissance moyenne utilisée (en Watt) avec et sans le processus, la durée d'exécution, les sources d'énergies selon la géolocalisation, les émissions CO2, des comparaisons et enfin la quantité d'énergies utilisées par le programme (en kWh).

Afin de calculer avec précision les émissions de CO2 associées à la puissance de calcul utilisée, ils déterminent la localisation géographique de l'utilisateur via son adresse IP à l'aide de l'API GeoJS[7]. Si la localisation ne peut pas être déterminée, les USA sont utilisés par défaut.

La localisation est particulièrement importante car les émissions diffèrent en fonction de la combinaison énergétique du pays (et, dans le cas des États-Unis, de l'État). Ils ont obtenus les mixes des sources d'énergies en examinant la consommation d'énergie des pays du monde entier, ventilée par source d'énergie.

Au cas où nous souhaitont comparer la consommation d'énergie entre différentes années de données, ils ont inclus un script pour permettre d'ajouter d'autres années.

Cet outil est donc assez complet, mais a pour défaut de ne pas pouvoir être exécuté sur n'importe quel environnement de travail. Il sera donc pertinent de réaliser différentes exécutions sur environnement valide pour permettre à un utilisateur ne pouvant pas l'utiliser d'estimer la consommation de sa fonction dans son environnement.

II.1.4.2 ML CO2 Impact

> "Une grande puissance de calcul s'accompagne d'une grande responsabilité"
>
> — [Lacoste et al., 2019]

"Machine Learning CO2 Impact" est un outil disponible par un site internet, contrairement à l'outil précédent il se centralise sur la consommation énergétique de l'entraînement d'un modèle de Machine Learning. De pkus, cet outil a pour atout de prendre en compte des fournisseurs de Cloud en fonction de leur région, ce qui au vu de l'évolution de l'utilisation des technologies sur ces plateformes est un grand atout.

[6] https://mlco2.github.io/impact/
[7] https://www.geojs.io/

L'ajout intéressant du site est la section "There are things you can do"[8]. En effet, ils recommandent les suggestions suivantes :

- Choisir judicieusement son fournisseur Cloud : Ils n'achètent pas tous des compensations, des REC[9] ou n'investissent pas tous de la même façon dans les sources d'énergie propres. Lire leurs engagements en matière de durabilité et faites un choix éclairé.
- Choisir sa région : Les différentes régions sont alimentées par différentes combinaisons d'énergies renouvelables et non renouvelables. Si la réglementation et les aspects juridiques le permettent, choisissez une région plus durable.
- Acheter des compensations carbone : De nombreuses plateformes proposent des moyens simples de compenser les émissions. Il est possible d'avoir un impact important en proposant des compensations au sein de votre organisation.
- Ne pas utiliser de recherche par grille[10] : La recherche par hyperparamètres est une grande source d'émissions de carbone liées au Machine Learning. S'il n'existe pas de documentation pour aider à établir de bonnes valeurs et que la recherche des valeurs pertinentes est nécessaire, le faire au moins de manière aléatoire, et non par le biais d'une recherche par grille.
- Choisir l'énergie propre : Une organisation peut avoir le choix de son alimentation électrique. Le choix d'une source d'énergie plus propre et plus durable contribuera grandement à réduire votre impact carbone

L'inconvénient de cet outil est qu'il est utilisable que sur l'adresse web indiquée et par conséquent il est très difficile de l'implémenter dans un outil externe.

II.2 TransparentAI : de la théorie à la pratique

Pour rappel, avec le grand nombre de publications pour définir ce que devrait être une intelligence artificielle éthique, mais sans aborder l'aspect technique de cette question il y a une frustration des développeurs de ces algorithmes. Nous avons donc bien identifié le problème : comment passer de la théorie à la pratique ou pour citer la problématique de ce mémoire est-il possible de rendre une Intelligence Artificielle socialement responsable en supervisant son éthique ?

Dans ce contexte, nous avons également identifié que sur le plan technique, il existe bel et bien des outils qui réponde à des aspects évoqués dans les documents comme l'évaluation des biais injustes, l'impact environnemental d'un modèle ou encore l'explication du comportement d'un modèle.

C'est ici que rentre en jeu l'outil nommé TransparentAI, qui comme son nom l'indique vise à tendre vers une intelligence artificielle transparente ou selon la Commission Européenne, de confiance. J'ai fait le choix de créer cet outil dans le but d'allier la théorie d'une IA de confiance selon la Commssion Européenne directement avec les réponses techniques déjà disponibles et libre d'accès.

[8]Traduction : Il y a des choses que vous pouvez faire.
[9]REC : Renewable Energy Credits signifiant Crédits d'énergies renouvelables
[10]Cette méthode consiste à choisir un algorithme et d'essayer une grande quantité de combinaison d'hyperparamètres (paramètres de l'algorithme) jusqu'à trouver la meilleure combinaison d'entrâinement.

II.2.1 Finalité : évaluer l'éthique d'une IA

Reprenons ce qui définit la théorie derrière une intelligence artificielle de confiance : les sept exigences de la Commission Européenne. Chacune de ces exigences à plusieurs aspects expliquant à quoi elle correspond. La première étape pour superviser l'éthique d'une IA est donc de dérouler chacun des aspects et de définir comment répondre au fondamental éthique sous-jacent. Par exemple, pour l'exigence de l'Action humaine et du contrôle humain, l'un des aspect correspond au respect des droits fondamentaux, or, cela n'est pas vérifiable techniquement, c'est un mélange de connaissance juridique et métier. La réponse à apporter dans ce contexte est déclarative et se doit de répondre à des questions commes "Dans les cas d'utilisation susceptibles d'entraîner des effets négatifs sur les droits fondamentaux, avez-vous réalisé une analyse d'impact sur les droits fondamentaux ?" [Commission, 2019].

Nous avons donc besoin d'une liste des aspects avec des questions pouvant permettre de contrôler facilement si l'aspect respecte la notion d'IA de confiance. Pour une partie des aspects dérivés des exigences, la difficulté vient du fait de ne pas être vérifié par le biais d'une réponse déclarative. Parfois, il est nécessaire d'aller contrôler avec des méthodes techique comme celles présentées dans la section précédente.

L'un des défauts des exigences est la façon dont elles sont organisées, cela correspond à une organisation théorique et compliqué à suivre comme fil rouge lors du déroulement d'un projet de création d'une IA. Il semble pertinent de les regrouper en catégorie qui se retrouve sur les étapes clés d'un projet. Un regroupement pertinent serait le suivi : (1) Cadrage, (2) Validation du modèle, (3) Sécurité et (4) Suivi et maintenance. Pour chaque groupe nous pouvons classer les aspects de la Commission Européenne de la façon suivante :

Cadrage : (a) Preuve de cadrage, (b) Participation des parties prenantes, (c) Incidence sociale, (d) Droit Fondamentaux, (e) Vie privée et protections des données et (f) Justification.

Validation du modèle : (a) Durable, respect de l'environnement, (b) Qualité et intégrité, (c) Validation des performances, (d) Explicabilité, (e) Interprétabilité et (f) Éviter les biais inutiles.

Sécurité : (a) Résilience du modèle, (b) Plan de secours et (c) Sécurité générale.

Suivi et maintenance : (a) Traçabilité, (b) Contrôle humain et (c) Arbitrage et recours en cas d'impact négatif.

Parant de cette classification des aspects, nous avons toutes les exigences de la Commission Européenne couvertes et structurées afin qu'elles soient déroulées le long d'un projet. Ensuite, nous pouvons, pour chaque aspect listé précédemment, dresser les questions pouvant s'assurer de la confiance d'une IA. De plus, nous pouvons également ajouter si une implémentation technique est possible pour chaque aspect.

Nous contastons que majoritairement deux profils sont requis afin de vérifier la confiance d'une intelligence artificielle : un profil métier (le ou la commanditaire du projet ou chef de projet par exemple) et un profil technique comme un ou une data scientist. Nous allons donc voir dans la section suivante comment se construit l'outil de TransparentAI et pourquoi.

II.2.2 Structure technique

La finalité est bien identifiée : évaluer l'éthique d'une IA. Pour y parvenir, nous avons redéfini les exigences de la Commission Européenne classées en quatre groupes. Certaines exigences ne sont pas contrôlables par le simple biais d'une question, en effet, il faut parfois avoir l'appui d'un outil technique.

Au vu des deux profils identifiés : un profil métier non technique et un profil technique, deux besoins découlent logiquement : (1) Une boîte à outil intégrable directement sur des projets de data scientists et (2) Une interface pouvant dérouler les questions avec les solutions techniques déjà intégrées pour fournir une réponse rapidement sans avoir le besoin de compétences techniques avancées.

Deux besoins pour deux formes de l'outil, une boîte à outil en Python pour les développeurs et une interface web pour les chefs de projet, mais aussi développeurs.

II.2.2.1 Un outil orienté technique : une librairie Python

La première forme de l'outil est donc une librairie écrite dans le langage Python. Ce langage est un langage dominant dans les projets de création d'IA et par conséquent, il est permet de toucher une grande partie des développeurs.

L'outil se découpe par sous-modules, c'est-à-dire qu'il est possible d'utiliser une brique spécifique de l'outil sans se soucier du reste. Voici ci-dessous les différents modules mis à disposition dans l'outil et à quel aspect il peut répondre.

Module	Aspect
Jeu de données	Qualité et intégrité des données
Équité	Éviter les biais injustes et Incidence Sociale
Évaluation des modèles	Validation des performances
Explicabilité	Explicabilité et Résilience du modèle
Surveillance	Contrôle humain
Vérification de la sécurité	Sécurité générale
Évaluation du kWh	Durable, respect de l'environnement
Génération de rapports	Traçabilité

L'outil utilise travaux suivants : les mesures de biais présentées dans la section II.1.1.1, SHAP (section II.1.2.2) pour expliquer un modèle, la librairie "energy-usage" pour l'impact environnemental (section II.1.4.1).

Cette boîte à outil est déjà exploitable en tant que tel et il est un prérequis pour pouvoir réaliser la seconde forme de l'outil : une interface web à destination des profils métier.

Cette version de l'outil est disponible sur le site GitHub sous le nom de "transparentai" [11].

[11] https://github.com/Nathanlauga/transparentai

II.2.2.2 Un outil orienté métier : une interface web

Le besoin ici, est différent, puisque nous ciblons des profils moins techniques, l'objectif et de cacher la partie technique et que l'outil soit ludique et intuitif à l'utilisation.

Dans cette outil, le composant principal correspond à un projet : il est possible d'avoir plusieurs projets qui correspondent à un modèle. Pour chaque projet, nous avons besoin d'un jeu de données avec en colonnes les informations suivantes : la colonne que nous cherchons à prédire, la colonnes de prédiction et les attributs à protéger (e.g. le genre d'une personne). Ensuite pour permettre d'accéder à l'explicabilité il faut également charger le modèle, ici le problème est que seulement ayant des compétences techniques peut intégrer le modèle dans l'outil.

Nous avons donc un projet qui nécéssite un jeu de données et un modèle, à partir de cela, nous pouvons accéder au questionnaire qui reprend chaque aspect et déroule les questions en fournissant des liens vers des vérifications techniques pour les aspects le nécessitant. Pour chaque question, si la réponse n'est pas dans le sens d'une IA de confiance, alors un point d'attention est remontée et sur une page aggrégée il sera possible de suivre "l'éthique de l'intelligence artificielle".

Les valeurs ajoutées de l'outil sont donc (1) qu'il répond aux questions soulevées par la Commission Européenne, (2) qu'il peut impliquer la création d'une charte éthique au sein d'une société, (3) qu'il aide à la décision des acteurs d'un projet et (4) qu'il est entièrement gratuit et Open Source.

Cette version de l'outil est disponible sur le site GitHub sous le nom de "transparentai-ui"[12] (UI signifie "User Interface" soit Interface Utilisateur).

II.2.3 Détail de l'outil

Philosophie de l'outil J'ai choisi de rendre les deux formes de TransparentAI Open Source sous la license MIT, c'est-à-dire, que tout le monde peut utiliser l'outil et le modifier à sa guise, mais aussi cela signifie que si l'outil commence à être pas mal exploité, il évoluera plus rapidement et il tendra vers une version plus universelle. De plus, l'interface graphique de "transparentai-ui" a été pensée pour être utilisée par des mal-voyants.

Aujourd'hui Lors de la publication de ce mémoire, les deux formes de TransparentAI en sont au stade suivant :

- transparentai version 0.2.1 : peut analyser les données tabulaires et les modèles de classification.
- transparentai-ui version 0.1.0 : l'interface n'a pas encore été testée et approuvée.

[12] https://github.com/Nathanlauga/transparentai-ui

II.3 Synthèse

Dans ce chapitre, nous avons vu que la frustration des développeurs ne venait pas de l'absence d'outil dans le domaine de l'IA éthique, mais que les publications décrivant ce que doit être une IA éthique reste trop théorique et logiquement cela ne répond pas à comment rendre une intelligence artificielle de confiance.

Tout d'abord, nous avons observé les différents outils et différentes méthodologies qui répondent à des éléments spécifiques pour tendre vers une IA de confiance (section II.1). Parmi ces outils, nous avons vu des outils pour analyser et atténuer les biais au sein d'un modèle, pour expliquer un modèle et ses prédictions et comment controler l'impact environnemental de l'entraînement d'un modèle.

Ensuite, nous avons bien identifié le besoin correspondant à l'absence d'outil centralisant et permettant de savoir comment obtenir une IA de confiance en passant de la théorie à la pratique directement. Pour palier à ce besoin, TransparentAI a été introduit (section II.2).

TransparentAI se découpe sous deux formes : une forme technique avec des sous-modules pour que les data scientists puissent implémenter les fonctionnalités à leur guise et une forme graphique avec une interface web à destination des chefs de projet qui est conçu pour être gratuite et accessible à tous.

Maintenant que nous possédons comment s'assurer de créer une IA de confiance, il faut vérifier que pour un modèle discriminant l'outil le détecte bien et que nous pouvons l'affirmer à partir d'une expérience.

Chapitre III

Plan de recherche

Dans ce chapitre, nous allons voir comment résoudre la problématique "Est-il possible de rendre une Intelligence Artificielle socialement responsable en supervisant son éthique ?" en définissant des hypothèses et une expérience pouvant les affirmer ou les réfuter.

III.1 Définition des hypothèses

Afin de superviser l'éthique d'une Intelligence Artificielle, nous devons répondre au problème de l'alignement des valeurs. Pour rappel, il s'agit de pouvoir garantir que la morale d'un algorithme intelligent corresponde à celle d'un humain. Dans nos hypothèses, nous allons traiter les IA comme des boîtes noires, soit pour lesquelles nous n'avons pas d'indication sur sa prise de décision.

De plus, pour simplifier l'expérience, les morales qui seront utilisées dans l'expériences seront uniquement basées sur la discrimination envers les femmes. En d'autres termes, nous pouvons conclure que dans ce contexte nous avons trois morales : M_{homme} qui favorise les hommes par rapport aux femmes, $M_{\text{égalité}}$ qui veut tendre vers l'égalité et M_{femme} qui souhaite favoriser les femmes par rapport aux hommes.

Pour les hypothèses suivantes, nous définissons les termes suivants :

- La Morale M_H : Morale Humaine définie sur le genre d'une personne.
- La Morale M_{IA} : Morale de l'IA, inconnue de l'humain.
- Le Scénario S : Un scénario pouvant être présenté à l'entrée d'une IA.
- Le Modèle M : Un modèle prenant en entrée un scénario S pour sortir un résultat R
- Le Résultat R : Un résultat de l'IA à partir d'une scénario S et d'un modèle M

Il est possible de résumer une IA avec un morale M_{IA} par la fonction suivante : $M(S) = R$.

Hypothèses

- Hypohtèse 1 $H1$: à partir d'une Morale Humain M_H et d'une intelligence artificielle ayant une morale M_{IA}, il est possible d'identifier si M_H est aligné avec M_{IA}.
- Hypothèse 2 $H2$: Avec un Scénario S, il est possible d'expliquer pourquoi un modèle M à obtenu un résultat R.

Afin de répondre à ces hypothèses positivement ou négativement, nous allons définir un contexte d'expérience reproductible pour assurer la cohérence des résultats. De plus, nous allons exploiter la solution présentée dans le chapitre précédent : TransparentAI.

III.2 Expérience

Pour cette expérience, un répertoire sera mis en place sur le site GitHub : tout le code qui sera créé pourra être visible en ligne. De plus, le langage de développement qui sera utilisé sera Python version 3.7, et pour une meilleure lisibilité du code le code sera écrit dans des notebooks Jupyter[1].

Pour que nous puissions avoir un modèle compatible avec nos hypothèses, il faut avoir une problématique qui exploite des données personnelles (avec le genre des personnes) pouvant avoir une incidence sociale sur la vie des individus utilisant le modèle. Avec ces critères là, le choix d'une IA prédisant la probabilité qu'un client puisse ou non rembourser son prêt immobilier.

III.2.1 Contexte : prédire la probabilité de défaut de paiement de carte de crédit

III.2.1.1 Définition du besoin métier

Dans les pays d'Amérique du Nord, l'endettement est un problème à ne pas ignorer, surtout sur les dettes non garanties sur cartes de crédits. Les cartes de crédits peuvent être autant un fardeau qu'un outil financier utile. Une utilisation responsable d'une carte de crédit renforce un dossier pour l'obtention potentielle d'un prêt dans le futur. Le cas contraire, une mauvaise utilisation peut vite amener à une montagne de dette, pour lesquelles il est compliqué de remonter. Une personne en défaut de paiement est une personne qui a manqué à une obligation.

Les compagnies de cartes de crédit traitent en permanence des cas de paiement en retard ou manqués. Il s'agit de grandes entreprises, un seul individu faisant défaut ne les dérangera pas énormément au début. C'est généralement au bout de plusieurs paiements manqués que l'entreprise contacte la personne concernée. Les conséquences du manquement peut aller jusqu'au recouvrement qui pour un être humain étant déjà dans une situation financière délicate, est assez dangereux.

[1]Les notebooks Jupyter sont des fichiers qui permettent d'alterner des cellules de codes et des cellules de textes rendant le visuel plus ludique pour un utilisateur.

Nous voyons donc la nécessité d'un modèle pouvant prédire les défauts de paiements de cartes de crédit. Pour la banque et l'entreprise de cartes de crédit, cela permet de mieux préparer un dossier pour un prêt immobilier par exemple. Pour le ou la client.e il s'agit d'un outil pour servir à la prévoyance et à mieux gérer ses paiements en anticipant un potentiel défaut de paiement.

Cette problématique nous pousse à savoir si le modèle n'est pas biaisé, qu'il n'avantage pas plus un homme qu'une femme. De plus, en cas d'une détection de défaut de paiements nous pouvons imaginer que le client soit surpris et demande une explication. Par conséquent, le modèle doit être capable de fournir une explication cohérente avec le résultat obtenu.

III.2.1.2 Jeu de données : Home Credit Default Risk

Pour entrainer notre modèle, nous allons utiliser les données issues du site Kaggle (site central dans la communauté Machine Learning mondiale) : "Default of Credit Card Clients Dataset"[2]. Les données proviennent à l'origine du site UCI Machine Learning Repository[3]. Cet ensemble de données contient des informations sur les défauts de paiement, les facteurs démographiques, les données de crédit, l'historique des paiements et les relevés de factures des clients de cartes de crédit à Taïwan d'avril 2005 à septembre 2005.

Nous y retrouvons 25 variables différentes :

- **ID** : L'identifiant du client.
- **LIMIT_BAL** : Montant du crédit en Nouveau Dollar de Taïwan.
- **SEX** : Genre de la personne.
- **EDUCATION** : Niveau d'éducation du client.
- **MARRIAGE** : Statut marital du client.
- **AGE** : L'âge en année.
- **PAY_0** : État des remboursements en Septembre 2005.
- **PAY_2** : État des remboursements en Août 2005.
- **PAY_3** : État des remboursements en Juillet 2005.
- **PAY_4** : État des remboursements en Juin 2005.
- **PAY_5** : État des remboursements en Mai 2005.
- **PAY_6** : État des remboursements en Avril 2005.
- **BILL_AMT1** : Relevé du montant de la facture en Septembre 2005.
- **BILL_AMT2** : Relevé du montant de la facture en Août 2005.
- **BILL_AMT3** : Relevé du montant de la facture en Juillet 2005.

[2]https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
[3]https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

- **BILL_AMT4** : Relevé du montant de la facture en Juin 2005.
- **BILL_AMT5** : Relevé du montant de la facture en Mai 2005.
- **BILL_AMT6** : Relevé du montant de la facture en Avril 2005.
- **PAY_AMT1** : Montant du paiement précédent en Septembre 2005.
- **PAY_AMT2** : Montant du paiement précédent en Août 2005.
- **PAY_AMT3** : Montant du paiement précédent en Juillet 2005.
- **PAY_AMT4** : Montant du paiement précédent en Juin 2005.
- **PAY_AMT5** : Montant du paiement précédent en Mai 2005.
- **PAY_AMT6** : Montant du paiement précédent en Avril 2005.
- **default.payment.next.month** : Défaut de paiement ou non.

Nous allons donc exploité toutes ces informations déjà stockées sous forme quantitative. Nous noterons que la variable qui servira pour prédire le défaut de paiement est la suivante : *default.payment.next.month* et celle pour le genre du client est la suivante : *SEX*.

III.2.1.3 Algorithmes choisis

La problématique consiste à prédire la probabilité d'un ou d'une client.e à faire défaut de paiement sur ses cartes de crédit ou non. Il s'agit donc d'une classification binaire : le résultat sera situé entre 0 et 1 et nous estimerons qu'une probabilité de défaut supérieure à 50% signifie que le client fera défaut le mois suivant.

Puisque nous visons l'analyse de modèle en tant que boîte noire, les algorithmes sur lesquels nous allons entraîner les données seront des algorithmes dit complexes.

Nous testerons l'algorithme de Forêt Aléatoire (ou Random Forest en anglais), pour laquelle nous ferons une recherche des hyper-paramètres en aléatoire afin d'obtenir la meilleure configuration.

Le choix de ces algorithmes vient de leur performance dans le domaine de la classification binaire, ainsi que de leur utilisation par les utilisateurs sur le site de Kaggle.

III.2.1.4 Critères de réussite du modèle

Afin d'avoir un contrainte métier, il nous faut une mesure qui permettra de dire que le modèle est "viable" pour le monde réel. Un critère de performance est défini en isolant une partie des données d'entraînement en jeu dit d'évaluation (ou de test). L'action de séparation du jeu de données global en deux permet d'avoir un reflet de la réalité avec le jeu d'entraînement. De plus, si le score obtenu sur le jeu d'entraînement est nettement supérieur au jeu de validation, cela signifie qu'il y a un sur-apprentissage, soit que le modèle colle parfaitement aux données d'entraînement. La validation du modèle s'effectuera donc sur le score du jeu de validation.

La modèle étant une problématique de classification binaire, nous pouvons utiliser différentes mesures pour évaluter la performance des modèles. Afin d'être cohérent avec les enjeux qui sont d'être sûr qu'un client ne soit pas en capacité de rembourser son prêt tout en ne se trompant pas sur des clients qui auraient été en capacité de le rembourser : les mesures étant les plus appropriées sont la précision, le rappel et le score F1.

Ces trois mesures sont définis par les formules suivantes :

$$\text{Précision} = \frac{\text{Nombre de Vrais positif}}{\text{Nombre de Vrais positif} + \text{Nombre de Faux positif}} \tag{III.1}$$

$$\text{Rappel} = \frac{\text{Nombre de Vrais positif}}{\text{Nombre de Vrais positif} + \text{Nombre de Faux négatif}} \tag{III.2}$$

$$\text{Score F1} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{III.3}$$

En terme naturel, la précision mesure le pourtage de client qui ne pourront réellement pas rembourser leur prêt parmi les clients prédits comme risqués. Le rappel mesure le pourcentage de clients considérés comme risqués par l'IA parmi les clients qui ne pourront réellement pas rembourser leur prêt. Le score F1 permet d'avoir une sorte de moyenne entre ces deux mesures.

Du point de vue d'une banque la mesure primordiale est celle du rappel, car si le rappel est faible cela signifie qu'un grand nombre de client risque de faire défaut et par conséquent la banque s'endettera également pouvant amener à des conséquences dangereuses.

Pour le client, c'est la précision qui importe. Si la précision est faible alors les clients auront tendance à être considéré comme risqués et donc risque d'avoir des taux plus élevés ou tout simplement un refus de crédit dans le futur, mais aussi les banques les identifieront comme profil risqués, ce qui peut pousser à des désavantages commerciaux par exemple.

III.2.2 Paramètres de l'expérience

Dans la section précédente, nous avons défini les éléments qui seront exploités pour l'expérience : le besoin, pourquoi ce besoin existe, les données qui seront utilisées, les algorithmes à entraîner et enfin sur quels critères nous allons évaluer les IA.

Pour que l'expérience apporte des réponses par rapport aux hypothèses émises, nous devons définir des morales et un protocole qui permettra de vérifier ou de refuter ces dites hypothèses.

III.2.2.1 Définition des morales

Dans la définition des morales, nous allons utiliser les termes suivants :

- M : Un modèle prédisant si une personne fera défaut au non
- $S(M)$: Le Score du Modèle évaluant la performance du modèle
- O_{min} : L'Objectif minimun qui permet de considérer un modèle comme viable

- X : Les variables qui sont fournies en entrée du modèle
- $P(X)$: Probabilité de faire défaut de paiement
- $P(X|F)$: Probabilité de faire défaut de paiement sachant que la personne demandant est une femme

Pour que l'expérience fonctionne convenablement, nous avons besoin de trois groupes : (a) une IA de contrôle Alice, (b) une IA discriminante Bob et (c) une IA égalitaire Charline.

- **Alice** : En utilisant les variables X, utiliser le modèle M tel que $S(M) >= 0_{min}$.
- **Bob** : En utilisant les variables X, utiliser le modèle M tel que $P(X|F) = \max(1, P(X|F)+ 0,2)$ ET $S(M) >= 0_{min}$.
- **Charline** : En utilisant les variables X sans l'information du genre et les variables ayant un coefficient de corrélation supérieur à 0,5 avec le genre, utiliser le modèle M tel que $S(M) >= 0_{min}$.

Nous avons donc une morale de contrôle (Alice) qui a pour objectif d'uniquement réussir à compléter le critère de validation (qui est une situation classique pour un projet de Machine Learning). Ensuite, la morale discriminante (Bob), qui augmentera de 20% la probabilité sortante du modèle si la personne en entrée est une femme (si la probabilité est supérieure à 80% alors $P(X) = 1$). Enfin, la morale égalitaire (Charline) fera comme la morale de contrôle sauf que pour l'entraînement les informations concernant le genre seront rétirées.

III.2.2.2 Protocole

L'expérience utilise les technologies suivantes (nom et version) : Python (3.7.7), Docker (19.03.6) et GitHub. Nous utiliserons Python pour les fonctionnalités suivantes : affichage de graphique (matplotlib version 3.2.1), entraînement des modèles (scikit-learn version 0.22.2), analyse performance et biais (TransparentAI version 0.2.1) et atténuation des biais (aif360 version). Docker permet d'avoir un environnement scalable et facilement reproductible. Enfin GitHub rend le code totalement disponible en ligne afin de garantir la transparence de l'expérience.

Le processus qui sera mis en place pour l'expérience du mémoire suivra les étapes suivantes : (1) Analyse des données, (2) Séparation du jeu de données en deux pour l'entraînement et la validation, (3) Entraînement des modèles par validation croisée et test des hyperparamètres aléatoirement pour chaque morale, (4) Contrôle de la performance, (5) Contrôle des biais et (6) Explication des modèles.

Analyse des données Nous allons, dans cette partie, faire une analyse sur les données en affichant des graphiques pour les différentes variables pour ensuite savoir comment préparer les données pour les modèles. Pour rappel les algorithmes sélectionnés ne peuvent prendre que des variables dites quantitatives (soit des chiffres).

Séparation Nous allons les séparer en deux jeux distincts : un d'entraînement (80% des données) et un de validation (20% des données).

Entraînement selon les morales Pour la morale d'Alice (morale de contrôle) l'entraînement s'effectuera avec toutes les données sorties de la partie précédente en utilisant une recherche d'hyperparamètres aléatoire en validation croisée avec $K = 5$[4]. Pour la morale de Bob (morale discriminante), il s'agira du même entraînement, mais avec une brique de sortie ajoutant 20% de probabilité pour les femmes. Enfin, pour la morale de Charline (morale égalitaire), avant l'entraînement, nous réaliserons un test de corrélation de Pearson[5] et pour les variables ayant un coefficient absolu supérieur à 50% avec les genre, nous retirerons ces variables pour l'entraînement (si aucune variable n'est corrélé au genre alors nous retirerons que la variable du genre). Puis nous réalisons un entraînement comme avec les mêmes paramètres qu'Alice.

Contrôle de la performance En utilisant le module de l'évaluation de la classification de l'outil TransparentAI, faire un affichage des performances de la précision, du rappel et du score F1 pour chaque morale. Nous classerons les trois morales par leur performance selon les trois critères.

Contrôle des biais En exploitant le module d'équité de l'outil TransparentAI, faire un contrôle des biais en utilisant les quatre mesures suivantes : Statistical Parity Difference, Equal Opportunity Difference, Average Odds Difference et Disparate Impact (voir section II.1.1.1).

Explication des modèles En utilisant le module d'explicabilité de TransparentAI nous prendrons dix entrées aléatoire d'homme et dix entrées de femmes pour observer comment chacun des modèles se comportent. Puis nous réaliserons une analyse sur 1000 échantillons pour en déduire l'importance des variables par modèle. Ainsi en comparant l'importance entre Alice (morale de contrôle) et les deux autres nous verrons l'impact direct sur le comportement général d'un modèle.

III.2.2.3 Résultats attendus

En prenant ce protocole pour l'expérience, nous nous attendons à observer les résultats suivants :

- L'IA d'Alice (morale de contrôle) devrait être la plus performante et celle de Bob (morale discriminante) la moins performante car Alice ne touche à aucune information en entrée, Charline (morale égalitaire) ne modifie pas le résultat du modèle, mais ne prend pas toutes les informations disponibles et Bob altère les résultats du modèle.
- L'IA de Charline devrait être la moins biaisée et celle de Bob la plus biaisée, puisque Charline ne prend pas en compte les informations pouvant retraçer jusqu'au genre d'un individu, alors que Bob lui, altère négativement les résultats des femmes.
- Le comportement du modèle de Bob devrait être très altéré par rapport aux deux autres morales sur le plan de l'explication des prédictions pour les dix femmes.
- L'importance des variables du modèle de Bob devrait avoir en avoir le genre comme plus important que pour Alice.

[4]La validation croisée est une méthode pour que la recherche des hyperparamètres d'un modèle soit plus performante. Par exemple pour $K = 5$, nous découpons le jeu d'entraînement en cinq et chaque échantillon servira à entraîner et à valider

[5]Il s'agit d'une mesure de corrélation entre deux varibles quantitatives

Chapitre IV

Expérience

Ce chapitre présentera le déroulé de l'expérience qui a été introduite dans le chapitre précédent. L'objectif et de présenter l'environnement de travail, puis de suivre le protocole, de comparer les résultats observés avec les résultats attendus et enfin nous pourrons en tirer une conclusion.

IV.1 Réalisation de l'expérience

L'expérience est accessible en ligne sur l'adresse suivante : https://github.com/Nathanlauga/experience-memoire. La mise en ligne de mes travaux se justifie par ma volonté de transparence des observations et des résultats présentés. Cela garantie la possibilité de reproduire l'expérience à l'identique.

IV.1.1 Environnement de travail

Avant de commencer à coder une seule ligne, il est nécessaire d'installer les différents outils. Comme précisé dans le plan de recherche le jeu de données a été téléchargé depuis le site de Kaggle.

Dans un premier temps, nous devons identifier les librairies Python avec les bonnes versions et les écrire dans un fichier de prérequis, le fichier s'intitule "*requirements.txt*", il s'agit d'une norme au sein des développement en Python.

Nous avons donc les bonnes librairies avec leur bonne version pour garantir la reproductibilité. Avec ces informations nous pouvons créer le "*Dockerfile*" qui est le fichier où nous intégrons les installations dans un conteneurs Docker. Voici le code du fichier :

```
FROM python:3.7

RUN apt-get update
RUN pip install --upgrade pip

COPY ./requirements.txt /
RUN pip install -r requirements.txt
```

```
RUN mkdir /nb
RUN cd /nb

CMD ["jupyter", "notebook", "--port=8888", \
  "--ip=0.0.0.0", "--allow-root", "--notebook-dir=/nb"]
```

Ce code nous permet de lancer un conteneur avec Python en version 3.7, pour ensuite intaller les prérequis et d'automatiquement lancer les notebooks Jupyter qui seront notre espace de travail. Pour clôturer l'installation de l'environnement, nous mettons en place un script qui permet de lancer le conteneur Docker qui lie le répertoire "nb" du conteneur avec le dossier courant du projet.

Nous avons donc tous nos prérequis installé, il est temps de passer à la création du notebook de l'expérience et à l'analyse des données.

IV.1.2 Analyse des données

Pour réaliser une première analyse du jeu de données, j'ai créé un rapport en exploitant l'outil pandas-profiling (version 2.8.0)[1]. Concernant les informations globales du jeu de données, nous pouvons retenir qu'il y a 25 variables dont 22 quantitatives, deux catégoriques et une binaire. Il est également important de noter qu'il n'y a pas de valeur manquante, ni de lignes en doublons, cela nous permettra de ne pas perdre de temps sur la préparation des données. Enfin, il y a 30 000 lignes dans ce jeu de données.

Le plus intéressant dans l'analyse et de regarder dans le détail les variables du genre (*SEX*) et celle du défaut de paiement ou non () *default.payment.next.month*). Nous notons qu'il y a une plus grande proportion de femmes par rapport aux hommes : 60,4% contre 39,6% et que le défaut de paiement ne représente que 22,1% contre 77,9% de paiements valides (voir Figure IV.1 et Figure IV.2).

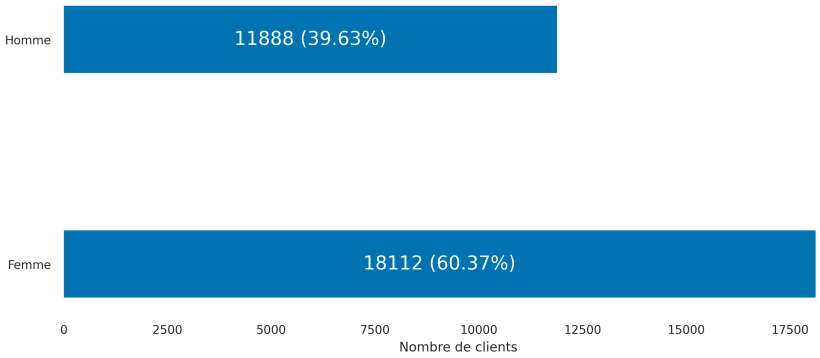

Figure IV.1: Affichage en barres de la variable *SEX*.

[1]Cet outil permet, en lui fournissant des données tabulaires en entrée, de générer une page HTML qui offre une analyse complète des données en question.

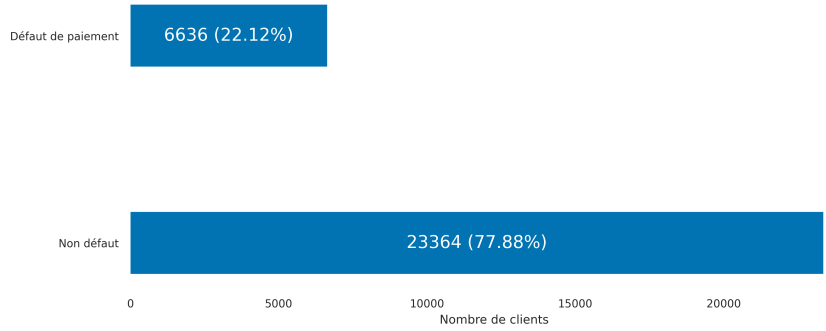

Figure IV.2: Affichage en barres de la variable *default.payment.next.month*.

Pour préparer la morale de Charline (soit la morale égalitaire), nous devons faire un test de corrélation avec les autres variables. Voici les cinq variables les plus liées au genre (les coefficients sont mis en valeur absolu afin de faire abstraction du signe et de se concentrer sur l'importance).

Variable	Coefficient de Pearson
SEX	1,000000
AGE	0,090874
PAY_2	0,070771
PAY_3	0,066096
PAY_4	0,060173

Nous pouvons constater qu'aucune variable si ce n'est elle-même n'est suffisamment corrélé avec le genre, l'âge étant le plus élevé avec seulement 9%. En conclusion pour la morale de Charline nous ne retirerons que la variable du genre.

IV.1.3 Entraînement des modèles et mise en place des morales

Pour entraîner les différents modèles nous devons dans un premier temps définir sur quels hyper-paramètres nous allons jouer afin de trouver les meilleurs. Nous allons faire la recherche aléatoire sur le nombre d'arbres qui seront présent dans la forêt, le nombre de variables maximales par arbre, la profondeur maximale, le nombre minimum d'échantillons par séparation, le nombre minimum d'échantillons par feuille et enfin si nous utilisons la technique du bootstrap[2] ou non. Nous pouvons donc mettre en place de dictionnaire Python suivant :

```
random_grid_rf = {
    'n_estimators': [500, 600, 700, 800, 900, 1000],
    'max_features': ['auto', 'sqrt'],
    'max_depth': [None, 10, 25, 40, 50, 60],
    'min_samples_split': [2, 5, 8],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}
```

[2]Il s'agit du fait de pouvoir tirer au hasard un échantillon qui a déjà été tiré au sort dans le passé.

La recherche des meilleurs hyper-paramètres se fait sur 50 itérations avec une validation croisée (avec $K = 5$) ce qui nous fait un total de 250 entraînements. La score a optimiser est le score F1, comme défini dans le protocole.

Après la recherche, nous obtenons les paramètres suivants :

```
{
   'bootstrap': True,
   'max_depth': 50,
   'max_features': 'sqrt',
   'min_samples_leaf': 2,
   'min_samples_split': 5,
   'n_estimators': 900,
}
```

Nous avons donc une forêt aléatoire de 900 arbres avec une profondeur maximale de 50. Grâce à cela nous avons déjà le premier modèle qui celui de la morale d'Alice. Pour obtenir la morale de Bob, nous devons mettre une fonction post prédiction : nous utilisons la probabilité prédite par le modèle de la morale d'Alice, si l'individu est une femme alors nous ajoutons 20% à la probabilité (sans dépasser 100%). Enfin pour la morale de Charline, nous récupérons les mêmes paramètres et nous entraînons le modèle en retirant la variable du genre.

IV.1.4 Contrôle de la performance selon les morales

A partir des modèles entraînés, nous pouvons récupérer les probabilités pour chaque morale et grâce à l'outil de TransparentAI, d'automatiquement calculer les mesures de perfomance choisies, soit le score F1, la précision et le rappel :

```
metrics = ['f1','precision','recall']

def compute_metrics(predictions):
    classification.compute_metrics(y_test,
                                   predictions,
                                   metrics)

perf_morale_a_rf = compute_metrics(y_pred_morale_a)
perf_morale_b_rf = compute_metrics(y_pred_morale_b)
perf_morale_c_rf = compute_metrics(y_pred_morale_c)
```

Grâce au code ci-dessus, nous obtenons le tableau de comparaison suivant :

Morale	Score F1	Précision	Rappel
Morale Alice	0,464478	0,651099	0,361005
Morale Bob	0,504587	0,557604	0,460777
Morale Charline	0,467545	0,650815	0,364813

Nous observons que concernant le score F1 la meilleure intelligence artificielle est celle de Bob, puis Charline et enfin celle d'Alice. Pour la précision, c'est Alice et Charline qui sont supérieures à Bob, mais sur le rappel Bob a l'ascendant sur ces dernières.

Sur le point de vue de la performance, le choix porterait sur le modèle de Bob, puisqu'il est plus performant, son défaut vient sur la précision, mais 55% reste abordable car cela signifie que seulement 45% des clients considérés avec un défaut de paiements potentiel auraient en réalité pu payer leurs dettes.

IV.1.5 Contrôle des biais selon les morales

Dans cette partie, nous utilisons le sous module *transparentai.fairness* de l'outil TransparentAI. Ce sous module nous permet de calculer automatiquement les quatre mesures de biais qui nous intéresse. Voici le code nous permettant de l'obtenir :

```
privileged_group = {'SEX': ['Male']}

def compute_bias(predictions):
    return fairness.model_bias(y,
                               y_pred_a,
                               df,
                               privileged_group,
                               pos_label=0)['SEX']

bias_a = compute_bias(y_pred_a)
bias_b = compute_bias(y_pred_b)
bias_c = compute_bias(y_pred_c)
```

Il faut préciser que la fonction *fairness.model_bias* a pour paramètre le terme de *pos_label* qui sert à spécifier quel résultat du modèle peut être considéré comme un résultat avantageux, ici c'est le fait de ne pas être détecté comme potentiellement risqué. Grâce à cela voici les résultats obtenus :

Morale	Statistical Parity Difference	Disparate Impact	Equal Opportunity Difference	Average Odds Difference
Morale Alice	0,009644	1,010255	-0,000120	0,003792
Morale Bob	-0,052856	0,943797	-0,005207	-0,139464
Morale Charline	0,009799	1,010432	-0,000004	0,003734

Pour rappel, concernant les mesures de *Statistical Parity Difference*, *Equal Opportunity Difference* et *Average Odds Difference*, aucun biais signifie un score de 0. Pour la mesure de *Disparate Impact* l'objectif est de 1.

Nous pouvons constater que le modèle le moins biaisé est celui de Charline qui possède le meilleur score sur chaque mesure. De plus, la morale d'Alice n'est pas loin derrière : nous pouvons considéré que ces deux morales ne sont pas du tout biaisées. Par contre pour Bob, une mesure demande à réfléchir : l'*Average Odds Difference*.

En effet, pour cette mesure, nous pouvons conclure que pour un homme, le modèle de Bob prédit correctement un client 13,95% plus souvent que pour une femme. Cette mesure de biais, bien qu'intéressant reste assez faible : il est donc possible qu'un arbitrage décide d'ignorer cette mesure puisque la performance du modèle est très intéressante.

IV.1.6 Explicabilité

Pour finir cette expérience, nous allons essayer de voir comment se comporte les modèle en général. Plus particulièrement nous regarderons les modèles d'Alice et de Bob pour comparer l'influence des variables surtout sur l'attribut du genre.

Nous utiliserons le sous module suivant *transparentai.models.explainers* de TransparentAI qui nous permettra d'expliquer les modèles globalement et localement. Ci-dessous l'influence des variables fournies par l'outil. La Figure IV.4 nous montre l'influence globale pour le modèle d'Alice et la Figure IV.3 pour Bob.

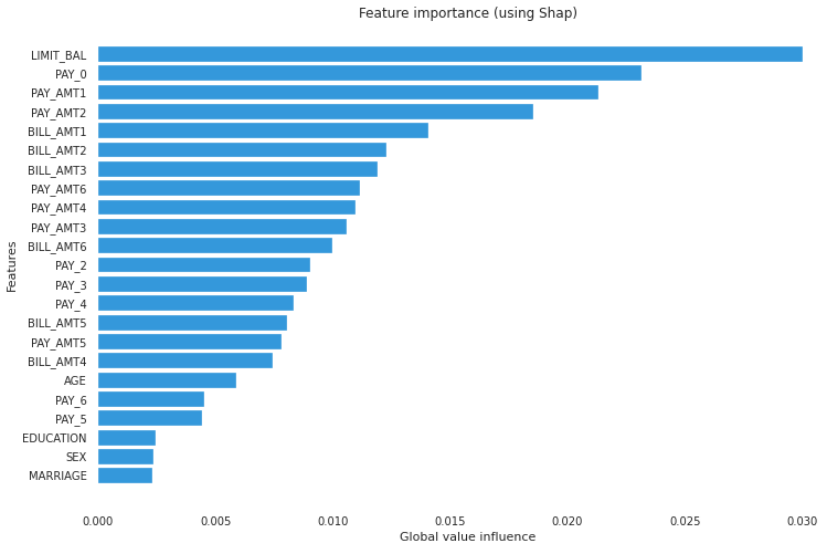

Figure IV.3: Importance des variables pour la morale d'Alice.

Nous pouvons constater que la différence fondamentale vient de la variable la plus importante. En effet avec le modèle de Bob l'attribut du genre domine largement alors que pour Alice le genre ne fait même pas parti des variables les plus influentes.

Si nous regardons dans le détail une prédiction spécifique pour une homme et une prédiction pour une femme, afin de mieux voir les différences voici ce que nous pouvons observer dans l'expérience. Pour ce même homme nous voyons que la morale de Bob ajoute une influence très importante pour le genre (ici environ -19%) contre une influence presque égale à 0 pour Alice.

Ensuite, nous prenons au hasard dix hommes et dix femmes et pour afficher l'influence du genre pour voir si l'influence est belle et bien proche de -20% pour les hommes. La Figure IV.5 nous montre deux lignes séparées pour la morale d'Alice et celle de Bob avec une différenciation entre les dix hommes et les dix femmes.

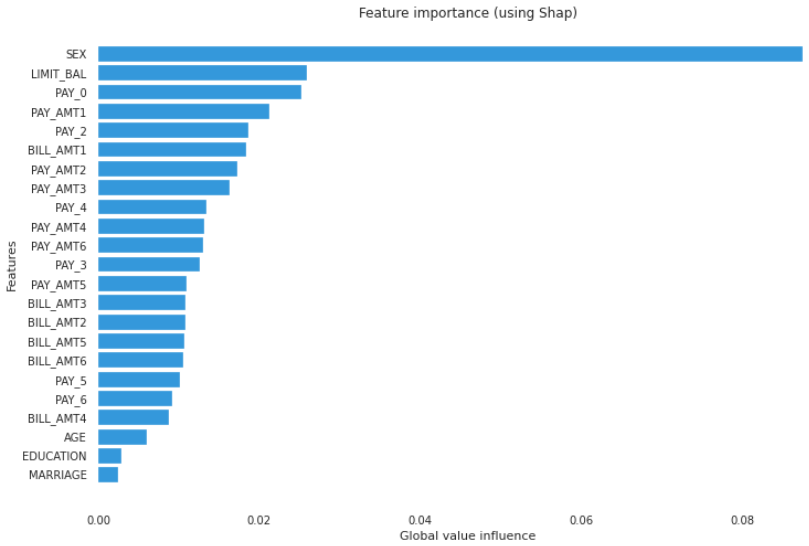

Figure IV.4: Importance des variables pour la morale de Bob.

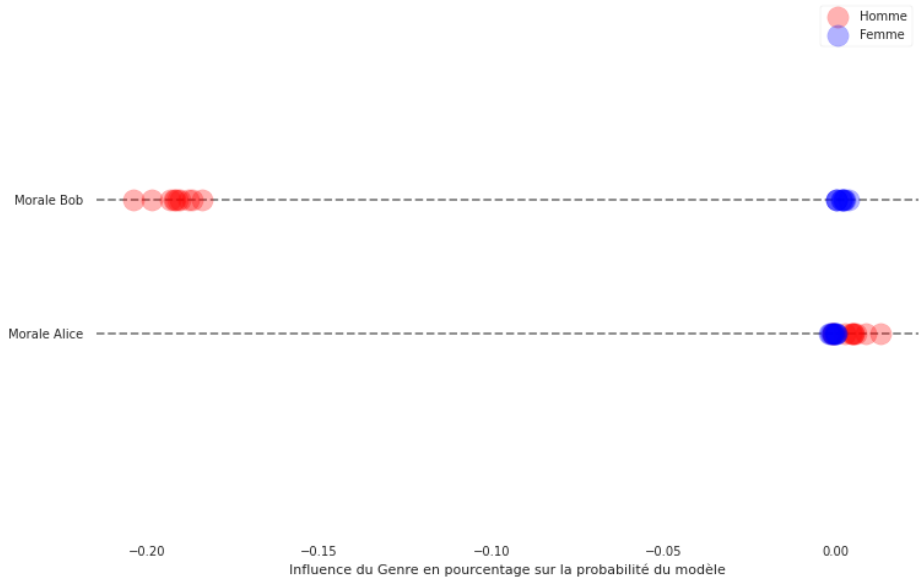

Figure IV.5: Comparaison entre les morales d'Alice et de Bob sur dix hommes et dix femmes.

IV.2 Résultats obtenus

Nous allons dans cette section comparer les résultats obtenus avec ce que nous attendions lors de la définition de l'expérience. L'analyse portera sur trois points : la performance, les biais et l'explicabilité.

Performance Nous nous attendions au résultat suivant : L'IA de contrôle d'Alice devait être la plus performante suivi de Bob et enfin de Charline, or, c'est basé sur les critères du score F1, de la précision et du rappel nous observons que c'est Bob (morale discriminante) qui est le plus performant (sauf sur la précision). Concernant les deux autres modèles, nous pouvons estimer qu'ils sont à égalité. Nous avons donc un modèle qui a été altété consciemment pour impacter négativement les femmes, mais qui plus performant donc par conséquent plus attractif d'un point de vue métier. Cette observation rajoute encore plus d'importance à la nécessité de détecter les biais.

Biais En toute logique nous devions observer un biais faible voire inexistant pour le modèle de Charline et à contrario un biais plus ou moins important pour Bob. Dans les faits, c'est bien le cas pour Charline et il en va de même pour Alice. Concernant Bob, nous observons bien des mesures biais plus éloignées par rapport à l'objectif équitable, mais les scores restent faibles et il serait même acceptable pour un chef de projet de considérer que le biais (étant présent sur une mesure sur quatre) puisse être ignoré au vu de la performance du modèle.

Explicabilité Dans cette dernière étape, les attentes se jouaient surtout entre Alice et Bob qui possédent la variable du genre, en comparant leur influence des variables globale, nous sommes forcés de constater qu'en effet la morale de Bob accorde bien plus d'importance au genre dans la prise de décision que pour Alice (1ère pour Bob contre 22ème pour Alice). Puis concernant les autres variables, l'ordre d'importance n'évolue pas. L'observation qui est très importante est celle concernant l'influence des variables localement, en effet pour dix hommes et dix femmes tirés au hasard, pour chaque homme avec le modèle de Bob nous observions une influence d'environ -20% dans la probabilité ; cela correspond exactement à ce que nous avions défini pour rendre le modèle discriminant, même si la règle était inversée (si c'est une femme alors nous rajoutons 20%).

IV.3 Conclusion de l'expérience

Pour conclure l'expérience, il est nécessaire de rappeler le contexte de l'expérience : il s'agit d'un modèle peu complexe avec une faible quantité de variables en entrée, ce qui n'est pas représentatif des modèles dans le monde réel.

Concernant les trois morales définies, nous avons pu observer dans ce cas précis qu'entre les algorithmes de contrôle et égalitaire, il n'y avait pas de grande différence : cela s'explique avec le fait que la cible du défaut de paiement ou non n'est pas très corrélé avec cette variable.

Dans ce contexte, selon les incentives qui sont utilisés, le modèle à choisir diffère. Si nous nous basons sur la performance (qui est le critère le plus important pour une décision métier) c'est la morale discriminante qui l'emporte, si nous analysons les biais, selon les décideurs sur un projet il est également possible de choisir la morale discriminante. Il faut attendre le

critère de l'explicabilité, pour être certain de rejeter le modèle, pour lequel nous identifions bien le biais présent pour favoriser les hommes par rapport aux femmes.

Ce que nous tirons de cette expérience positivement, c'est que dans un premier temps l'outil est simple d'utilisation pour un développeur d'IA (maîtrisant Python). Ensuite, l'outil a identifié le biais et à même quantifié : 20% en moins pour les hommes correspondant à notre définition de la morale discriminante.

Cette expérience possède également des limites : comme précisé au-dessus, ces données et cette problématique n'est pas représentative, par exemple, nous n'avons pas pu voir le cas où un développement d'un algorithme visant uniquement la performance sans vouloir discrimer un groupe social était en réalité biaisé (l'équivalent de la morale de contrôle ici). De plus, sans sensibilisation autour des enjeux de l'éthique un modèle est très souvent validé uniquement sur le critère de la performance, même s'il est complètement opaque. Enfin, nous pouvons imaginer qu'un développeur malvaillant puisse ruser en exploitant une combinaison de différents modèles en analysant leur explicabilité. Par exemple, ici il aurait pu choisir d'utiliser le modèle de contrôle si l'entrée était un homme et le modèle discriminant pour une femme, car l'explicabilité était constaté uniquement si l'individu en entrée était un homme.

Bien que nous pouvons voir des limites, si une équipe en charge de construire un algorithme intelligent est sensible au sujet de l'éthique, nous pouvons affirmer que TransparentAI permet de fournir une facilité à l'accès aux réponses sur certains aspects plus qu'importants pour pouvoir assurer la transparence d'un modèle et par conséquent de tendre vers une intelligence artificielle de confiance.

Conclusion

Tout au long ce 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 I.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 refait 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 I.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 I.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 mémoire a introduit une solution (voir section II.2). Les méthodes qui ont été utilisées ont deux axes, celui de la création et de la définition de l'outil Transparentai. Le second sera axé sur une expérience qui exploitera l'outil en tant que tel pour observer la détecter d'une intelligence artificielle discriminante.

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.

Suite à la création de l'outil et à la réalisation de l'expérience, nous avons observé qu'en effet l'outil offre une facilité à l'accès pour réaliser une IA éthique et par ce biais, TransparentAI permet de rapprocher la théorie à la pratique. L'expérience a également identifié des limites, mais en bref, si l'équipe d'un projet est consciente des enjeux de l'éthique, alors l'outil répond à un besoin plus qu'indiscutable.

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 pallier aux problématiques d'ordre moral. TransparentAI répond à ce besoin moderne et bien qu'il ne soit, aujourd'hui (Juillet 2020) qu'uniquement disponible dans sa version pour des utilisateurs technique, la version graphique pour une utilisation universelle est prévue dans l'année. De plus, grâce à la philosophie de TransparentAI, si l'outil gagne en utilisateur, alors il deviendra naturellement très fonctionnel et mon espoir porte vers une utilisation quotidienne dans les entreprises du monde entier pour garantir une intelligence artificielle de confiance.

Pour conclure, bien que les travaux 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 : l'humain contrôle la machine.

Bibliography

Algorithm Watch. AI Ethics Guidelines Global Inventory by AlgorithmWatch, April 2020. URL `https://inventory.algorithmwatch.org`. Library Catalog: inventory.algorithmwatch.org.

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018. URL `http://arxiv.org/abs/1806.08049`.

Aristote. *Ethique à Nicomaque*. Vrin, revised edition, January 1994. ISBN 978-2-7116-0022-9.

Rita Astuti. La moralité des conventions : tabous ancestraux à Madagascar. *Terrain. Anthropologie & sciences humaines*, (48):101–112, February 2007. ISSN 0760-5668. doi: 10.4000/terrain.5041. URL `http://journals.openedition.org/terrain/5041`. ISBN: 9782735111312 Number: 48 Publisher: Association Terrain.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The Moral Machine experiment. *Nature*, 563(7729):59, November 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0637-6. URL `https://www.nature.com/articles/s41586-018-0637-6`.

Nicolas Baumard, Olivier Mascaro, and Coralie Chevallier. Preschoolers are able to take merit into account when distributing goods. *Developmental Psychology*, 48(2):492–498, 2012. ISSN 1939-0599(Electronic),0012-1649(Print). doi: 10.1037/a0026598. Place: US Publisher: American Psychological Association.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv:1810.01943 [cs]*, October 2018. URL `http://arxiv.org/abs/1810.01943`. arXiv: 1810.01943.

Franz Boas. Museums of Ethnology and Their Classification. *Science*, 9(228):587–589, 1887. ISSN 0036-8075. URL `https://www.jstor.org/stable/1762958`. Publisher: American Association for the Advancement of Science.

Thomas J. Bouchard and Matt McGue. Genetic and environmental influences on human psychological differences. *Journal of Neurobiology*, 54(1):4–45, January 2003. ISSN 0022-3034. doi: 10.1002/neu.10160.

Vannevar Bush. As We May Think, July 1945. URL https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

Cambridge. DATA | signification, définition dans le dictionnaire Anglais de Cambridge, 2020. URL https://dictionary.cambridge.org/fr/dictionnaire/anglais/data. Library Catalog: dictionary.cambridge.org.

J. Clement. Global social media ranking 2019, 2019. URL https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Nicolas Cointe. *Ethical Judgment for decision and cooperation in multiagent systems*. Theses, Université de Lyon, December 2017. URL https://tel.archives-ouvertes.fr/tel-01851485.

Louis Columbus. 25 Machine Learning Startups To Watch In 2019, May 2019. URL https://www.forbes.com/sites/louiscolumbus/2019/05/27/25-machine-learning-startups-to-watch-in-2019/. Library Catalog: www.forbes.com Section: Innovation.

UE Commission. Ethics guidelines for trustworthy AI, April 2019. URL https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Daniel Crevier. *AI: the tumultuous history of the search for artificial intelligence*. Basic Books, New York, NY, 1992. ISBN 978-0-465-02997-6 978-0-465-00104-0. OCLC: 26858345.

Data for Good. Serment d'Hippocrate pour data scientist, 2018. URL https://www.hippocrate.tech.

DeepMind. AlphaGo, 2016. URL https://deepmind.com/research/alphago/.

Jeff Desjardins. Infographic: What Happens in an Internet Minute in 2018?, May 2018. URL https://www.visualcapitalist.com/internet-minute-2018/.

Jeff Desjardins. What Happens in an Internet Minute in 2019?, March 2019. URL https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/. Library Catalog: www.visualcapitalist.com.

Louis Dorad. Machine Learning Canvas, 2016. URL https://www.louisdorard.com/machine-learning-canvas.

Jean Ducat. Du vol dans l'éducation spartiate. In *Dossier : Alexandre le Grand, religion et tradition*, Mètis, pages 95–110. Éditions de l'École des hautes études en sciences sociales, Paris, June 2017. ISBN 978-2-7132-2599-4. URL http://books.openedition.org/editionsehess/2103.

Amnon H. Eden, James H. Moor, Johnny H. Soraker, and Eric Steinhart, editors. *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Springer, New York, 2012 edition edition, April 2013. ISBN 978-3-642-32559-5.

Figure_Eight. The State of AI and Machine Learning Report, May 2019. URL https://www.figure-eight.com/the-state-of-ai-and-machine-learning-report/. Library Catalog: www.figure-eight.com Section: eBooks.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00069 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1806.00069. arXiv: 1806.00069.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When Will AI Exceed Human Performance? Evidence from AI Experts. *arXiv:1705.08807 [cs]*, May 2017. URL http://arxiv.org/abs/1705.08807. arXiv: 1705.08807.

A. G. Greenwald and M. R. Banaji. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27, January 1995. ISSN 0033-295X.

Alexei Grinbaum. « Le jugement éthique est une affaire d'humains, pas de robots », May 2019a. URL https://usbeketrica.com/article/jugement-ethique-affaire-humains-pas-robots-hasard. Library Catalog: usbeketrica.com.

Alexei Grinbaum. La Conversation scientifique, Des machines pourront-elles apprendre le bien et le mal ?, March 2019b. URL https://www.franceculture.fr/emissions/la-conversation-scientifique/des-machines-pourront-elles-apprendre-le-bien-et-le-mal.

Alexei Grinbaum. *Les robots et le mal*. Desclée De Brouwer, January 2019c. ISBN 978-2-220-09594-3.

Jonathan Grudin. AI and HCI: Two Fields Divided by a Common Focus. *AI Magazine*, 30(4):48, September 2009. ISSN 0738-4602, 0738-4602. doi: 10.1609/aimag.v30i4.2271. URL https://aaai.org/ojs/index.php/aimagazine/article/view/2271.

Michael Gurven. To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences - BEHAV BRAIN SCI*, 27, August 2004. doi: 10.1017/S0140525X04000123.

J. Kiley Hamlin, Karen Wynn, and Paul Bloom. Social evaluation by preverbal infants. *Nature*, 450(7169):557–559, November 2007. ISSN 1476-4687. doi: 10.1038/nature06288.

Yuval N Harari and Pierre-Emmanuel Dauzat. *Sapiens: une brève histoire de l'humanité*. Albin Michel, Paris, 2015. ISBN 978-2-226-25701-7. OCLC: 1082432348.

Michael Haupt. Who should get credit for the quote 'data is the new oil'? - Quora, 2006. URL https://www.quora.com/Who-should-get-credit-for-the-quote-data-is-the-new-oil.

Lê Nguyên Hoang. A Roadmap for Robust End-to-End Alignment. *arXiv:1809.01036 [cs]*, September 2018. URL http://arxiv.org/abs/1809.01036. arXiv: 1809.01036.

Feng-hsiung Hsu, Murray S. Campbell, and A. Joseph Hoane, Jr. Deep Blue System Overview. In *Proceedings of the 9th International Conference on Supercomputing*, ICS '95, pages 240–244, New York, NY, USA, 1995. ACM. ISBN 978-0-89791-728-5. doi: 10.1145/224538.224567. URL http://doi.acm.org/10.1145/224538.224567. event-place: Barcelona, Spain.

Geoffrey Irving and Amanda Askell. AI Safety Needs Social Scientists. *Distill*, 4(2): 10.23915/distill.00014, February 2019. ISSN 2476-0757. doi: 10.23915/distill.00014. URL https://distill.pub/2019/safety-needs-social-scientists.

Peter Jackson. *Introduction to Expert Systems*. Addison-Wesley Longman Publishing Co., Inc., USA, 3rd edition, 1998. ISBN 978-0-201-87686-4.

Jeremy Jordan. Organizing machine learning projects: project management guidelines., September 2018. URL https://www.jeremyjordan.me/ml-projects-guide/.

Immanuel Kant and Victor Delbos. *Fondements de la métaphysique des mœurs*. Libraire Delagrave, Paris, 2007. ISBN 978-2-206-00155-5. OCLC: 612252890.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv:1812.04948 [cs, stat]*, December 2018. URL http://arxiv.org/abs/1812.04948. arXiv: 1812.04948.

Garry Kasparov. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. John Murray, June 2017. ISBN 978-1-4736-5350-4.

Lawrence Kohlberg and Richard H. Hersh. Moral development: A review of the theory. *Theory Into Practice*, 16(2):53–59, April 1977. ISSN 0040-5841, 1543-0421. doi: 10.1080/00405847709542675. URL http://www.tandfonline.com/doi/abs/10.1080/00405847709542675.

Adam Kramer, Jamie Guillory, and Jeffrey Hancock. Correction for Kramer et al., Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(29):10779–10779, July 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1412583111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1412583111.

Charles Krauthammer. Be Afraid, May 1997. URL https://www.weeklystandard.com/charles-krauthammer/be-afraid-9802.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Jeff Larson and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm, May 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Mickaël Launay. *Urnes interagissantes*. thesis, Aix-Marseille, June 2012. URL http://www.theses.fr/2012AIXM4775.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Matthew Mayo. The Machine Learning Project Checklist, December 2018. URL https://www.kdnuggets.com/the-machine-learning-project-checklist.html/.

Gabriele Medeot, Srikanth Cherla, Katerina Kosta, Matt McVicar, Samer Abdallah, Marco Selvi, Ed Newton-Rex, and Kevin Webster. StructureNet: Inducing Structure in Generated Melodies. In *ISMIR*, 2018.

Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

Michel de Montaigne. *Essais de Montaigne*. Volland, 1789. Google-Books-ID: 5I0tAAAAMAAJ.

G.E. Moore. Cramming More Components Onto Integrated Circuits. *Proceedings of the IEEE*, 86(1):82–85, January 1998. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.1998.658762. URL http://ieeexplore.ieee.org/document/658762/.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going Deeper into Neural Networks, June 2015. URL http://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *arXiv:1905.06876 [cs]*, September 2019. URL http://arxiv.org/abs/1905.06876. arXiv: 1905.06876.

newsflash. 'AI is very, very stupid,' says Google's AI leader, at least compared to humans - CNET, November 2018. URL https://newsflash.one/2018/11/14/ai-is-very-very-stupid-says-googles-ai-leader-at-least-compared-to-humans-cnet/.

Friedrich Nietzsche. *La Généalogie de la morale*, volume Œuvres complètes de Frédéric Nietzsche, vol. 11. Mercure de France, 1900. URL https://fr.wikisource.org/wiki/La_G%C3%A9n%C3%A9alogie_de_la_morale/Premi%C3%A8re_dissertation.

Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. Turing Test: 50 Years Later. *Minds and Machines*, 10(4):463–518, November 2000. ISSN 1572-8641. doi: 10.1023/A:1011288000451. URL https://doi.org/10.1023/A:1011288000451.

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

Gil Press. A Very Short History Of Big Data, May 2013. URL https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/. Library Catalog: www.forbes.com Section: Innovation.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24, 2019.

Robert Redfield. *The Primitive World and Its Transformations*. Cornell University Press, 1965. Google-Books-ID: KHQLAAAAYAAJ.

David Reinsel, John Gantz, and John Rydning. The Digitization of the World from Edge to Core. page 28, November 2018. URL https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 1939-1471, 0033-295X. doi: 10.1037/h0042519. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0042519.

Jean-Jacques Rousseau. *Rousseau : Oeuvres complètes, tome 4*. Gallimard, Paris, April 1969. ISBN 978-2-07-010491-8.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Jean Servier. *Méthode de l'ethnologie*. Que sais-je ? PUF, Paris, 2e éd. rev. edition, 1993. ISBN 978-2-13-045905-7.

L. S. Shapley. 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, volume 2. Princeton University Press, Princeton, 1953. ISBN 978-1-4008-8197-0. doi: 10.1515/9781400881970-018. URL https://www.degruyter.com/view/books/9781400881970/9781400881970-018/9781400881970-018.xml.

David Shepardson and Alexandria Sage. Waymo gets first California OK for driverless testing without... *Reuters*, October 2018. URL https://in.reuters.com/article/us-autos-selfdriving-waymo-idINKCN1N42S1.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, October 2017. ISSN 1476-4687. doi: 10.1038/nature24270. URL https://www.nature.com/articles/nature24270.

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

Herbert A. Simon and Allen Newell. Heuristic Problem Solving: The Next Advance in Operations Research. *Operations Research*, 6(1):1–10, February 1958. ISSN 0030-364X. doi: 10.1287/opre.6.1.1. URL https://pubsonline.informs.org/doi/abs/10.1287/opre.6.1.1.

R. J. Solomonoff. The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management*, 5(2):149–153, January 1985. ISSN 0167-2533. doi: 10.3233/HSM-1985-5207. URL https://content.iospress.com/articles/human-systems-management/hsm5-2-07.

William Graham Sumner. *Folkways, a study of the sociological importance of usages, manners, customs, mores, and morals*. Boston, Ginn, 1906. URL http://archive.org/details/folkwaysstudyofs00sumnuoft.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36(4):1–13, July 2017. ISSN 07300301. doi: 10.1145/3072959.3073640. URL http://dl.acm.org/citation.cfm?doid=3072959.3073640.

TheEconomist. Progress and its perils | Dec 19th 2009, December 2009. URL https://www.economist.com/weeklyedition/2009-12-19. Library Catalog: www.economist.com.

Van-Tinh Tran. *Selection Bias Correction in Supervised Learning with Importance Weight*. Artificial Intelligence [cs.AI], Université de Lyon, July 2017. URL https://tel.archives-ouvertes.fr/tel-01661470.

Morgane Tual. A peine lancée, une intelligence artificielle de Microsoft dérape sur Twitter. *Le Monde.fr*, March 2016. URL https://www.lemonde.fr/pixels/article/2016/03/24/a-peine-lancee-une-intelligence-artificielle-de-microsoft-derape-sur-twitter_4889661_4408996.html.

Matteo Turilli. Ethical protocols design. *Ethics and Information Technology*, 9(1):49–62, March 2007. ISSN 1572-8439. doi: 10.1007/s10676-006-9128-9. URL https://doi.org/10.1007/s10676-006-9128-9.

A. M. Turing. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236): 433–460, October 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://academic.oup.com/mind/article/LIX/236/433/986238.

Alan Turing. Browse the Turing Digital Archive, May 1951. URL http://www.turingarchive.org/browse.php/B/5.

Cédric Villani. *Donner un sens à l'intelligence artificielle: pour une stratégie nationale européenne : [Mission parlementaire du 8 septembre 2017 au 8 mars 2018*. 2018. ISBN 978-2-11-145700-3. OCLC: 1030337149.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M. Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, Timo Ewalds, Dan Horgan, Manuel Kroiss, Ivo Danihelka, John Agapiou, Junhyuk Oh, Valentin Dalibard, David Choi, Laurent Sifre, Yury Sulsky, Sasha Vezhnevets, James Molloy, Trevor Cai, David Budden, Tom Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Toby Pohlen, Yuhuai Wu, Dani Yogatama, Julia Cohen, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Chris Apps, Koray Kavukcuoglu, Demis Hassabis, and David Silver. *AlphaStar: Mastering the Real-Time Strategy Game StarCraft II*. 2019. URL https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/.

Norbert Wiener. *Cybernetics; or, Control and communication in the animal and the machine*. M.I.T. Press, New York, 1961. ISBN 978-0-262-23007-0 978-0-262-73009-9. OCLC: 1284210.