Université de Montréal

Tensions en éthique de l'intelligence artificielle (IA)

Un guide herméneutique pour les décideurs politiques

Par

Anne Boily

Département de science politique, Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade *Philosophiæ Doctor* (Ph.D)

Décembre 2020

Université de Montréal

Département de science politique, Faculté des arts et des sciences

---

Cette thèse intitulée

Tensions en éthique de l'intelligence artificielle (IA)

Un guide herméneutique pour les décideurs politiques

Présentée par

Anne Boily

A été évaluée par un jury composé des personnes suivantes

Jean-François Godbout
Président-rapporteur

Charles Blattberg
Directeur de recherche

Pascale Devette
Membre du jury

Charles Ess
Examinateur externe

Résumé

L'éthique de l'intelligence artificielle (IA) constitue un domaine de recherche pluridisciplinaire en expansion. Cette thèse s'inscrit dans le champ de l'éthique de l'IA et en propose une nouvelle interprétation. D'entrée de jeu, les nombreux liens qu'entretiennent les défis et opportunités que présentent les systèmes employant l'intelligence artificielle avec le domaine politique y sont exposés. Une hypothèse clé animant cette thèse est que, puisque la politique consiste à répondre au conflit par le dialogue, les décideurs politiques aux prises avec des questions concernant l'IA peuvent tirer profit des orientations fournies par différentes traditions éthiques. Afin de faciliter un dialogue optimal, une prise de position métaéthique particulière est avancée au niveau théorique, soit le « monisme non orthodoxe », tandis qu'une série de questions ciblées est proposée au niveau de la pratique.

Divisée en trois sections, la thèse débute avec une exploration métaéthique des fondements des approches éthiques principales qui sont à l'œuvre dans les réflexions contemporaines concernant l'IA. Les écoles éthiques étudiées, en faisant appel au continuum qui distingue le monisme et le pluralisme, sont l'éthique de la vertu, l'utilitarisme et l'éthique déontologique. Ces démarches monistes sont ensuite placées en contraste avec le pluralisme des valeurs, une approche souvent employée, mais rarement nommée de manière explicite.

La deuxième section consiste en une analyse métaéthique d'une vingtaine de directives éthiques émises par des compagnies privées, la société civile ainsi que des organisations à multiples partenaires, de même que par des instances gouvernementales ou intergouvernementales. C'est ce portrait qui révèle à quel point le pluralisme des valeurs est récurrent dans ces directives. En outre, il se mêle souvent à d'autres approches éthiques pour générer des versions en situation de « tension métaéthique », bien que cela ne se produise souvent que de manière implicite. En conséquence, les propositions associées à ces approches sont parfois contradictoires, tant en ce qui concerne leur formulation que dans la manière dont elles seraient mises en œuvre.

Une approche éthique alternative est proposée dans la troisième section. Cette approche est formée d'éléments dérivés spécialement de l'éthique de la vertu et du pluralisme. Ils fondent le

socle sur lequel, dans le chapitre final de la thèse, un « guide dialogique » est développé pour l'usage des décideurs politiques. La prudence, une sensibilité profonde au contexte, une orientation téléologique « douce » vers le bien commun ainsi que l'ouverture à la possibilité de dilemmes insolubles caractérisent cette approche éthique. La philosophie herméneutique est également mise à contribution pour justifier l'articulation d'une série de questions destinées à guider le dialogue des décideurs politiques. En effet, l'herméneutique encourage une logique de questions, plutôt qu'une logique dérivée de principes ou de théories.

Mots-clés : éthique, intelligence artificielle (IA), politique, décideurs, dialogue, herméneutique, monisme, pluralisme, dilemme, tension.

Abstract

The ethics of artificial intelligence (AI) is a growing field of multidisciplinary research. This thesis falls within AI ethics and suggests a new interpretation. The numerous challenges and opportunities generated by AI systems are described at the outset. A key assumption of the thesis is that, given that politics consists of responding to conflict with dialogue, policy-makers dealing with the questions surrounding AI can benefit from the guidance provided by different ethical traditions. Furthermore, in order to facilitate an optimal dialogue, this thesis puts forward a particular metaethical position, that of "unorthodox monism" at the theoretical level, and one consisting of a series of pertinent questions at the practical level.

Divided into three main sections, the thesis begins by exploring the metaethical foundations at work in most contemporary thinking about AI. Using a continuum that distinguishes between monism and pluralism in ethics, the particular schools examined are virtue ethics, utilitarianism, and deontology. These monistic approaches are then contrasted with value pluralism, which is an approach that is often employed and yet rarely identified explicitly.

The second section consists of a metaethical analysis of a sample of some twenty sets of AI ethical guidelines produced within private companies, civil society and by multi-partner organizations, as well as by governmental and intergovernmental bodies. It is this portrait which reveals how value pluralism recurs in many of these sets of guidelines. Moreover, it is also often implicitly combined with other ethical approaches to generate "mixed" versions, which exhibit metaethical tensions. As a result, the proposals associated with these approaches are sometimes contradictory, both in regards to how they are formulated as well as to how they are to be implemented.

Third, an alternative ethical approach is proposed. It consists of elements derived in particular from virtue ethics and value pluralism. They form the basis upon which, in the thesis' final chapter, a "dialogical guide" is formulated for use by policy-makers. Prudence, a deep sensitivity to context, a "soft" teleological orientation towards a common good, along with openness to the possibility of intractable dilemmas characterize this approach. Hermeneutical

philosophy is also drawn upon in order to justify the articulation of a series of questions meant to guide the dialogue of policy-makers. Indeed, hermeneutics calls forth a logic of questions rather than one based on principles or theories.

Keywords: ethics, artificial intelligence (AI), politics, policy-makers, dialogue, hermeneutics, monism, pluralism, dilemma, tension.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ACM : Association for Computing Machinery

AoIR : Association of Internet Researchers

COMEST : Commission mondiale d'éthique des connaissances scientifiques et des technologies de l'UNESCO

DARPA: Defense Advanced Research Projects Agency

FEM : Forum économique mondial (pour WEF: World Economic Forum)

FLI : Future of Life Institute

GEAH : Groupe d'experts ad hoc pour l'élaboration d'un projet de recommandation sur l'éthique de l'intelligence artificielle de l'UNESCO

GEHN IA : Groupe d'experts de haut niveau sur l'intelligence artificielle de la Commission européenne (pour HLEG AI : High-Level Expert Group on Artificial Intelligence)

G7 : Groupe des sept

G20 : Groupe des vingt

IA : Intelligence artificielle

IAG (AGI) : Intelligence artificielle générale/généralisée

IBM : International Business Machines

IEEE : Institute of Electronics and Electrical Engineers, Incorporated

ITU: International Telecommunication Union of the United Nations

IVADO : Institut de valorisation des données

LPRPDE : Loi sur la protection des renseignements personnels et les documents électroniques

Mila: Montreal Institute of Learning Algorithms

MAIEI : Montreal AI Ethics Institute

OCDE : Organisation de coopération et de développement économiques (pour OECD: Organisation for Economic Co-operation and Development)

ONU : Organisation des Nations Unies

PAI : Partnership on Artificial Intelligence

RGPD : Règlement général sur la protection des données de l'Union européenne

SIA : Système d'intelligence artificielle/employant l'intelligence artificielle

UE : Union européenne

UNDG : United Nations Development Group

UNESCO : Organisation des Nations Unies pour l'Éducation, la Science et la Culture

XAI : Explainable Artificial Intelligence

À mes parents, dont la vie est le modèle éthique le plus inspirant qui soit.

Remerciements

Il faudrait un chapitre entier pour remercier toutes les personnes qui m'ont entourée et appuyée lors de mon parcours doctoral. Je commencerais par mon directeur de recherche, Charles Blattberg. Quand je jette un regard sur les dernières années, le mot principal qui me vient à l'esprit est « croissance ». Charles m'a fait découvrir de nouveaux penseurs et présenté une vision enrichie de la réalité. Il m'a, par le fait même, enseigné le métier de philosophe politique. Sa patience et sa grande ouverture, de même que sa disponibilité indéniable, ont été pour moi un appui solide. Charles représente pour moi un alliage de sincérité, d'audace et d'ouverture dans le monde de la pensée, et c'est un modèle que j'aimerais imiter dans les années à venir.

J'aimerais remercier les professeurs Charles Ess, Pascale Devette et Jean-François Godbout pour leur lecture attentive de la thèse, leurs commentaires et suggestions, de même que pour leur grande générosité herméneutique. Leurs conseils m'ont beaucoup aidée à améliorer et préciser mon propos. Les échanges que nous avons eus ont été particulièrement gratifiants. Je remercie aussi Marc-Antoine Dilhac pour ses commentaires sur la première ébauche de mon projet.

Ma famille a été absolument cruciale dans mon parcours, et je veux remercier tout spécialement mes parents, mes frères et sœurs pour leur écoute, leur intérêt, leur appui et leurs prières. Un merci spécial à Jane, qui a été une oreille plus que patiente, encourageante et qui a cru en moi, ainsi qu'à Inma, Kirstie, Sonia, Susana, Judy, Mich, Rebeca et Natalie. Votre expérience et votre sagesse m'ont beaucoup servie. Toute ma reconnaissance à Virginia, Marie-Chantale, Tripti, Yin et Diana pour vos bons soins. Merci à Flora pour ton optimisme, tes encouragements et ta solidarité dans nos aventures intra et extra-universitaires, à Anjeza pour ton écoute, ta douceur et ta sagesse, à Vertika pour ton sens de l'humour et ta joie de vivre, à Katherine pour les découvertes que tu m'as fait faire, à Natasha pour avoir été la meilleure compagne d'étude qui soit, à Semra et Florence pour votre gentillesse, que j'espère côtoyer davantage dans le futur. Merci également à Allison, Océane, Carlène et Emma d'avoir partagé ces moments agréables des derniers mois et années.

Je dois remercier tout particulièrement Catherine pour son temps, sa patience et son aide dans la révision de la thèse. J'ai senti que je parcourais les derniers miles de cette aventure en compagnie d'une amie. Merci à Vincent Perreault pour la relecture de la partie plus « technique » sur l'IA. J'aimerais de nouveau remercier Jean-François Godbout, qui a été présent pour éclaircir

mes doutes et me prodiguer des conseils au fil de mon parcours doctoral, ainsi que François Charbonneau, que je considère toujours comme mon premier mentor dans le monde académique. Merci aux créateurs d'Optimal Work, qui a constitué un instrument de développement personnel insoupçonné. Enfin, merci à la Fondation Robert-Bourassa, au GRIPP et au fonds de la FESP pour le soutien financier lors de mes études.

Introduction

Lorsque le célèbre critique culturel Neil Postman écrivait que « les conséquences de l'évolution technologique sont toujours vastes, souvent imprévisibles et largement irréversibles » [Traduction libre] (1998, 4), il ne songeait probablement pas à l'intelligence artificielle. Pourtant, il est dorénavant difficile de penser à quelque essor technologique que ce soit sans en tenir compte. Au tournant des années 2010, ce que nous connaissons comme l'« IA » a pris une ampleur considérable dans le domaine de la recherche comme dans la sphère publique. En tant que discipline, l'intelligence artificielle a émergé de la rencontre de la philosophie, de la linguistique, du génie informatique, des mathématiques et de l'économie, de la cybernétique, ainsi que de la psychologie et de la neuroscience (Russell et Norvig 2010, 5-16).[1] C'est précisément cette interdisciplinarité qui fait son succès (Nilsson 2010, 657). Définie et redéfinie de multiples manières, l'IA est, selon le professeur de droit Ryan Calo, « mieux comprise comme un ensemble de techniques visant à approcher un aspect de la cognition humaine ou animale en utilisant des machines » [Traduction libre] (2017, 4). Plus précisément, elle consiste en « […] un vaste domaine d'étude consacré à octroyer aux ordinateurs la capacité de faire des choses intelligentes » [Traduction libre] (Gerrish 2018, 6).

Le développement de ce champ d'investigation est évidemment animé par des finalités bien concrètes. La professeure de sciences cognitives Margaret A. Boden en relève deux : l'une de nature technologique, qui renvoie à l'utilisation des ordinateurs pour une utilité définie, et l'autre de la nature scientifique, visant à contribuer à la connaissance sur l'intelligence de tous les êtres (2018, 2). Il m'apparaît important de fournir ici un très rapide (et forcément incomplet) survol de l'historique de l'intelligence artificielle comme discipline, question de fournir quelques éléments clés à l'horizon de compréhension du lecteur.[2] Il sera ensuite temps de situer la question de l'IA dans le champ de la pensée politique, pour ensuite donner une idée de ce que j'aimerais considérer comme une contribution de la thèse, au plan académique.

---

[1] Une édition plus récente (2020) de l'ouvrage vient de paraître, mais je n'ai malheureusement pas pu y avoir accès par les services de la bibliothèque avant la fin de cette thèse.

[2] Les quelques lignes présentées ici traitent surtout du développement de l'IA en Occident; pour un regard plus complet, qui inclut l'apport du Japon, par exemple, on pourra se référer à l'exhaustif ouvrage de Nilsson (2010), qui retrace la chronologie de l'IA.

1. Bref historique de la discipline de l'intelligence artificielle

Même si l'humanité rêve de « machines » ou d'automates semblables aux humains, et ce, depuis l'Antiquité (Nilsson 2010, 19), le terme d'« intelligence artificielle » a d'abord été employé aux États-Unis, au milieu des années 1950, pour désigner ce qui s'appelait, auparavant, la « simulation informatique » (*computer simulation*) (McCarthy, Minsky, Rochester et Shannon 2006 [1955]). John McCarthy est ainsi considéré par certains comme le « père de l'IA » (Childs 2011, s.p.). Ce dernier, en compagnie de son équipe de dix chercheurs, s'est cloisonné, lors de l'été 1956, au Collège Dartmouth, au New Hampshire, pour élucider

> […] comment faire en sorte que les machines utilisent le langage, façonnent des concepts [dans] l'abstraction, résolvent des types de problèmes actuellement réservés aux humains et parviennent à s'améliorer. [Traduction libre] (McCarthy et al. dans Russell et Norvig 2010, 17)

À Dartmouth, l'équipe rencontre Allen Newell et Herbert Simon, qui sont à l'origine du GPS (*General Problem Solver*). C'est la réunion des « grands » qui révolutionneront la recherche en IA (Russell et Norvig 2010, 3, 18).

Quelque dix années avant cela, l'intelligence artificielle était déjà « en gestation » (Russell et Norvig 2010, 16). On peut même remonter à un siècle avant cette date pour retracer les origines de l'IA. La mathématicienne Ada Lovelace travaillait à l'élaboration d'une machine conçue par Charles Babbage en 1834, qui allait devenir l'ancêtre de l'ordinateur digital. Puis, au milieu du 20^{e} siècle, le mathématicien Alan Turing développe la « machine de Turing universelle » (*Universal Turing Machine*), un système imaginaire fonctionnant selon une logique binaire. Ce dernier a ensuite proposé le fameux jeu de l'imitation ou « Test de Turing », qui permettrait de déterminer si une machine pouvait simuler l'intelligence humaine au point de confondre un agent humain (Turing 1950). La combinaison de ses travaux avec la logique propositionnelle de Bertrand Russell, elle aussi binaire, ainsi que la théorie des synapses neuronales octroie à cette époque un élan au développement des recherches en IA (Boden 2018, 6-8).

Une sorte de schisme vient néanmoins fracturer le champ de l'IA dans les années 1960, séparant les chercheurs adonnés au calcul symbolique et ceux centrés sur la cybernétique. Les chercheurs en IA symbolique ont la mainmise sur la discipline, si l'on veut, jusqu'au milieu des années 1980, lorsque commencent à percer les réseaux profonds et la méthode du traitement parallèle (*parallel distribution processing*) (Boden 2018, 15-17). Malgré cela, plusieurs années devront s'écouler avant que l'on reconnaisse la performance de ces réseaux neuronaux. En effet, dans les années 1970 et 1980, l'IA fonctionnait avec des schémas de représentation du monde et du savoir (Russell et Norvig 2010, 22-24). D'une certaine façon, l'IA a d'abord été appréhendée de manière atomiste (avec des informations à programmer), avant d'être traitée de façon plus holiste (avec des réseaux), si l'on désirait parler en termes méréologiques.

À la fin des années 1980, le financement de la recherche en IA est en baisse. On parle d'un « hiver de l'IA » (Calo 2017, 4). On sait qu'actuellement — depuis 2006 environ — ce sont justement les réseaux de neurones artificiels qui ont mené aux avancées les plus patentes en intelligence artificielle, notamment dans la reconnaissance d'images et du langage (*natural language processing* [NLP]) (Gerrish 2018, 125). Également, plusieurs méthodes autres que celles de l'apprentissage machine sont employées à l'heure actuelle dans la recherche en IA (Calo 2017, 4-5). Malgré cela, il est vrai que les méthodes symboliques ont été abandonnées, au profit des démarches statistiques de l'apprentissage machine. Les méthodes symboliques portent un certain opprobre, venant du fait qu'elles n'aient pas fourni les résultats escomptés. Néanmoins, Stuart J. Russell et Peter Norvig affirment que les approches symboliques et connexionnistes ne sont pas rivales, mais plus efficaces lorsqu'agencées pour en tirer le potentiel complémentaire (2010, 25).

Distincts de l'IA symbolique (qu'on a appelée « *Good old fashioned » AI* [GOFAI]) (Haugeland 1989, 112), les réseaux de neurones artificiels (ou connexionnisme) sont souvent employés dans l'approche de l'apprentissage profond (*deep learning*). Concrètement,

> l'apprentissage profond […] consiste en de multiples couches en cascade, modelées sur le système nerveux humain […]. Les architectures d'apprentissage profond permettent à un système informatique de s'entraîner en utilisant des données […] en reconnaissant des modèles et en faisant des inférences probabilistes. [Traduction libre] (Malli, Jacobs et Villeneuve 2018, 4)

Cet apprentissage peut être supervisé ou non, semi-supervisé ou renforcé (*Ibid*, 4). C'est précisément l'apprentissage profond qui permet le fonctionnement de technologies présentement répandues comme les assistants vocaux (Siri, Alexa) ou encore les « maisons connectées » (pensons à Google Home) (*Ibid*., 6).

L'augmentation de la puissance de calcul des ordinateurs, depuis quelques décennies, ainsi que la disponibilité de données très abondantes (glanées sur Internet entre autres), a permis des avancées importantes en analyse prédictive. Ces données, que l'on appelle fréquemment données volumineuses, données massives, mégadonnées ou gigadonnées (*big data*) constituent

> […] des ensembles de données qui peuvent être recherchés, agrégés et triangulés avec d'autres ensembles de données. […] Ces mégadonnées prennent la forme d'artefacts de communication, tels que des photographies, le microciblage de profils, le contenu des réseaux sociaux et les métadonnées. [Traduction libre] (Shorey et Howard 2016, 5033)

Des percées dans les domaines de la sécurité, de l'éducation, de la justice pénale, de la santé, de l'utilisation des ressources énergétiques, de l'emploi et plus largement dans le monde des affaires ont été rendues possibles (Malli, Jacobs et Villeneuve 2018, 6; Reisman, Schultz, Crawford et Whittaker 2018, 3). Ce qui génère l'enthousiasme et les investissements massifs en recherche et en développement de l'IA est que la précision et la rapidité des systèmes employant l'intelligence artificielle (SIA) sont supérieures à celles de l'humain dans certains domaines (Gerrish 2018, 125). Tous ces SIA sont des exemples d'IA « étroite » ou « faible », contrairement à une IA qui serait « forte » ou « générale » (Searle 1980). Autrement dit, l'intelligence artificielle sert ici des tâches concrètes qui ne peuvent pas être généralisées et qui sont souvent répétitives (Malli, Jacobs et Villeneuve 2018, 4). Elles fonctionnent à l'aide des algorithmes qui sont, pour le dire simplement, des « […] séries d'étapes utilisées pour accomplir une tâche » (Shorey et Howard 2016, 5033).

Je traiterai de la question de l'intelligence artificielle générale (ou généralisée) dans le premier chapitre, dans le contexte de la revue de la littérature. Cependant, force est d'admettre que les travaux en IA tendent aujourd'hui à frapper l'imaginaire, lequel se dirige très facilement vers des scénarios dystopiques dans lesquels robots et *cyborgs* sont les protagonistes. Après tout, le mot « robot » a une consonance d'origine négative, puisqu'il désignait, dans la pièce de théâtre de Karel Capek, en 1920, des travailleurs humains sans âme qui s'emparaient des emplois des autres

(Leben 2018, 1). Pour en rester aux faits, c'est l'IA non généralisée, l'IA faible, sur laquelle il est pressant de se pencher. Le chercheur en IA Nils J. Nilsson résumait, en 2010, les avancées de l'IA en quatre catégories, soit 1) des systèmes IA complets (DeepBlue, une voiture autonome, système de reconnaissance vocale, par exemple), 2) des « architectures » (ou modèles), 3) des « processus » (comme certains algorithmes et systèmes de planification) et 4) des représentations (incluant les réseaux de neurones, des vecteurs, ou encore des réseaux sémantiques) (2010, 656-657). On s'éloigne ici, du moins au premier coup d'œil, d'agents autonomes maléfiques visant à prendre les rênes du monde.

Si l'IA peut fonctionner à l'intérieur de « machines physiques », elle œuvre plus généralement dans des systèmes virtuels (Boden 2018, 3). De même, il existe des robots mécaniques qui ne fonctionnent pas grâce à l'intelligence artificielle. Les deux réalités ne sont pas interdépendantes, malgré l'imaginaire populaire. Dans la thèse, je ferai fréquemment référence à des systèmes employant l'intelligence artificielle (SIA), soit « […] tout système informatique utilisant des algorithmes d'intelligence artificielle, que ce soit un logiciel, un objet connecté ou un robot » (Comité d'élaboration de la Déclaration de Montréal IA Responsable 2018a, 20).

2. L'intelligence artificielle, l'éthique et la politique

D'aucuns pourraient se demander quel lien l'IA entretient avec la politique et, plus précisément, l'étude de cette dernière. Plus encore, comme il s'agit d'une thèse de pensée politique, il conviendrait de mettre au jour la relation qu'entretient un domaine scientifique avec une réflexion de nature philosophique sur le bien commun. Cette thèse porte sur l'éthique de l'intelligence artificielle, et se veut un travail de réflexion à l'intention des décideurs politiques. Cette réflexion se fait également dans le contexte où je conçois la politique comme essentiellement dialogique face au conflit (Crick 1993,18). De plus, il s'agit d'un dialogue qui vise le bien commun, à savoir, la vérité articulée par des interprétations.

a) Décideurs politiques et communauté politique

Les « décideurs politiques » auxquels je ferai référence, tout au long de la thèse, sont entendus comme les représentants élus qui parlent au nom d'une « communauté politique ». Par cette définition même, mon propos normatif ne s'adresse qu'aux décideurs au sein de régimes démocratiques. Il s'agit d'une restriction conceptuelle que j'assume entièrement et que je juge pertinente pour les recommandations que je mettrai de l'avant à la fin de la thèse, puisqu'elles impliquent le dialogue. Ces décideurs ou élus politiques forment ainsi un groupe de citoyens, appartenant à un État, une province, voire une municipalité, un syndicat étudiant, etc., qui prennent des décisions susceptibles d'avoir une incidence concrète sur le bien commun.

Par exemple, dans le cas canadien, on pourrait penser aux députés de la Chambre des communes à Ottawa, sans oublier les autres acteurs avec lesquels ils entreront en dialogue (des sénateurs, des professionnels de la fonction publique, des experts techniques, des citoyens ou d'autres intervenants d'horizons distincts). Leurs décisions auront une portée non négligeable sur l'éthique de l'IA qui sera privilégiée dans la communauté donnée. Elles font partie de ce qu'on appelle la « politique de l'IA » (*AI Policy*) et donc, de manière encore plus vaste, de la « politique technologique ». La politique technologique, quant à elle, renvoie « […] à un ensemble de décisions que les sociétés, par l'entremise de leurs gouvernements, prennent sur ce qu'elles veulent et ne veulent pas permettre, sur ce qu'elles désirent ou ne désirent pas encourager. » [Traduction libre] (Brundage et Bryson 2016, 2). Ainsi, la thèse ne se veut pas une proposition de réforme du Parlement canadien pour repenser le dialogue, mais une proposition aux élus (du Parlement, par exemple) pour faciliter leur dialogue sur les politiques technologiques à élaborer et ce, à l'intérieur des institutions existantes. Évidemment, cette proposition concerne leur dialogue face à l'avènement de l'intelligence artificielle.

b) Éthique et métaéthique

Il sera clair, au fil de la thèse, que le propos ne porte pas sur des enjeux précis à l'intérieur du champ de l'IA, ou comme une forme d'éthique appliquée, dont je fournirai plusieurs exemples

au premier chapitre. C'est plutôt sur la compréhension même de l'éthique et par conséquent sur la métaéthique que je me penche, puisque je considère qu'elles ont une incidence réelle sur la manière de dialoguer sur les défis concrets que l'on trouve dans le champ de l'éthique de l'IA. Comme le dit Charles Ess, la métaéthique consiste en « […] nos manières de second niveau de conceptualiser et d'analyser les questions éthiques de premier niveau » [Traduction libre] (2020, 554). Et plus précisément, selon Geoff Sayre-McCord,

> la métaéthique est la tentative de comprendre les présuppositions et les engagements métaphysiques, épistémologiques, sémantiques et psychologiques de la pensée, de la parole et de la pratique morales. [Traduction libre] (2014, §1)

Martin Gibert (2020a), quant à lui, voit dans la métaéthique une avenue permettant de comparer les approches éthiques entre elles, ce que je ferai au fil des chapitres.

Par ailleurs, le lecteur remarquera que j'emploie le terme « éthique » plutôt que celui de « moralité ». Mis à part leurs origines différentes (grecque et latine), je ne me prononcerai pas sur le débat sur la différence entre l'éthique et la moralité, puisque là n'est pas l'essence de mon questionnement. Je comprends l'éthique comme une réflexion sur le caractère (bon ou mauvais) d'actions et de décisions, de même que sur ce qu'il est bon d'*être* (Ricœur 1996; Williams 2011). Cela étant dit, c'est l'éthique politique qui m'intéresse ici, et j'étofferai le contenu de cette compréhension de l'éthique au chapitre 6, en présentant mon propre ancrage métaéthique, ainsi que, dans le chapitre 7, mes propositions aux décideurs politiques. Pour le moment, il suffit de comprendre l'éthique politique comme touchant aux individus « réunis en communauté politique », qui « […] adopte une forme ou une autre par l'entremise de ses représentants élus » [Traduction libre] (Luño s.d.a, 1-2).

c) Politique et pensée politique

Peut-être pourrait-on penser que l'éthique (et surtout la métaéthique) est l'apanage des philosophes, qui eux ne se trouvent que dans les départements de philosophie. Or, l'éthique peut — et devrait — être traitée également en science politique comme dans presque tous les domaines du savoir. En effet, comme le dit Aristote dans son *Éthique*, la science politique est la science

maîtresse (2014, Livre I, 1, 1094a25). L'éthique a sa pertinence en pensée politique non seulement « parce qu'elle suppose quelque chose sur la nature de l'autorité […] » [Traduction libre] (Tronto 2011, 408), mais aussi parce que l'éthique et la politique « […] sont comme les branches étroitement imbriquées d'une tige commune » [Traduction libre] (Garner 1907, 2010). En outre, la politique est un atout pour l'éthique de l'IA puisque, selon Calo, l'éthique, prise seule, peut être édulcorée et manquer de « mordant » dans son application (2017, 6-7). L'éthique et la politique doivent se parler pour entourer le développement de l'IA et elles ont le potentiel de le faire de manière efficace (ÓhÉigeartaigh et al. 2020, 2). La pensée politique a tout intérêt à se tourner vers l'avenir pour découvrir de nouveaux objets d'étude, en faisant appel à sa capacité de « voir la forêt derrière les arbres » (Susskind 2018, 10-11). Plus encore,

> […] pour plusieurs raisons, la théorie [qui est une forme de pensée] politique est bien adaptée à l'examen de l'interaction entre la technologie et la politique. […] Le canon de la pensée politique contient une sagesse qui a survécu aux civilisations. Il peut nous éclairer sur nos difficultés futures et nous aider à identifier les enjeux. [Traduction libre] (Susskind 2018, 9)

Bref, la pensée politique — son nom l'indique — outille à penser et à parler du politique (Susskind 2018, 10) et, dans ce cas-ci, de son rôle face au développement et à l'emploi des SIA.

Si l'on tient pour acquis que « […] l'IA est déjà suffisamment mature technologiquement pour avoir une incidence sur des milliards de vies des milliards de fois par jour » [Traduction libre] (Brundage et Bryson 2016, 1; voir aussi Floridi 2019, 185), il semble naturel, pour ne pas dire indispensable, que les décideurs politiques se penchent sur son encadrement éthique. La croissance économique, la distribution des ressources, les retombées positives et négatives sur des franges de la population (comme la manipulation du comportement) (Harari 2018, §9; Schiff et al. 2020, 153; Boddington 2017, 60; Müller 2020, §2.2) sont des enjeux d'importance capitale dont je traiterai au chapitre suivant. L'ampleur des changements générés par les percées technologiques est telle que la décrivait Postman d'entrée de jeu. Ces transformations ont une incidence sur l'usage de notre esprit et de notre corps, ainsi que sur notre perception du monde, ajoute-t-il (1998, 3). L'intelligence artificielle aurait en somme la capacité de

> […] perturber notre réflexion sur nous-mêmes, nos natures, nos capacités et notre place dans la société et dans le monde d'une manière qui pourrait causer un bouleversement

> majeur de la pensée éthique sous-jacente aux codes d'éthique. [Traduction libre] (Boddington 2017, 61)

Pour cette raison, il va sans dire que les questions éthiques de l'IA ne peuvent être abandonnées au secteur privé, aux développeurs et aux concepteurs de SIA, pas plus qu'à seulement ceux qui les mettent en marché ou en tirent un profit. Luciano Floridi et ses collaborateurs affirment qu'un tel geste serait « inacceptable en raison d'un déficit de responsabilité sociale et politique » [Traduction libre] (2018b, 507). Mark Coeckelbergh reprend cette idée en parlant d'« irresponsabilité » (2020a, 176-177). Dans un rapport concernant le développement responsable de l'IA, le géant du Web Google explique que

> bien que les chercheurs en IA puissent jeter les bases de ce qui est techniquement faisable, la portée de l'IA dans la pratique dépendra de l'appétit exprimé par l'industrie et la société, ainsi que des directives et des limites fixées pour son application par le gouvernement. *Les décideurs sont ainsi essentiels à l'élaboration de la vision et à l'établissement des cadres qui sous-tendront le développement de l'IA*. [Traduction libre, je souligne] (2019b, 6)

Microsoft abonde dans le même sens, en assurant qu'une IA centrée sur les intérêts des humains a tout à gagner en misant sur l'inclusion des élus politiques (2018a, 138). Dans son guide « non technique » sur la nature et les défis éthiques de l'IA à l'intention des décideurs politiques, Branka Panic souligne que « seule une politique réfléchie et responsable nous permettra d'intégrer avec succès l'IA dans notre société » [Traduction libre] (2020, n^{o} 24). Le chercheur au Oxford Internet Institute Brent Mittelstadt a cadré la situation succinctement :

> *l'éthique de l'IA est en fait un microcosme des défis politiques et éthiques auxquels la société est confrontée*. Il est insensé de penser que des questions normatives très anciennes et complexes peuvent être résolues par des solutions techniques ou une « bonne » conception uniquement. […] L'éthique n'est pas censée être facile ou basée sur des formules. *Il faut s'attendre à des désaccords de principe irréductibles et les accueillir favorablement*, car ils reflètent à la fois une considération éthique sérieuse et une diversité de pensée. […] L'éthique est un processus, pas une destination. [Traduction libre, je souligne] (2019, 11)

Je souscris à cette façon de voir exposée par Mittelstadt, qui donne déjà à penser qu'une certaine irréductibilité sera à l'œuvre dans la réflexion éthique. On peut y déceler un écho à la notion de « tension » qui est centrale au travail que je présente ici. Plus encore, il m'apparaît que l'approche éthique choisie pour affronter les défis que pose l'IA aura une incidence sur le dialogue des

décideurs politiques. Il s'agit d'une autre raison illustrant l'importance de documenter la relation entre l'éthique et la politique en ce qui a trait à l'intelligence artificielle. J'exposerai, dans la thèse, la nature et les conséquences potentielles de ce lien.

3. Plan de la thèse

Trois sections de deux chapitres chacune divisent la thèse, qui débutent dès le deuxième chapitre. Le premier chapitre est essentiellement dédié à la revue de la littérature en éthique de l'intelligence artificielle. Il se termine sur la formulation des hypothèses et des questions de recherche qui guident le propos. Ainsi, le corps de l'ouvrage est formé de sept chapitres. La première section, « Fondements », est spécialement consacrée à la métaéthique. Dans les chapitres deux et trois, j'explore avec le lecteur les bases de trois approches éthiques monistes en Occident, soit l'utilitarisme, l'éthique déontologique et l'éthique de la vertu, puis les fondements du pluralisme des valeurs (tel que compris, surtout, par Isaiah Berlin et Bernard Williams).

La deuxième section, « Portrait », se voue à l'analyse d'un échantillon de vingt directives éthiques publiées en 2016 et 2020, émanant des entreprises privées, de la société civile et de différents organismes internationaux. Ces directives sont essentiellement des prises de position, chartes, réflexions et rapports concernant l'éthique de l'IA. Je propose une lecture de ces documents guidée par les catégories métaéthiques approfondies dans la section précédente. En effet, on a assisté à une véritable « prolifération » de telles directives dans les dernières années, qui ont entraîné certaines formes d'« incohérence et de confusion » quand vient le temps de les évaluer (Floridi 2019, 186). Le chapitre quatre recense les démarches « monistes » et le chapitre cinq, les « pluralistes », ainsi que celles qui exhibent une certaine tension métaéthique.

Ce segment de la thèse se veut une sorte de « portrait » de l'éthique de l'IA contemporaine dans ses ancrages métaéthiques. La notion de tension, qui y est mise de l'avant et abondamment illustrée, est compréhensible grâce à la première section, et « dépassable », à mon sens, par la troisième et ultime section. Cette démarche permet d'illustrer les dynamiques à l'œuvre dans les réflexions contemporaines sur l'éthique de l'IA, et ce, en provenance de plusieurs secteurs de la

société. Elle consiste en une avenue de réponse à la confusion possible des décideurs face à cette panoplie d'options éthiques qui proposent de « réglementer » l'IA. La notion de réglementation, voire celle de l'encadrement, peuvent sous-entendre une forme de contrôle. Elles peuvent aussi être comprises comme une manière plus algorithmique que dialogique d'aborder un problème. Je leur préférerai une avenue impliquant le dialogue, dans la troisième et dernière section.

Cette dernière section contient les chapitres six et sept. Elle consiste en des « Propositions ». D'abord, au sixième chapitre, j'amène une critique des courants éthiques dominants en IA actuellement, pour proposer une approche alternative. Cette dernière forme le socle sur lequel j'élabore ensuite une sorte de soumission à l'intention des décideurs politiques, au chapitre sept. Cette proposition se veut un guide pour leurs dialogues sur l'éthique de l'IA, informé par mes réflexions métaéthiques ainsi que par la philosophie herméneutique, qui ouvre la voie à une conversation optimale entre ces décideurs, dans le but d'atténuer la tension que je relève en éthique de l'IA. Elle combine un but commun avec le meilleur respect possible de la pluralité des points de vue qui s'exprimeront dans la pratique du dialogue. La thèse se conclut avec un document et un illustré de cette proposition. Ainsi, on aura compris qu'il s'agit ici d'une thèse certes descriptive, mais surtout éminemment normative.

Chapitre 1 — Revue de littérature

« […] *l'IA a réussi à capter l'imagination des décideurs politiques suffisamment tôt dans son cycle de vie pour qu'il y ait un espoir que nous puissions encore la canaliser vers l'intérêt public*. » — Ryan Calo [Traduction libre] (2017, 28)

Introduction

La réflexion éthique sur l'avènement de différentes technologies est loin d'être chose nouvelle. Remontant à Platon, pour qui la technologie « imite la nature », et qui propose la figure du démiurge comme artisan, pour ensuite s'inscrire dans la « philosophie de la technologie » après la Seconde Guerre mondiale, on la trouve dans les travaux de Martin Heidegger, Jacques Ellul, Herbert Marcuse, Lewis Mumford, Hans Jonas et Albert Borgmann (Platon s.d.; Franssen, Lokhorst et van de Poel 2018, §1.1; Vallor 2016, 28), auxquels j'ajouterais ceux d'Ivan Illich et Neil Postman. Dans le contexte un peu plus récent de la pensée éthique sur les médias numériques, Charles Ess voit dans la réflexion éthique une occasion de développement double pour les êtres humains, à savoir la compréhension de nos propres schèmes de moralité, ainsi que ceux des autres n'appartenant pas nécessairement à la même culture (2009, xiii). Le défi est que cet exercice se fait dans le contexte où les technologies émergentes présentent de nouvelles questions éthiques (Ess 2009, 169). L'éthique de la technologie, dont découle l'éthique de l'intelligence artificielle (Gibert 2020, s.p.) est elle-même le produit d'une myriade d'influences, incluant des codes institutionnalisés, des cultures professionnelles, des capacités technologiques, des pratiques sociales et de la prise de décisions individuelles (Ananny 2016, 96).

L'éthique de l'intelligence artificielle est un domaine de recherche pluridisciplinaire en pleine expansion. Les chercheuses Carina Prunkl, du Future of Humanity Institute à Oxford et Jess Whitlestone, du Leverhulme Centre for the Future of Intelligence à Cambridge, soutiennent même qu'une nouvelle communauté de recherche est en pleine émergence. Elle fait appel à des disciplines aussi diverses que la philosophie, l'informatique, la science politique, la sociologie, le droit, l'économie et les relations internationales (Prunkl et Whittlestone 2020, 1). Le philosophe et

éthicien à l'Institut de valorisation des données Martin Gibert (2019) abonde dans le même sens en ce qui a trait la pluridisciplinarité. Cette dernière est, au fond, souhaitable, puisque les enjeux éthiques qui sont soulevés par ces nouvelles technologies ne peuvent être entièrement appréhendés par une seule lunette. C'est au sein de cette communauté de recherche que cette thèse s'inscrit et désire apporter quelques nouvelles réflexions.

Gibert (2019) conçoit l'éthique appliquée à l'intelligence artificielle comme « […] le domaine de l'éthique qui se demande ce qui est bon, juste ou vertueux dans la mise en œuvre des systèmes d'intelligence artificielle », domaine qu'il appelle « éthique artificielle ». Dans cette thèse, je préfère parler « d'éthique de l'intelligence artificielle »[3] plutôt que d'« éthique artificielle », pour la simple raison que cette dernière expression peut à mon sens porter à confusion, à savoir suggérer que l'éthique elle-même soit artificielle. L'éthique de l'intelligence artificielle, dans la perspective de philosophie politique que j'ai adoptée, consiste en une une réflexion éthique adressée aux décideurs politiques, dans le but de les outiller à porter une attention particulière aux répercussions sociétales (soit dans « la vie sociale et politique » [Maclure et Saint-Pierre 2018, 746]) que posent les SIA, au moyen du dialogue entre eux. Une réflexion de la sorte est cruciale dans le contexte où actuellement, « ce qui est numérique est politique », même « hyper-politique » [Traduction libre] (Susskind 2018, 6, 13, 16; voir également Gibert 2020, pour qui les questions morales de l'IA sont des questions politiques).

La préoccupation pour les répercussions sociétales n'émane toutefois pas que des sciences sociales. Plusieurs chercheurs et développeurs en intelligence artificielle se préoccupent des retombées de ces technologies sur la société (Bengio 2017). Néanmoins, des travaux portant sur la politique de l'IA abondent depuis environ le milieu des années 2010 (Dutton 2018a). Bien sûr, les axes de recherche touchant à l'éthique et à la politique se chevauchent parfois, puisque « […] les gouvernements prennent désormais conscience de l'effet perturbateur de l'IA et veulent prendre les devants », notamment pour des raisons de concurrence dans la recherche et l'économie (Dutton 2018a).

[3] On devrait même parler d'« éthique *pour* l'intelligence artificielle », ce qui se lirait peut-être mieux en français. Toutefois, cette expression peut renvoyer à la sous-catégorie plus précise de l'éthique des machines, alors que l'éthique sur laquelle je réfléchis s'adresse non seulement ni principalement aux SIA, mais aux décideurs politiques. C'est la raison pour laquelle je parle d'éthique de l'intelligence artificielle à l'intention des décideurs politiques.

Ce chevauchement entre les préoccupations éthiques et les différentes disciplines qui sont invoquées dans ces réflexions fait que l'éthique de l'IA est un domaine, somme toute, un peu flou. Ses contours ne sont pas encore tracés clairement, en ce que des problèmes très distincts sont traités au sein de ce champ. Gibert (2019) propose de sectionner en deux les types de problèmes que l'on trouve en éthique de l'IA : 1) des problèmes épistémiques, comme celui de l'opacité des algorithmes et 2) les problèmes éthiques, qui mettent en relation un agent et un patient moraux. De mon côté, j'observe que quatre types d'éthique se profilent dans ces démarches : 1) celle des principes programmés dans les machines ou SIA (« l'éthique des machines » [en anglais « *machine ethics* »), 2) celle des concepteurs et développeurs (comme un code d'éthique professionnel, tel qu'analysé, entre autres, par Boddington [2017, 40–47] et par McNamara, Smith et Murphy-Hill [2018, 3]), 3) celle des utilisateurs de SIA (qui renvoie à l'éthique personnelle ou aux lois d'un État donné) puis 4) celle pour les décideurs politiques qui encadrent la conception, le déploiement et l'utilisation de telles technologies. En éthique de l'IA, on retrouve souvent ces quatre groupes de destinataires quelque peu entremêlés dans les documents. Cela n'est pas surprenant si l'on pense que ces quatre catégories sont difficilement séparables dans l'absolu. Quiconque s'intéresse à l'aspect politique de l'éthique de l'intelligence artificielle s'adressera logiquement à ces quatre catégories de destinataires, sans pour autant s'ancrer dans « l'éthique des machines » en tant que telle, ou encore à l'éthique personnelle des individus employant des SIA dans leur vie quotidienne, par exemple.

1. Une approche thématique à la littérature

Le passage en revue suivant vise à donner au lecteur une idée des thèmes traités en éthique de l'IA du côté du monde académique principalement. Les documents en éthique de l'IA émanant des cercles gouvernementaux, des entreprises privées ou de la société civile seront explorés dans la deuxième section de cette thèse, aux chapitres quatre et cinq. Des démarches en éthique de l'IA telles que celles du IEEE (Institute of Electronics and Electrical Engineers, Incorporated) chevauchent toutefois le monde académique et ces autres cercles, et rendent sa catégorisation plus ardue. À ce sujet, l'étude de 2019 du IEEE, qui énonce des principes et des recommandations concernant l'alignement des valeurs éthiques dans le développement de l'IA, dans l'optique d'un

monde technologique en mouvance (2019, 2, 4, 171), représente la synthèse des études de 2016 et 2017. J'analyserai ces deux études avec davantage de précision dans la deuxième section de la thèse. Il faut néanmoins noter que l'esprit avec lequel le dernier rapport du IEEE est réalisé est celui du dialogue, par exemple avec des traditions éthiques non occidentales (2019, 3, 36-49), ainsi que de la sensibilité à la fois aux contextes changeants, mais également aux droits humains[4] (2019, 171). Je chercherai, par l'entremise de telles analyses à venir, à exposer au lecteur des exemples de cohabitations surprenantes d'approches éthiques au sein d'une même directive.

Cela étant dit, le regard de ce chapitre se veut porté vers la littérature académique en éthique de l'IA. Je m'appuie pour ce faire sur la catégorisation faite par l'éthicien Vincent C. Müller qui a recensé le champ de l'éthique de l'IA en dix sections (2020, §2). Je reprends cette même structure de sections (avec un contenu différent à plusieurs égards), car elles me semblent bien dépeindre les thèmes qui sont présents dans cette littérature. Cependant, j'y ajoute quelques précisions ou explicitations inspirées de la science politique. Étant donné que ce champ d'études est très large et en développement exponentiel, cela le rend difficile à circonscrire ou à en capturer un instantané.

Ainsi, j'ai uni les enjeux de sécurité internationale à ceux des systèmes autonomes. J'ai par ailleurs fusionné les catégories d'agents moraux artificiels et celle de la « singularité », et traiterai des sources portant sur l'intelligence artificielle générale dans cette section. J'ai élargi la catégorie de la manipulation du comportement pour y ajouter les effets de l'IA sur les régimes politiques, notamment la démocratie. À l'opacité des systèmes, j'ai ajouté l'enjeu de l'attribution de la responsabilité, et aux interactions entre humains et robots (je parlerai de « machines »), j'ai greffé celui de la « société technologique ». La catégorie sur l'automatisation et l'emploi est aussi élargie pour inclure l'effet plus général sur l'économie. Enfin, je propose une dernière catégorie (ce qui m'en donne une dizaine comme Müller), à savoir, les compilations qui ont été faites des prises de position et de documents éthiques émanant de toutes sortes de sources (gouvernements, société civile, associations professionnelles, entreprises privées, etc.). Évidemment, ces catégories ne sont pas des compartiments étanches : elles se recoupent. Un même ouvrage ou article peut traiter de

[4] Dans la thèse, j'emploie consciemment l'anglicisme « droits humains » plutôt que « droits de l'Homme ». Cette même traduction est employée dans la version française du livre de Sandel, *Justice*, Albin Michel : 2016. Le calque de l'anglais désigne mieux, et de façon plus inclusive, ce que je cherche à désigner.

plusieurs thèmes à la fois et les sujets eux-mêmes sont parfois ardus à catégoriser. C'est pour la clarté que cet exercice est mené de cette manière.

Le lecteur doit être conscient que, malgré l'intérêt de tous les sujets qui seront mis de l'avant, deux considérations sont cruciales. Tout d'abord, cette revue de la littérature n'est pas exhaustive. L'interdisciplinarité du champ de l'éthique de l'IA, de même que la diversité des enjeux soulevés, rend impossible cette aspiration à l'intégralité. La grande actualité de ce domaine d'étude est un défi supplémentaire, en plus d'être une richesse : des quantités importantes de documents sont régulièrement publiées sur le sujet lors des dernières années, et ce, de manière constante. J'ai donc cherché à tracer un portrait aussi complet que possible en ce qui a trait aux thèmes et aux enjeux, tout en sachant que les auteurs mentionnés, bien que centraux, ne représentent pas l'entièreté du champ.

Ensuite, mon propos, dans la thèse, prendra une tout autre direction que celle des thèmes mentionnés dans la revue de littérature. C'est parce que je m'inscris dans les débats académiques touchant à l'éthique de l'IA qu'il m'apparaît primordial de documenter les enjeux suivants. Néanmoins, mon analyse ne se situera pas en « éthique appliquée » à un ou plusieurs de ces défis, mais plutôt, en amont, au niveau de l'analyse métaéthique. Autrement dit, je me pencherai sur les différentes approches éthiques normatives à l'œuvre, comme les théories conséquentialistes, déontologiques et l'éthique de la vertu. C'est dans l'optique où « […] un exposé de l'éthique reposera explicitement ou implicitement sur des conceptions sous-jacentes des agents moraux — nous — et de notre place dans le monde » (Boddington 2017, 11) que j'ai choisi de mener mes travaux de recherche en éthique de l'IA. C'est la raison pour laquelle la littérature éthique que j'explorerai avec davantage de profondeur au cours des prochains chapitres est la littérature que l'on pourrait appeler métaéthique, ou plus classique.

Par ailleurs, ce sont les propositions d'éthique appliquée dans la littérature académique, mais surtout dans les documents éthiques publiés par les entreprises privées, la société civile et des organisations internationales qui feront l'objet de mon analyse, à la lumière de mes recherches sur l'éthique. Évidemment, je suis convaincue qu'ultérieurement, une fois ce travail d'analyse

accompli dans cette thèse, il sera fructueux de se pencher sur les enjeux éthiques appliqués de plus en plus diversifiés qui sont posés par l'intelligence artificielle, dans des projets de recherche à venir.

a) Les systèmes autonomes et la sécurité internationale

Il est clair que le développement de l'intelligence artificielle, en pleine renaissance, ne bénéficie pas également à tous les pays du monde. Il pourrait même renforcer, voire consacrer, des dynamiques du Nord industriel riche et du « Sud global » en décalage technologique et économique, souvent exploité pour sa main-d'œuvre peu coûteuse (Shestakofsky 2017). On parlerait, entre les pays (mais même au sein d'une société donnée), de « fracture numérique » (Gibert 2019). La question de la sécurité des systèmes autonomes est probablement l'une des plus visibles et controversées en éthique de l'IA, puisque son contenu est fréquemment mis de l'avant à titre d'exemple de scénarios potentiellement dystopiques. Des systèmes autonomes pourraient piloter non seulement des voitures, mais des armes létales (Boulanin 2016 et Scharre 2016 dans Dutton 2018c; Calo 2017, 13; Metz 2019). En 2018, Google comptait des milliers d'employés récalcitrants à travailler sur le développement d'un programme du Pentagone (*Project Maven*) visant à améliorer des drones militaires (Shane et Wakabayashi 2018). Ces derniers ont signé une lettre de protestation dans laquelle ils écrivent :

> nous estimons que Google ne devrait pas être impliqué dans l'économie de la guerre. Par conséquent, nous demandons que le Projet Maven soit annulé et que Google élabore, publie et applique une politique claire stipulant que ni Google, ni ses sous-traitants, ne construiront jamais de technologies de guerre [Traduction libre] (« Letter to Google's C.E.O. » 2018).

Google s'est retiré du projet et a éventuellement mis au point une série de principes éthiques devant entourer le développement de l'IA (Metz 2019, s.p.; Google 2018). Toujours en 2018, et cela suivant l'initiative du Future of Life Institute (FLI) de Boston, des milliers de chercheurs en intelligence artificielle ont également signé une lettre d'engagement à bannir les « robots tueurs », soit les armes létales autonomes ou « LAWS » (« *lethal autonomous weapon systems* ») (Sample 2018).[5] Vingt-huit pays membres de l'Organisation des Nations Unies (ONU) se sont

[5] En août 2020, on comptait 247 organismes signataires ainsi que 3253 individus (Future of Life Institute s.d.).

engagés à renoncer à ces armes (Future of Life Institute s.d.b). À l'automne 2019, les pourparlers étaient ouverts à l'Organisation des Nations Unies, sans pour autant en arriver à une entente, contrairement à l'Union européenne (Brzozowski 2019; Conn 2018). Le sérieux de l'enjeu se voit au fait que ces armes létales autonomes ont été qualifiées de troisième révolution dans l'armement (dépassant les armes nucléaires) (Russell et al. 2015, 415). Un des arguments en leur faveur est que ces armes autonomes, en raison de leur plus grand degré de précision que les armes portées par des humains, sauveraient des vies humaines au bout du compte. C'est l'argumentaire du chercheur Matt Zeiler, fondateur de la compagnie new-yorkaise Clarifai, qui œuvrait sur le même projet du Pentagone que Google (Metz 2019, s.p.). Le souhait sous-jacent au développement d'armes létales autonomes est que si la guerre est inévitable, elle n'implique plus des humains, mais seulement des machines (Tegmark 2017, 110).

Le risque que des États faillis ou des groupes terroristes ne tombent en possession de telles armes est réel (Sample 2018), tout comme celui des puissances mondiales qui en auraient en leur possession, qui pourraient s'en servir pour des « attaques préventives » (Russell et al. 2015, 415-416). Les budgets qui sont alloués au développement de systèmes autonomes dans l'armée sont faramineux, et pourront déboucher sur un « capitalisme autonome », puisque ces systèmes seront faciles à produire et à reproduire (Omohundro 2014, 304-305). Plus encore, de tels systèmes autonomes « rationnels » exhibent des caractéristiques devant lesquelles nous met en garde l'informaticien Steve Omohundro : l'autopréservation, l'efficacité, l'acquisition de ressources et la reproductibilité, car elles peuvent déboucher sur des comportements dangereux (2014, 304).

Les systèmes autonomes ne comprennent pas que l'armement. D'autres systèmes en développement et en test font couler beaucoup d'encre. C'est le cas de la voiture autonome (ou « sans chauffeur » [Nurock 2019, 73]). Malgré la promesse qu'elle réduira le nombre d'accidents (Maclure et Saint-Pierre 2018, 745), plusieurs enjeux sont soulevés, le plus fréquent étant le dilemme du tramway (Foot 1978; Thomson 2003) que l'on applique au cas de la voiture autonome et pour lequel toutes sortes d'analyses ont été proposées (par exemple Bonnefon, Shariff et Rahwan 2016, dans Gibert 2020; Millar 2016; Etzioni et Etzioni 2016a, 150; Martin 2017; McAleer 2018; Leben 2018; De Luca-Baratta 2019; Elish 2020; Della Foresta 2020, Müller 2020), au point où la professeure de théorie politique Vanessa Nurock parle de « tramwaylogie » (2019,

69). Je ne ferai pas exception et je reviendrai sur ce type de dilemme éthique dans un chapitre ultérieur.

Parmi les analyses éthiques sur le sujet, celle de Nørskov et Rodogno suggère que l'automatisation des voitures pourrait mener à une banalisation de certaines tragédies (par exemple un accident causant la mort d'un passager ou d'un piéton, en raison d'une collision inévitable) ou encore à une dissolution de la responsabilité humaine face à une perte qui a été causée (s.d., 1, 11). Pour prévenir ce type de tragédie (que ce soit pour une voiture ou encore un avion), le professeur de droit Ryan Calo soutient que des standards de sécurité clairs doivent être mis de l'avant et respectés (2017, 14). La déresponsabilisation humaine est aussi un argument invoqué contre l'emploi de systèmes autonomes, comme des armes létales (Müller 2020, §2.7.2). Cette idée de la banalisation du mal par l'emploi de l'IA plus largement est également reprise par Coeckelbergh (2020a, 180). Bien que la tendance, dans la littérature, soit à l'invocation du dilemme du tramway, on trouve toutefois des voix critiques qui s'élèvent pour dire que l'accent mis sur ce type d'enjeu masque le contexte éthique plus global, ou encore les questions éthiques périphériques tout aussi importantes et profondes, qui requièrent que l'on s'éloigne de la tentation de la prétendue neutralité dans l'analyse éthique (Boddington 2017, 88; Nurock 2019, 72-73).

Les systèmes autonomes, en fin de compte, doivent prendre des décisions à caractère éthique qui ont des répercussions importantes sur les êtres humains (Millar 2016, 789). Qu'il soit question de robots « tueurs », de véhicules autonomes, d'assistants médicaux robotisés (« *carebots* ») ou de systèmes de domotique (Millar 2016, 787–789), des questions éthiques se posent quant aux répercussions des « choix » de ces systèmes sur les humains à l'échelle globale, ainsi que sur le maintien de la paix mondiale. Une autre série d'interrogations est soulevée par la prise de décision de systèmes autonomes : celles sur le type d'éthique qu'il faudra « programmer » dans de tels systèmes d'intelligence artificielle robotisés (c'est-à-dire possédant un « corps ») ou non. C'est l'épineuse question de l'éthique des machines qu'il convient à présent de traiter.

b) L'éthique des machines

Les réflexions sur l'éthique des machines (« *machine ethics* ») sont probablement les plus emblématiques du champ de l'éthique de l'IA. L'éthique des machines porte également d'autres noms dans la littérature. Par exemple, Gibert (2020) préfère l'appeler « éthique des algorithmes », puisqu'elle porte sur l'implantation — selon ses termes — des principes éthiques dans les SIA, plutôt que sur le comportement humain face à l'IA, qu'il considère plus globalement comme l'éthique de l'IA. Cette façon de voir diffère de celle d'Ananny pour qui les algorithmes en eux-mêmes sont moins importants que leurs assemblages, que les réseaux et catégories qu'ils forment par leur fonctionnement (2016, 97-99). Le professeur de génie informatique Roman V. Yampolskiy recense une panoplie de sous-domaines traitant d'éthique qui ont émergé dans le champ de l'informatique lors des dernières années : « […] éthique des machines, éthique informatique, éthique des robots, ethicALife, morale des machines, éthique des cyborgs, éthique informatique, robotique, droits des robots et morale artificielle […] » [Traduction libre] (2019, §1). Il s'agit d'une discipline qui se trouverait à la croisée de l'informatique et de la philosophie, donc également très interdisciplinaire (Leben 2018, 6), puisqu'il est nécessaire d'obtenir la coopération de plusieurs experts de domaines distincts pour assurer l'alignement entre les valeurs humaines et celles qui sont promues par les SIA.

Cela explique peut-être pourquoi l'éthique des machines, comme sous-champ de l'éthique de l'IA, est très large et difficile à baliser, à tel point que Müller soutient qu'

> […] il n'est pas certain qu'il existe une notion cohérente d'« éthique des machines », car les versions plus faibles risquent de réduire la notion d'« avoir une éthique » à des notions qui ne seraient normalement pas considérées comme suffisantes (par exemple, sans « réflexion » ou même sans « action »); les notions plus fortes qui vont vers des agents moraux artificiels peuvent décrire un ensemble — actuellement — vide [Traduction libre] (2020, §2.8).

Parallèlement, Virginia Digum souligne qu'une difficulté supplémentaire résiderait dans notre définition même de l'intelligence, qui serait, à son avis, trop centrée sur l'humain et trop étroite (2019, 10). Bref, si l'on veut parler d'« agents moraux artificiels », il faut s'entendre sur une définition de ce que l'on considère comme l'intelligence, et Dignum estime que notre démarche serait plus fructueuse en parlant de propriétés comme la proactivité, la sociabilité et la réactivité,

qui devraient être combinées (2019, 10). Au fond, notre compréhension de l'intelligence bénéficierait d'être élargie et plus dynamique.

Un autre point digne d'intérêt, en éthique des machines, est que si un robot est programmé selon des règles éthiques spécifiques, l'inverse peut également être vrai : une programmation peut être faite pour des règles qui ne seraient pas moralement acceptables (Müller 2020, §2.8). Toutes sortes de propositions de règles ou de programmes ont été lancées, parfois même inspirées par les lois d'Isaac Asimov sur la robotique, mises de l'avant dans sa nouvelle « Cercle vicieux », en 1942 (Hébert-Dolbec 2020). Les lois d'Asimov sont les suivantes :

> 1) « Un robot ne peut pas blesser un être humain ou, par inaction, permettre qu'un être humain se fasse infliger un mal », 2) « Un robot doit obéir aux ordres qui lui sont donnés par des êtres humains, sauf si ces ordres sont en conflit avec la première loi » et 3) « Un robot doit protéger sa propre existence tant que cette protection n'est pas en conflit avec la première ou la seconde loi » [Traduction libre] (Anderson M. 2017, §5)

Oren Etzioni, le directeur général du Allen Institute for Artificial Intelligence aux États-Unis, se sert de ces lois comme point de départ pour en élaborer trois autres devant encadrer le développement de l'intelligence machine, ainsi que leur régulation par les gouvernements et la sphère privée : ces lois portent sur la transparence des systèmes, mais aussi sur la responsabilité humaine à les gérer (Etzioni 2017, s.p.). L'idée d'un système IA qui serait responsable d'assurer l'éthique d'un autre a été mise de l'avant par quelques chercheurs, notamment Etzioni et Etzioni (2016a), dans ce qui s'appellerait un « *AI Guidance Program* » ou encore un « *AI Ethics bot* » (150, 152). C'est la question du choix des valeurs à programmer dans ces systèmes qui devient centrale. Ces derniers préconisent que dans de tels cas, les machines soient programmées moralement suivant une logique « communautarienne » plutôt que « libertarienne », de façon à ce que les valeurs des communautés locales soient respectées tout en n'alourdissant pas les systèmes avec des possibilités de choix trop personnalisés. Non seulement ces systèmes éthiques guideront d'autres SIA, mais ils pourraient même servir de guides éthiques aux êtres humains eux-mêmes (Etzioni et Etzioni 2016a, 150-152, 154).

D'autres suggestions sont mises de l'avant pour la programmation de SIA « éthiques ». Une approche qui gagne en popularité depuis quelques années est le « design éthique » (Millar 2016; Leikas, Gotcheva et Koivisto 2019; Millar et al. 2019). De telles études sont souvent menées au

moyen de scénarios prospectifs, cherchant à prévoir l'incidence d'une technologie sur un milieu donné au moyen des valeurs portées par les différents agents ou technologies et les potentiels conflits entre elles (Millar et al. 2019). Il existe différentes approches de *design* éthique, qui sont reprises dans les collaborations avec les ingénieurs : le « *value-sensitive design* » (VSD), le « *life-based design* » (LBD), le « *responsible research and innovation* » (RRI), dans la tradition du « *human-centered design* » (Leikas, Gotcheva et Koivisto 2019, 2, 9). Évidemment, une telle méthode de *design* éthique de machines implique que les valeurs soient identifiables d'avance, tout comme les différentes « parties prenantes », et qu'elles soient mises en relation dans différents contextes hypothétiques, parfois même au moyen de questions (Leikas, Gotcheva et Koivisto 2019, 3, 8). Un avantage de ces méthodes est qu'elles facilitent la coopération interdisciplinaire que le philosophe Mark Coeckelbergh appelle de ses vœux (2020a, 177-179).

Pour l'ingénieur et philosophe de la technologie Jason Millar (2016), les ingénieurs et les éthiciens auraient besoin d'un outil d'évaluation éthique pour mener à bien une telle démarche de design. Cet outil devrait répondre à cinq exigences : 1) préconiser une approche de proportionnalité entre les besoins des concepteurs (*designers*) et des usagers, 2) être centré sur les besoins de l'usager, 3) tenir compte des implications psychologiques d'une relation entre l'humain et le robot, 4) assurer le respect des principes du code d'éthique sur les interactions entre humains et robots (HRI) et 5) savoir faire la part des choses entre les attributs acceptables et inacceptables dans la conception (Millar 2016, 793-799).

L'idée de développer des codes d'éthique à partir de scénarios prospectifs pourrait sembler s'apparenter à un exercice imaginatif de science-fiction. Pourtant, dans son étude commandée par le Future of Life Institute, la philosophe Paula Boddington (2017) soutient que c'est le lot des chercheurs en éthique de l'IA, puisque le paysage technologique change continuellement, et à une vitesse importante, tant et aussi longtemps que de tels codes d'éthique ne sont pas simplement développés pour l'apparat (2017, 36-37, 99). Il s'agirait donc d'une approche proactive, plutôt que réactive à l'éthique de l'IA (Coeckelbergh 2020a, 168-169). Dans cette optique, une manière d'assurer le succès de l'entreprise des codes d'éthique pour l'IA est l'implication d'une diversité de personnes. Il faut également que les ingénieurs puissent traduire le contenu des codes d'éthique dans des « étapes réalisables » (Boddington 2017, 79-80, 100-101, 101).

Un enjeu fréquemment soulevé dans la littérature récente (2019-2020) est justement l'écart entre l'abstraction des principes éthiques publiés lors des dernières années et leur mise en pratique qui préoccupe les chercheurs en éthique de l'IA (Jobin, Ienca et Vayena 2019, 3; Mittelstadt 2019; Coeckelbergh 2020a, 169-174; Johnson 2020; Madaio et al. 2020; Morley, Floridi, Kinsey et Elhalal 2019; Raji et al. 2020). C'est la question de l'« applicabilité » (Tegmark 2017, 280) des principes éthiques dans les IAs qui est évoquée, jusqu'à parler d'une « coupure » entre les chercheurs en éthique de l'IA et ceux en apprentissage machine (Johnson 2020, s.p.), en raison de la « nature abstraite » des principes mis de l'avant (Madaio et al. 2020, 1). Un exemple emblématique de ce problème est l'étude qu'ont menée les chercheurs Andrew McNamara, Justin Murphy-Hill et Emerson Smith, visant à documenter l'influence du code d'éthique de l'Association for Computing Machinery (ACM), mis à jour en 2018, lors de la vague de publications de positionnements éthiques émanant des compagnies du secteur privé. Ces derniers ont relevé qu'un tel code professionnel n'avait aucune incidence sur le processus décisionnel des ingénieurs (2018, 1, 4).

En vue de remédier à ce type de décalage, une équipe de chercheurs de Microsoft ont mis au point une liste de contrôle pour opérationnaliser des principes énoncés dans une déclaration éthique comme celle de l'Organisation de coopération et de développement économiques (OCDE) (2019). La liste a été développée en partenariat avec une douzaine d'entreprises technologiques différentes et 48 ingénieurs, avec une attention particulière placée sur l'enjeu d'équité (Johnson 2020, s.p.; Madaio et al. 2020, 1). C'est justement moyennant une méthode de « co-conception » (« co-design ») que l'équipe a pu mettre au point la liste de contrôle, par l'entremise de quelques questions (Madaio et al. 2020, 4). Ce qui ressort de cet exercice est que la notion d'équité change selon les contextes et les « parties prenantes » consultées et que la priorisation de l'équité impliquera des compromis dans la programmation, ce qui enjoint les ingénieurs à se servir de leur liste avec une profonde sensibilité au contexte, si cela est possible (Madaio et al. 2020, 15).

Plus récemment, la chercheuse Jessica Cussins Newman (2020), au « Center for Long-Term Cybersecurity » de l'Université de Californie à Berkeley, a mené une étude sur la mise en pratique des principes éthiques de l'IA moyennant trois études de cas. La troisième d'entre elles porte sur

l'Observatoire de l'Organisation de coopération et de développement économiques (OCDE) des politiques relatives à l'IA, lancé en février 2020, qui cherche à assurer la mise en pratique des principes éthiques adoptés en 2019. Cette création est orientée vers une certaine « uniformisation » de l'application concrète de ces principes entre les pays du monde (Newman 2020 2, 40). En outre, l'OCDE possède un outil pédagogique pour les parlementaires qui devront légiférer sur l'IA, le « OECD Global Parliamentary Network » (Newman 2020, 40; OECD 2020a) ainsi qu'un portail informatif sur la politique de l'IA (OECD 2020b). Newman recense également les outils existants visant à « traduire » les principes éthiques pour les mettre en application en IA, dans le but d'atténuer l'écart entre la théorie et la pratique (2020, 4-8).

Je reviendrai sur les influences métaéthiques derrière les différentes propositions pour l'éthique de l'IA. Je montrerai par exemple, dans les chapitres subséquents, qu'une approche de « design éthique » peut être informée par une métaéthique empruntée à l'école du pluralisme des valeurs. D'autres traditions éthiques sont invoquées pour l'éthique des machines, comme l'éthique de la vertu (sur laquelle je reviendrai en détail ultérieurement). C'est le cas par exemple de Martin Gibert qui, insatisfait par les propositions éthiques informées par l'éthique déontologique et le conséquentialisme, propose que les algorithmes soient programmés selon les résultats de questionnaires que l'on aurait fait passer à un échantillon de personnes vertueuses de la société (environ 1 % ou moins) (Hébert-Dolbec 2020, §9; Gibert 2020). Cet échantillon de personnes devrait par ailleurs représenter les « […] diversités culturelles, de sexe et de genre, de revenus, d'éducation, de structures familiales, d'âges et autres » (Hébert-Dolbec 2020, §12; Gibert 2020).

La philosophe de la technologie Shannon Vallor (2016) avait préalablement proposé d'arrimer l'éthique de la vertu aux enjeux que posent les SIA, mais son injonction à adopter la vertu s'adresse aux humains plutôt qu'aux algorithmes. Vallor propose une éthique de la vertu « technomorale » globale (2016, 50), inspirée de la tradition éthique occidentale (avec la figure centrale d'Aristote), mais aussi des traditions orientales confucéenne et bouddhiste. Elle met ainsi de l'avant une liste de douze vertus technomorales (2016, 120) dont la pratique pourrait faciliter le développement de la technologie pour le bien humain. Vallor publie également, deux ans plus tard, un guide éthique pour la conception des nouvelles technologies, contenant sept outils : 1) le « balayage des risques éthiques », 2) l'« exécution de pré et post-mortems éthiques »,

3) « l'expansion du cercle éthique », 4) des « analyses basées sur des cas », 5) le « souvenir des avantages éthiques du travail créatif », 6) l'importance de « penser aux gens terribles » et 7) en guise de « fermeture de la boucle », la « rétroaction et l'itération éthiques » [Traduction libre] (Vallor 2018, 3).

Critique des approches éthiques comme l'éthique de la vertu et l'éthique du *care*, qu'il juge non orientées vers l'action, le philosophe Derek Leben propose une éthique « algorithmique » des robots (donc de SIA qu'il définit comme « autonomes » [2018, 1, 5]). Il tire son inspiration du réalisme moral, des jeux d'échecs,[6] des jeux de coopération, des théories de l'acteur rationnel, du principe du *Maximin* de John Rawls (qu'il adapte pour donner le *Leximin*, un terme emprunté aux jeux informatiques [2018, 87-89]) et du conséquentialisme, dans l'emploi de la raison instrumentale (Leben, 2018). Son approche, qu'il qualifie de « contractualiste » (« *contractarianistic* »), se veut une théorie morale universelle qui puisse être entrée dans les robots pour guider l'action en étant « programmable » dans des algorithmes (Leben 2018, 3, 4, 43, 73, 147). L'universalisme de la théorie est possible, selon Leben, puisqu'il applique la théorie darwinienne de l'évolution à la moralité, ce qui impliquerait que « […] les jugements moraux sont le produit de pressions évolutionnaires pour un comportement coopératif dans des organismes par ailleurs intéressés » [Traduction libre] (2018, 38).

La possibilité de pouvoir programmer des systèmes d'IA de manière éthique, voire d'« implanter » des principes et des valeurs dans les SIA, génère du scepticisme. Yampolskiy (2019) soutient qu'il n'est pas avisé de penser que des machines pourront prendre des décisions éthiques. D'abord, il critique la littérature en éthique des machines, qu'il trouve trop centrée sur les débats concernant les « bonnes valeurs morales » à inculquer aux machines. Il s'agit pour lui d'une discussion sans fin puisqu'aucune approche éthique ne serait universelle. De même, il est plus important que les machines soient sécuritaires et programmées selon le respect de la loi, plutôt qu'elles puissent débattre sur le bien et le mal (2019, §1-3). L'on pourrait suggérer que le respect de la loi est un enjeu politique, puisque l'élaboration même de ces lois doit être précédée d'une réflexion éthique par les décideurs politiques.

[6] Une autre comparaison entre la moralité et le jeu d'échecs peut être explorée dans l'article de Nicholas Denyer (1982), « Chess and Life: The Structure of a Moral Code ».

À la défense des chercheurs en éthique des machines, il ne semble pas que l'optique de ces derniers soit de faire des SIA des agents moraux qui pourraient philosopher, mais bien de s'assurer que les algorithmes vont fonctionner de manière éthiquement bonne. La volonté de Yampolskiy de s'éloigner de la philosophie pour donner la parole aux ingénieurs et concepteurs de SIA se voit tempérée par ce que Madaio et al. affirment au sujet de la nature de l'éthique, comme étant « un concept fondamentalement socioculturel » (2020, 2). Sans la perspective des sciences humaines et sociales (2020, 3), une éthique des machines pourrait devenir trop étroite et verser dans le « lavage éthique » (« *ethics washing* » ou, dans le cas de l'IA, « *ethics bluewashing* » [Floridi 2019, 187]) : il s'agit d'une sorte de rhétorique éthique qui ne s'en tient qu'aux belles paroles (Madaio et al. 2020, 2). En outre, un des enjeux dont traitent les penseurs en éthique des machines est la transparence désirable des SIA, qui permet l'attribution de responsabilité en cas d'erreur ou de faute morale. Il s'agit de la troisième sous-catégorie de la littérature en éthique de l'IA.

c) L'opacité des SIA et l'attribution de la responsabilité

Un des enjeux récurrents en éthique de l'IA est très certainement celui de l'opacité des SIA et des algorithmes qui les composent. Le fameux problème de la « boîte noire », un « […] système dont le fonctionnement est mystérieux » et qui enregistre tout (Pasquale 2015, 12-13; voir aussi Manheim et Kaplan 2019, 155; Floridi et al. 2018a, 4; Guiton 2020) parle de lui-même : les systèmes d'intelligence artificielle fonctionnant grâce aux réseaux de neurones artificiels sont parfois des énigmes pour les utilisateurs des produits, mais également pour leurs concepteurs (Manheim et Kaplan 2019, 154). L'opacité des algorithmes peut être maintenue pour des raisons diverses, constituant ainsi une sorte de « miroir sans tain » pour les utilisateurs et consommateurs des technologies, qui génère des « pouvoirs invisibles », une « aristocratie numérique » (Pasquale 2015, 21-22, 282, 318), voire « un nouveau cléricalisme [interprétant] […] une logique qui dépasse la connaissance humaine » [Traduction libre] (Crawford 2019, §21). Plus encore, il devient difficile d'exiger, comme le veut la logique du gouvernement représentatif, que cette caste politique rende des comptes, en raison même de l'opacité (Crawford 2019, §19, 21).

Le désir de transparence dans l'IA est porté par des motivations importantes, puisque les systèmes d'intelligence artificielle influencent réellement la vie d'individus, pris seuls ou en

groupes (Bostrom et Yudkowsky 2014, 316-317; Allo et al. 2016, 1; Shorey et Howard 2016, 5034; Brundage et Bryson 2016, 1). L'immense champ de recherche d'analyse des données [« data analytics »] emploie l'IA pour établir des prédictions concernant les consommateurs « […] sur la base de leurs données financières, démographiques, ethniques, raciales, concernant leur santé, sociales, ainsi que d'autres données » [Traduction libre] (Manheim et Kaplan 2019, 120). Ce que le professeur d'éthique Dominic Martin appelle des décisions « moralement chargées », qui seraient prises par des machines (ou encore des prédictions faites par des machines qui auraient une incidence sur des décisions subséquentes) implique tout ce qui peut causer du tort aux personnes (physiquement et psychologiquement) ainsi qu'à des réalités matérielles comme des infrastructures (Martin 2017, 3). Par exemple, des systèmes d'IA sont employés pour déterminer le coût des assurances pour un particulier, selon sa catégorie de risque, moyennant les données sur sa santé (Allo et al. 2016, 3; Manheim et Kaplan 2019, 121), ou encore le pourcentage de probabilité de récidive d'une personne reconnue coupable d'une faute criminelle (Müller 2020, §2.4; Manheim et Kaplan 2019, 156-157). Par conséquent, un problème épineux se pose quand des SIA sont employés par des compagnies d'assurance, des cabinets d'avocats ou d'autres entreprises qui ne savent pas comment ils ont été entraînés ni au moyen de quelles données (Manheim et Kaplan 2019, 155-156).

Les algorithmes sont appelés à accomplir plusieurs tâches impliquant des décisions partiales, comme « la classification, la priorisation, l'association et le filtrage », de manière à ce que l'information soit, en fin de compte, retravaillée (Shorey et Howard 2016, 5033 [Traduction libre]). Ils ne sont donc pas neutres, puisque les données dont ils sont nourris ne le sont pas non plus, mais l'algorithme est appelé à les généraliser, sur la base des récurrences identifiées (Allo, et al. 2016, 1, 3). De plus, l'algorithme, par son fonctionnement même, filtre l'information qui est ensuite transmise aux utilisateurs, créant une sorte de paradoxe. Allo et al. relèvent en effet que, si le contenu est « personnalisé » à l'utilisateur grâce à l'algorithme, cela peut à la fois alimenter et entraver l'autonomie de l'utilisateur en question. Trop d'information rend difficile l'exercice de l'autonomie, mais force est de constater que « […] décider quelles informations sont pertinentes est intrinsèquement subjectif » (au sens de « partial ») [Traduction libre] (Allo et al. 2016, 9). Il est toutefois facile, voire tentant, de croire à la neutralité des algorithmes : le chercheur Matthew B. Crawford pense que

> notre volonté d'accepter la notion de contrôle anonyme est certainement due, en partie, à notre idéal de l'équité procédurale, qui exige que le pouvoir discrétionnaire individuel exercé par les personnes au pouvoir soit remplacé par des règles, chaque fois que cela est possible, car l'autorité sera inévitablement abusée. C'est le noyau originel du libéralisme, datant de la Révolution anglaise. [Traduction libre] (2019, §2, voir aussi § 49-51)

Le but que constitue la transparence se heurte malheureusement à l'impératif de sécurité qui exige que certaines informations demeurent confidentielles. À ces injonctions de restriction de l'information au nom de la sûreté, le professeur de droit Frank Pasquale (2015) apporte quelques bémols. Les géants de la finance et de la technologie accumulent nos renseignements personnels en banque de mégadonnées (ou gigadonnées), sans nous laisser voir ce qui en est fait (Pasquale 2015; Werner 2019). Des lois entourant la protection de la vie privée peuvent cacher 1) des secrets réels, 2) des secrets juridiques (ou professionnels), ou encore 3) créer du secret par la technique de l'« obfuscation », qui « […] consiste à rendre illisible pour un humain un programme, tout en le gardant pleinement fonctionnel, ou en dissimulant une information sous un déluge de données » (Pasquale 2015, 17). La logique de l'obfuscation peut être observée dans les contrats en ligne, par exemple (Frischmann et Selinger 2018, 291-294). Calo prédit lui aussi que l'accès aux données possédées par les gouvernements et par les entreprises sera rendu difficile sous le prétexte de la protection de la vie privée du consommateur. Les firmes pourraient esquiver, de la sorte, l'impératif de divulguer leurs données d'entraînement, voire de les partager (Calo 2017, 19).

Toutefois, transparence et éthique de l'IA ne sont pas synonymes, soutient Mike Ananny. Si l'interprétation de la boîte noire est certainement nécessaire, elle ne garantit pas à elle seule que les algorithmes soient « éthiques » (2016, 109). Ce sont leurs retombées, une fois assemblés en « algorithmes d'information en réseau » (« *networked information algorithm* »), qui devrait faire l'objet de recherches (Ananny 2016, 97, 107). En effet, ces derniers consistent en

> […] un assemblage de codes informatiques, de pratiques humaines et de logiques normatives institutionnellement situés qui créent, soutiennent et symbolisent les relations entre les personnes et les données par une action semi-autonome observable au minimum [Traduction libre] (Ananny 2016, 99).

Il convient alors de se demander au service de quoi et de qui ces assemblages d'algorithmes fonctionnent (Ananny 2016, 99-100). C'est au moyen d'une combinaison des trois écoles éthiques

dominantes (éthique déontologique, conséquentialisme et éthique de la vertu) qu'Ananny propose que cela soit fait (109).

L'autre versant de l'enjeu de l'opacité (ou de la transparence, dans son acception positive) est l'attribution de la responsabilité. La « traçabilité » de la logique d'une décision algorithmique peut se révéler difficile et entraver de ce fait la possibilité d'attribuer la responsabilité de cette décision (Allo et al., 2016, 5, 11; voir aussi Bostrom et Yudkowsky 2014, 317). Des outils d'évaluation de la portée des prises de décisions guidées ou menées par des algorithmes (« *algorithmic impact assessment* ») ont été développés dans le but d'aider les gouvernements et acteurs publics à assumer la responsabilité de ces décisions (Reisman et al. 2018, 3-4; Gouvernement du Canada 2020). Aux États-Unis, le Parti démocrate a déposé le projet de loi sur la responsabilité algorithmique, pour tâcher de réagir aux biais raciaux des algorithmes : on lui reproche néanmoins, parmi d'autres choses, de manquer de « dents » pour forcer les compagnies à le mettre en pratique (Kaminski et Selbst 2019).

D'autres défis se posent : Oren et Amitai Etzioni se demandent comment il est possible de déterminer, dans le cas où une décision automatisée aurait causé du tort à un individu ou à un groupe, si ce tort était intentionnel, dans le contexte où la logique de la décision elle-même est difficile à saisir, même pour les experts techniques (2016a, 150; Müller 2020, §2.3). Ceci peut consister en une menace à la démocratie, qui se transformerait en régime « algocratique » ou en « autoritarisme digital » (Danaher 2016, dans Müller 2020, §2.3; Manheim et Kaplan 2019, 111). De toute façon, estime Martin, les concepteurs et ingénieurs en IA ne peuvent être chargés seuls de la responsabilité que les machines puissent prendre de bonnes décisions morales (2017, 13) et les « traduire » en langage adapté à la compréhension humaine. La notion de « moralité distribuée », empruntée à l'épistémologie qui avait celle de « connaissance distribuée » est mise de l'avant par Floridi (2014, 261). Dans cette façon de voir, la moralité est partagée : elle est hétérogène, « […] socialement distribuée, de sorte que des acteurs placés différemment dans une situation ont des rôles moraux différents à jouer […] » [Traduction libre] (Boddington 2017, 23).

Peut-être cette idée trouve-t-elle un écho dans le champ de l'éthique de l'IA précisément en raison de son interdisciplinarité intrinsèque. Au-delà de la diversité d'agents humains, dans les

situations où la prise de décision éthique est hybride, soit partagée entre l'humain et la machine (ce que Floridi appelle des « systèmes multi-agents » [2014, 265]), il est possible de parler de moralité distribuée si les agents interagissent (267). L'éthique devient applicable non seulement à des systèmes, ou à des individus, mais à toute une infrastructure : il s'agit alors d'« infraéthique », qui devient positive si un nombre suffisant d'agents agissent comme « facilitateurs moraux » (Floridi 2014, 272, 275). C'est dans le même sens qu'abonde Nurock dans son traitement de l'éthique de l'IA : c'est l'infrastructure qu'il faut repenser, pas simplement des enjeux « accidentels » comme la diversification des données (Nurock 2019, 74).

Une partie du problème de l'opacité peut être résolue en faisant la promotion de la reproductibilité dans les modèles algorithmiques : si les données qui alimentent le SIA sont disponibles en accès libre (« *open access* »), non seulement la compréhension globale de la décision pourrait être facilitée et le domaine de l'IA pourrait progresser comme science, selon l'informaticien Zach Scott (2018). Une telle injonction est valable dans les cas où les données ne sont pas confidentielles. En revanche, si la reproductibilité n'est pas possible en ces termes, un autre facteur qui peut favoriser la transparence est de garantir ce que le professeur de mathématiques et sciences physiques Philip B. Stark (2018) appelle, avec un néologisme, la « préproducibilité ». Peut-être que cet appel à la « préproducibilité » participerait de la logique de l'infraéthique amenée par Floridi (2014). La première transparence à favoriser serait, dans l'optique de Stark, celle de l'expérience elle-même, avant de voir son potentiel de reproductibilité.

L'importance de la transparence en éthique de l'intelligence artificielle est centrale pour une autre série de raisons touchant à l'équité. En effet, des décisions automatisées peuvent impliquer : 1) la possibilité de prendre une décision sans évidence, 2) des décisions prises sans que le lien entre la conclusion et le processus décisionnel n'apparaisse clairement et 3) des données biaisées qui influent sur les décisions prises (Allo et al. 2016, 4-5). Ces problèmes épistémiques peuvent entraîner des injustices dans la pratique, comme du profilage criminel, ou encore de nouvelles catégorisations du monde (Allo et al. 2016, 5; Ananny 2016, 97, 101; Dilhac 2018). C'est ainsi que se glissent les difficultés que représentent les biais algorithmiques qui débouchent sur la discrimination.

Quelques exemples sont récurrents dans la littérature en éthique de l'IA, comme le moteur de recherche Google qui affichait des photos de gorilles lorsque l'utilisateur cherchait des personnes de couleur noire, ou encore l'algorithme aidant au recrutement d'employés chez Amazon.com inc. qui pénalisait systématiquement les curriculum vitae de femmes ayant posé leur candidature (voir par exemple Crawford 2016; Medhora 2018; Manheim et Kaplan 2019, 159; Serebrin 2019; Nurock 2019, 65-66; Müller 2020, §2.4). C'est sur cette question que le propos se dirige à présent.

d) Les risques de biais et de discriminations

On a vu que la sélection de l'information par les algorithmes implique un tri à caractère partial : « les algorithmes peuvent être incapables de reproduire "la découverte spontanée de nouvelles choses, idées et options" qui apparaissent comme des anomalies par rapport aux intérêts profilés d'un sujet » [Traduction libre] (Allo et al. 2016, 9). Cette dynamique ressemble au biais cognitif de confirmation, imposé cette fois par un algorithme. Le traitement des individus par un système de calcul peut se faire sur la base de leur appartenance aux groupes que ce dernier a créés et dans lequel il les a placés (Allo et al. 2016, 10).

On peut établir un parallèle avec l'argumentaire de la politologue Virginia Eubanks (2017), pour qui les discriminations sont inscrites dans les systèmes automatisés employés pour la prestation de services sociaux. Grâce à son analyse comparative d'un modèle de prédiction de la violence faite aux enfants en Pennsylvanie, d'un système d'aide sociale dans l'État de l'Indiana et d'un registre électronique de personnes sans logement à Los Angeles (2017, 10), Eubanks démontre que de tels systèmes automatisés reproduisent les inégalités sociales au point de les réifier au plan digital. Évoquant « les maisons des pauvres » du 19e siècle (que l'on retrouve dans certains romans de Dickens, par exemple *La petite Dorrit*), dans lesquelles étaient logées les personnes sans ressources ou endettées, Eubanks sonne l'alarme en dénonçant des systèmes automatisés qui reproduisent ces maisons en version 2.0 ou 3.0.

Des biais inhérents aux structures des SIA eux-mêmes pourraient être à l'origine de discriminations sociales, comme l'a suggéré, dès la fin des années 1990, la chercheuse Alison

Adam (1998). L'intelligence artificielle porte en elle une certaine vision du monde « genrée » qui ne frapperait pas le regard, mais existerait de manière subtile, dit-elle (1998, 1). Il y aurait donc le risque que les SIA « […] automatis[ent] des structures sociales de domination » (Nurock 2019, 74). Dans la même veine, Safiya Umoja Noble propose une analyse de ce type de biais, s'inscrivant dans l'approche épistémologique des *Black feminist technology studies* (BFTS), recensant les relations de pouvoir déséquilibrées à l'œuvre dans les algorithmes (2018, 171). Freya Werner (2019) voit dans l'algorithme de reconnaissance faciale des photos de profil de comptes Twitter, qui ne peut identifier que trois races et deux genres, une illustration de ce phénomène. Dans cette perspective, ce n'est pas l'IA en tant que technologie qui pose problème, mais les structures et inégalités qu'elle reproduit, selon Werner. D'autres estiment, au contraire, que si les technologies IA n'arrivent pas à fonctionner sans commettre des bévues équivalant à de la discrimination, il vaudrait mieux ne pas les employer du tout (Wadell 2019).

Les biais cognitifs des humains et les biais algorithmiques peuvent être comparés, et cela même sans tracer une adéquation entre les humains et les machines en tant que telles (Johnson s.d., 21-22). Puisque l'intelligence artificielle est « […] un enjeu non seulement technologique, mais aussi, et surtout social, éthique et politique », elle « […] nous tend le miroir de notre histoire patriarcale […], de nos biais et discriminations » (Nurock 2019, 61, 62). Les assistants vocaux d'Amazon, Apple et Windows ont tous des voix de femmes, de faire remarquer Nurock : Alexa, Siri et Cortana (2019, 66). La manière de traiter les enjeux éthiques associés à l'intelligence artificielle, en plus de la discipline « technique » elle-même, est teintée par les discriminations de genre, soutient-elle. Par exemple, la récurrence de dilemmes moraux comme celui du tramway traduit une volonté de penser l'éthique « d'un "point de vue de nulle part" », à partir d'une « fausse neutralité », ce qui renvoie à une manière de faire plutôt masculiniste que féministe (Nurock 2019, 72-73, citant Gilligan 2008).[7] (Cette idée de « fausse neutralité » de l'IA est également relevée par d'autres penseurs, comme le chercheur et informaticien Sabelo Mhlambi [2020, 26]). Ces dilemmes sont en réalité des scénarios moraux dans lesquels l'autonomie est exaltée au-dessus du soin des relations interpersonnelles, ce qui aurait pour effet d'« artificialiser » l'éthique (Nurock 2019, 73-74).

[7] Nurock soutient que l'éthique du *care* de Carol Gilligan (2008) n'est pas féminine, mais féministe (2019, 73). Ma lecture de Gilligan me pousse à penser le contraire, mais je tenais à rapporter le propos de Nurock dans son intégrité.

Au-delà de la structure interne des systèmes d'intelligence artificielle, il y a certes les courants sociétaux, mais surtout la malice des êtres humains. Ce seraient nos biais qui engendreraient des discriminations automatisées, écrit Leben :

> si nous utilisons les jugements humains pour modéliser les jugements des machines, les robots intégreront inévitablement les biais et les incohérences de notre propre psychologie : préférence pour les personnes familières ou génétiquement apparentées, ignorance des effets de nos actions sur les personnes très éloignées, et recours à de fausses croyances sur les types d'actions qui sont nuisibles. Au cours de l'histoire, un nombre considérable d'êtres humains ont été élevés pour approuver des choses horribles comme le génocide, le viol, l'esclavage, la torture et la maltraitance des enfants, pour n'en citer que quelques-unes. [Traduction libre] (Leben 2018, 3)

Les machines pourraient donc mieux fonctionner moralement que les humains, affirme Leben (2018, 3, 147).

Sur le plan de la programmation, ce qui rend la tâche difficile aux concepteurs de SIA, c'est la qualité des données et la manière d'entraîner ces derniers. En effet, « […] l'IA adapte ses algorithmes jusqu'à ce qu'il en trouve un qui donne toujours, ou presque toujours, le même résultat que celui des cas d'entraînement » [Traduction libre] (Manheim et Kaplan 2019, 158). En conséquence, un algorithme entraîné sur des images de personnes de couleur blanche fonctionnera mieux sur les caucasiens et logiquement, commettra des erreurs sur les autres (Calo 2017, 10). À ce sujet, les biais associés aux dynamiques raciales sont très discutés dans la littérature en éthique de l'IA. Certains SIA sont dénoncés comme étant racistes, comme un algorithme employé au Royaume-Uni depuis 2015 pour l'attribution de visas aux migrants qui favorisait l'accueil de personnes de couleur blanche (McDonald 2020). Dans le domaine de la santé, l'algorithme « Optum », employé par la firme « UnitedHealthcare », aurait aussi entretenu un biais discriminatoire à l'endroit des patients de couleur noire (Shapiro et Blackman 2020, §1).

Les biais sont difficiles à détecter de manière proactive, plutôt qu'une fois les dégâts commis. Par exemple, il est possible que des algorithmes biaisés soient toujours employés, puisqu'au moment de leur conception, on n'accordait pas énormément d'importance aux questions d'équité ou de représentativité dans les données (Field 2020, § 4). Plus encore, selon Müller (2020) les solutions techniques aux biais se heurtent à la difficulté de devoir traduire des notions comme

l'équité ou la « race » en langage mathématique (§2.4). Les chercheurs Wachter, Mittelstadt et Russell établissent un constat similaire en ce qui a trait à la notion d'équité dans la législation européenne. La mesure statistique de l'équité ne se comprend pas de la même manière que le concept juridique d'équité tel qu'employé, par exemple, à la Cour européenne de justice. Par conséquent, les manques d'équité dans l'automatisation, qui débouchent sur la discrimination, sont parfois difficiles à détecter, et c'est la raison pour laquelle ces chercheurs visent à combler l'écart par de nouvelles mesures statistiques plus adaptées (Wachter et al. 2020).

Au-delà des solutions techniques visant à contrer la discrimination, d'autres chercheurs proposent de repenser la manière dont l'éthique de l'IA est menée, de manière à déraciner, dès le début, toute tendance discriminatoire. C'est dans cet ordre d'idées que le chercheur Sabelo Mhlambi appelle à penser l'éthique de l'IA à partir de la tradition africaine ubuntu, qui n'aurait pas en germe les défauts des approches occidentales, comme l'éthique kantienne (Weinberger et Mhlambi 2020, §6). Le propos de Vallor (2016) s'approche et s'éloigne à la fois de cette prise de position, en ce qu'elle voit dans les approches occidentales et orientales un noyau commun, la vertu. Ess abonde dans le même sens lorsqu'il souligne les parallèles entre l'éthique de la vertu, le confucianisme, le bouddhisme et l'islam au moyen de la notion d'harmonie, par exemple (2020, 564). Mhlambi soutient quant à lui que l'éthique ubuntu est en porte à faux avec les éthiques dites occidentales et qu'elle est idéale pour approcher les enjeux que pose l'IA (Weinberger et Mhlambi 2020; Mhlambi 2020). Elle préconise la compassion et l'équité (Mlhambi 2020, 7) la restauration des relations plutôt que la rétribution des fautes et la « relationalité » plutôt que la rationalité comme conception de la personne; elle place les personnes avant l'efficacité et le profit (Weinberger et Mhlambi 2020, §8-10, 16, 18; voir aussi, au sujet de la relation, Hongladarom 2017 dans Ess 2020, 563).

La conception occidentale de la personne comme rationnelle aurait dévié au cours des siècles, selon Mhlambi, pour aboutir à un idéal positiviste, à l'exaltation de ce qui est mesurable et à l'individualisme (2020, 2-3, 5, 7). Plus encore, il critique le fait que, dans sa compréhension, les non-Européens n'aient pas été considérés comme des personnes selon ces idéaux occidentaux (2020, 5). Ainsi, soutient-il,

> la prétendue infaillibilité et suprématie de la rationalité, en particulier lorsqu'elle est administrée par des machines, exacerbe la marginalisation de ceux qui, dans la société, ont été rejetés sur la base de critères soi-disant rationnels ou « productifs » [Traduction libre] (Mhlambi 2020, 5)

En conséquence, « […] nous avons besoin d'humains "dans la boucle" ("*in the loop*"), pour fournir le contexte nécessaire qui fait souvent défaut aux machines et aux données. Un ensemble diversifié d'humains » [Traduction libre] (Weinberger et Mlhambi 2020, §15). On peut mentionner que Charles Ess apporte une nuance en déclarant que, pour une théorie ou une approche éthique, « […] les origines occidentales seules n'offrent pas les conditions suffisantes pour établir l'impérialisme du concept *a priori* » [Traduction libre] (2020, 564).

Ce que Mhlambi décrit serait au fond un aveuglement culturel dans l'éthique de l'intelligence artificielle. ÓhÉigeartaigh, Whittlestone, Liu, Zeng et Liu (2020) observent également que davantage de ponts interculturels pourraient être bâtis pour penser l'éthique et la gouvernance de l'IA. Bien qu'ils se concentrent sur les tensions générées par les problèmes de perception entre la Chine et les États-Unis, le fond du propos demeure, à leur avis, pertinent pour toutes les cultures : il est possible de s'entendre sur la mise en pratique de certaines politiques concernant l'IA, même si l'on n'est pas entièrement d'accord sur la manière de concevoir les principes et les valeurs qui les sous-tendent. La collaboration interculturelle devrait par ailleurs, selon ÓhÉigeartaigh et al., être promue par le monde académique, notamment par l'entremise de la traduction des travaux sur l'éthique de l'IA en plusieurs langues.

L'un des grands points de divergence qui est fréquemment invoqué pour marquer la différence de culture entre la Chine et les États-Unis est la question de la vie privée et de la surveillance. Le système de crédit social chinois est souvent mis de l'avant comme représentant l'empiétement sur la vie privée des personnes, mais aussi comme ayant un effet sur le comportement social des individus (ÓhÉigeartaigh et al. 2020, 5, 8; Smellie 2019; Larson 2018; Sproule 2018, §23). Les deux prochaines sections de la revue de littérature permettront d'aborder plus largement ces enjeux.

e) La protection de la vie privée et la surveillance

La possibilité que la structure des SIA, les données ou encore la culture qui informe la programmation des algorithmes soit biaisée et mène à des discriminations est bien établie dans la littérature. Cet enjeu est étroitement lié à celui de la surveillance, et difficile à dissocier dans l'absolu, puisque les algorithmes peuvent aussi être employés de manière à enfreindre le respect de la vie privée des individus, et ce, en perpétuant certaines discriminations sociétales (par exemple Harari 2018, §23-26). Ce peut être illustré par le fait qu'à New York et à Los Angeles, on fait appel aux outils de la « police prédictive » pour anticiper les endroits où les crimes pourraient avoir lieu (selon les données recueillies par le passé). Du personnel policier est déployé dans les quartiers jugés plus à risque (Ananny 2016, 106). Cette pratique n'est pas nécessairement une atteinte à la vie privée des individus, mais elle court le risque de perpétuer des biais associant certains lieux — et donc les personnes y demeurant — avec le crime. Plus encore, le fait d'accroître les patrouilles de surveillance dans un lieu donné peut mener à détecter davantage de crimes dans cette zone que dans une autre, moins surveillée (Müller 2020, §2.4).

Ce qui verse manifestement dans la discrimination, ce sont des pratiques comme celles que l'État de Pennsylvanie envisageait d'adopter, soit de

> […] permettre aux juges d'utiliser des estimations statistiques pour les futurs délits afin de déterminer la peine actuelle d'un détenu — le punissant non seulement pour les crimes qu'il a commis, mais aussi pour ceux que les algorithmes pensent qu'il pourrait commettre [Traduction libre] (Ananny 2016, 106)

Müller fait remarquer que plusieurs craignent les dérives de la police prédictive, qui pourraient « […] conduire à une érosion des libertés civiles, car elle peut enlever du pouvoir aux personnes dont le comportement est prédit » [Traduction libre] (2020, §2.4).

Le droit à la vie privée implique plusieurs autres droits, comme celui de maintenir la confidentialité sur ses informations personnelles ou encore de prendre ses propres décisions (Manheim et Kaplan 2019, 116). Les notions d'autonomie et de libre arbitre sont donc intimement connectées à la protection de la vie privée, puisque sans cette dernière, des tiers peuvent empiéter sur l'autonomie des individus par toutes sortes de moyens, allant de l'influence à la coercition, en

passant par la manipulation (Manheim et Kaplan 2019, 117). Par exemple, les téléphones « intelligents » présentent plusieurs enjeux touchant à la vie privée : des applications de traçage (par exemple Bengio et Dilhac, 2020; Cohen et Gupta 2020; Deglise 2020; Pilon-Larose 2020; Clark 2020), aux applications de partage photos et vidéos, des systèmes de domotique aux outils de navigation GPS (Maclure et Saint-Pierre 2018, 745-746; Manheim et Kaplan 2019, 123) : des gigaquantités de données à caractère personnel circulent chaque jour et sont agrégées. Ces gigadonnées (ou mégadonnées) (« *big data* ») sont en réalité

> […] de grandes quantités d'informations recueillies sur de nombreuses personnes utilisant de multiples appareils. Plus que la taille, [les gigadonnées] caractérisent des ensembles de données qui peuvent être recherchées, agrégées et triangulées avec d'autres ensembles de données. […] Les gigadonnées prennent la forme d'artefacts de communication, tels que des photographies, le microciblage de profils, le contenu des réseaux sociaux et les métadonnées. [Traduction libre] (Shorey et Howard 2016, 5033).

La raison pour laquelle les applications des téléphones, de même que les services (comme les moteurs de recherche) en ligne sont gratuits, soulèvent le professeur de droit Karl M. Manheim et l'avocate Lyric Kaplan (2019), c'est que nos données sont le coût d'utilisation de ces services, puisque les compagnies les revendent à d'autres parties pour leur usage (124). Ce peut être pour personnaliser la publicité qui arrive à chaque consommateur, et ce, parfois de manière cachée, ce qui s'apparente à de la manipulation, de soutenir Kaplan et Manheim (2019, 130). Face à ce problème, Frischmann et Selinger suggèrent que les compagnies de médias sociaux adoptent un modèle commercial comme celui de la British Broadcasting Corporation (BBC), qui ne dépend pas des données de ses utilisateurs et des publicités qui leur sont ensuite envoyées, et qui possède par ailleurs une certaine indépendance par rapport au gouvernement (Frischmann et Selinger 2018, 281-282).

D'autres technologies employant l'intelligence artificielle pour le traitement des données qu'elles recueillent sont désignées par le néologisme « Internet of Things (IoT) », soit l'« Internet des objets ». Il s'agit au fond d'

> […] un écosystème de capteurs électroniques présents sur notre corps, dans nos maisons, bureaux, véhicules et lieux publics […] qui se voient attribuer une adresse Internet et transfèrent des données sur un réseau sans interaction entre deux humains ou entre un humain et un ordinateur [Traduction libre] (Manheim et Kaplan 2019, 122).

Le but est évidemment d'améliorer l'assistance technologique dans plusieurs aires de nos vies (par exemple, les commandes vocales pour des appareils à la maison), mais la contrepartie est que ces données peuvent être employées pour nous influencer, voire nous contrôler (Manheim et Kaplan 2019, 123). Non seulement les appareils personnels, mais ceux liés à « l'économie du partage » fonctionnent souvent grâce à l'IA : c'est le cas pour des compagnies comme Lyft et Uber, par exemple (Brundage et Bryson 2016, 1). L'omniprésence de la collecte de données est à prendre en compte dans le contexte actuel où « l'intelligence artificielle est de plus en plus capable de déduire ce qui est intime de ce qui est disponible » [Traduction libre], ce qui brouille la distinction entre ce qui est public et ce qui est privé (Calo 2017, 17).

Les « villes intelligentes » (« *smart cities* ») présentent encore le risque de surveillance et d'entrave à plusieurs droits liés à la vie privée, notamment par l'emploi de caméras situées dans les rues, ou de détecteurs de plaques d'immatriculation (Manheim et Kaplan 2019, 124). Ce type de surveillance, qui peut impliquer des technologies de reconnaissance faciale, peut servir les fins d'un gouvernement (Calo 2017, 18), ou encore celles des services de police (Ducas 2020; Robertson, Khoo et Song 2020). En Espagne, un dispositif de reconnaissance faciale a été installé au terminal de bus du sud de Madrid, sans que cela ne soit connu ou diffusé, en vue de partager les informations recueillies à la police. Selon cette dernière, le système a été mis en place pour la défense des intérêts publics et pour assurer la sécurité (López-Molina 2020). En revanche, au pays de Galles, la Cour d'appel a reconnu que les outils de reconnaissance faciale employés par la police enfreignaient « […] les droits à la vie privée, les lois sur la protection des données et les lois sur l'égalité », et que la police n'avait pas vérifié, au préalable, que le SIA ne comportait pas de biais dans son fonctionnement (Liberty 2020).

Ces « villes intelligentes » ne sont pas conçues — ou du moins elles ne prétendent pas l'être — dans un but de surveillance, mais plutôt pour améliorer la vie des citoyens. Crawford, dans sa critique du « capitalisme de surveillance » (qu'il reprend de Zuboff 2019), reconnaît que la planification urbaine pourrait bénéficier de telles innovations technologiques, par exemple « […] en étant capable de prédire la demande de chauffage et d'électricité, de gérer l'attribution des capacités de circulation et d'automatiser l'élimination des déchets » [Traduction libre] (2019, §26; voir aussi Zuboff 2019, 5-7). Au fond, un « autoritarisme bénin », le consentement à quelques

formes de surveillance par des agences de l'État, pourrait garantir de meilleurs services et c'est le compromis que plusieurs feraient, estime-t-il (Crawford 2019, §27).

En ce qui concerne les enjeux de la surveillance en ligne, c'est la collecte de « cookies » dont il faut se préoccuper (Manheim et Kaplan 2019, 124). Ces derniers qui peuvent être comparés aux miettes que laissait, dans le conte repris par les frères Grimm, le petit Hansel derrière lui et sa sœur sur le chemin de la forêt, question que l'on puisse déterminer leur parcours exact. Dans le cas des « cookies » informatiques, ce sont les serveurs qui obtiennent ces informations (Manheim et Kaplan 2019, 124). Les cookies font donc partie de ces données qui sont amassées et transférées à des tiers sans que le citoyen moyen ne se rende bien compte de ce qui s'opère (Calo 2017, 17). Il faut d'ailleurs savoir que le processus d'anonymisation (ou de « pseudonymisation ») des données n'est pas toujours efficace. Il est possible, avec un peu de travail, de « re-personnaliser » des données, et ainsi permettre le profilage et la surveillance, à toutes sortes de fins, notamment commerciales, partisanes et politiques (Shorey et Howard 2016, 5034; Manheim et Kaplan 2019, 111, 128). Une autre forme de « surveillance en ligne » qui tend à se manifester, selon Crawford, est le risque que des médias sociaux comme Facebook enfreignent la liberté d'expression au nom de la rectitude politique (2019, §31, 39).

Il existe un paradoxe concernant le respect de la vie privée, à savoir l'inquiétude que la collecte de données en ligne suscite et, malgré tout, la persistance des utilisateurs à maintenir leurs habitudes sur la toile (Serebrin 2019). On pourrait alléguer que, en raison de notre attachement à l'efficacité (Selinger 2019, §7), rares seraient les individus qui se donneraient la peine de vérifier la confidentialité de leurs informations tout au long de leurs recherches Web. Des solutions ont été proposées afin d'améliorer le respect de la confidentialité. Parmi elles, la fiducie de données (« *data trust* »), mise de l'avant, par exemple, par Element AI (Serebrin 2019). L'avantage de ce modèle est que les données seraient gérées par une tierce partie, et non directement par les grandes entreprises technologiques (Element AI 2020). Toutefois, ce modèle n'est pas encore adopté par des géants de l'Internet comme les GAFA : Google, Apple, Facebook et Amazon. Si l'on y ajoute quelques compagnies comme IBM, Microsoft et Baidu en Chine, il y a là les institutions les plus puissantes dans le monde de l'IA (Calo 2017, 5). La concentration des données en ces quelques compagnies soulève la question de la parité des données sur le marché, de même qu'un certain

« exode des cerveaux » en IA vers le secteur privé, là où les ressources abondent (Calo 2017, 19). En conséquence,

> […] les applications utilisant l'apprentissage machine seront systématiquement orientées vers les objectifs des entreprises à but lucratif et non vers ceux de la société dans son ensemble. Les entreprises posséderont non seulement davantage d'informations et les exploiteront, mais elles auront aussi le monopole de leur analyse sérieuse. [Traduction libre] (Calo 2017, 19)

Il s'agirait donc d'un « oligopole » des compagnies de technologie sur le marché (Wright 2019). Une « économie de la surveillance », basée sur une « mutation dévoyée du capitalisme », constituant « un rejet de la souveraineté du peuple » (Zuboff 2019, s.p.) est ainsi mise en place.

La sociologue Shoshana Zuboff explique que le « capitalisme de surveillance » « […] revendique unilatéralement l'expérience humaine comme une matière première libre à traduire en données comportementales », servant de la sorte à nourrir un marché de prédictions du comportement des consommateurs (Zuboff 2019, 8). Non seulement ces technologies sont développées sur la base de savoirs accumulés sur les consommateurs, mais elles influencent également leurs habitudes et leurs décisions, dans le but, dit Zuboff, de « nous automatiser » (Zuboff 2019, 8). On assiste alors à l'émergence d'un nouveau pouvoir, qu'elle appelle « l'instrumentarisme » (« *instrumentarianism* ») (Zuboff 2019, 8). Les enjeux de la surveillance sont étroitement liés à ceux de la manipulation du comportement, que ce soit au plan individuel, par des incitatifs dirigés envers des individus, ou encore à plus grande échelle, sur le plan politique. On l'aura compris, si le fonctionnement de ces systèmes n'est pas saisi par le citoyen moyen, le risque qu'il encourt est de se faire manipuler par ceux qui les administrent (Susskind 2018, 25).

f) La manipulation du comportement et les effets sur les régimes politiques

L'usage de l'intelligence artificielle présente quelques menaces à l'autonomie individuelle, mais aussi, par ricochet, à la démocratie elle-même. Certains voient même dans cette technologie une incompatibilité fondamentale avec ce qu'ils conçoivent comme les principes sous-jacents à la démocratie : l'équité, la responsabilité et la transparence (Manheim et Kaplan 2019, 133). Cette

inquiétude n'est pas très loin de celle de Jamie Susskind, qui identifie quelques concepts de base de la pensée politique (le pouvoir, la liberté, la démocratie et la justice sociale [2018, 10]) pour penser l'avenir de la politique. Pour lui, ce n'est pas que la technologie soit incompatible avec la démocratie, mais que si les fondements de la politique sont bouleversés de façon profonde et sans que quelque chose soit fait pour l'empêcher, la politique elle-même pourrait perdre son sens (Susskind 2018, 25).

Les exemples abondent dans la littérature. L'IA peut servir à pirater des élections de plusieurs manières, par exemple en manipulant les frontières de circonscriptions électorales, pour favoriser un ou plusieurs groupes. Elle peut être employée, comme cela a été le cas aux États-Unis en 2016, pour répandre de la propagande ciblée et de fausses nouvelles, et ainsi miner la confiance des électeurs envers leurs institutions politiques et médiatiques (Manheim et Kaplan 2019, 133-134, 137, 144, 150-151). Le problème inhérent aux fausses nouvelles est que les compagnies derrière les médias sociaux sont orientées vers la recherche du profit au détriment de la vérité des contenus, comme l'expliquent Frischmann et Selinger (2018, 280). En outre, certaines couches de la population, par exemple les aînés, seraient plus susceptibles à leur influence (Brashier et Schacter 2020). La propagande pose un problème différent en ce que la vérité ou la fausseté des contenus est au contraire très importante, car l'influence et la diffusion de ces contenus sont le but recherché (Frischmann et Selinger 2018, 280). L'intelligence artificielle pourrait jouer un rôle très efficace dans la dissémination de faux contenus. En 2019, l'organisme OpenAI a décidé de ne pas publier son générateur de texte GPT-2. Alimenté de huit millions de pages Web et entraîné de manière non supervisée, GPT-2 performe si bien que les textes qu'il produit sont fort crédibles et, entre de mauvaises mains, il pourrait faire des ravages, notamment en incluant dans ses textes des figures publiques réelles (Pringle 2019). Cela dit, une étude de Neil L. Levy et Robert M. Ross (2020), du côté de la neuroscience, tend à tempérer l'idée selon laquelle les fausses nouvelles attireraient la croyance d'un grand nombre de personnes. Encourager le dialogue pourrait par ailleurs contribuer à diminuer leur portée (Levy et Ross 2020, 14).

Le tristement célèbre scandale de la firme Cambridge Analytica, en partenariat avec Facebook, met en relief les risques associés à la collecte de données personnelles. Sans obtenir au préalable l'assentiment des utilisateurs du média social, elle a pu dresser 230 millions de portraits

psychographiques d'Américains pour que Facebook bombarde ces derniers avec de la publicité politique ciblée et personnalisée, en vue d'influencer leur comportement électoral (Manheim et Kaplan 2019, 139; Müller 2020, §2.2). La manipulation du comportement à des fins politiques ou commerciales est un enjeu rendu réel « […] par la capacité de l'IA à détecter des tendances ("*patterns*") dans un monde complexe » [Traduction libre] (Calo 2017, 19). La manipulation électorale par le gouvernement ou une agence gouvernementale aurait été observée dans 48 pays en 2018 (Manheim et Kaplan 2019, 141). Toutefois, en août 2020, Facebook, Google et YouTube ont annoncé avoir mis en place des filtres permettant de ne pas diffuser des vidéos liées aux élections américaines dont le contenu provient de piratage (Agence France-Presse [AFP] 2020b).

Au plan global, l'emploi de l'intelligence artificielle peut avoir une portée majeure, non seulement sur les citoyens, mais sur les régimes politiques eux-mêmes. Le spécialiste de neuroscience Nicholas Wright suggère que la démocratie libérale sera remplacée par des gouvernements qui assoiront leur autorité et leur contrôle des citoyens moyennant la technologie (Wright 2019; Harari [2018] parle de « dictature numérique »; voir aussi Wright 2020 sur le « défi autoritaire »). Un exemple d'un tel mode d'« autoritarisme numérique » se retrouve en Chine, suggère Wright. Il est vrai que l'emploi de la technologie par le régime politique chinois génère de nombreuses interrogations dans la littérature en éthique de l'IA, par exemple en ce qui concerne le contrôle serré de la navigation Internet. Également, il est allégué que le gouvernement de Pékin a recours à des outils de reconnaissance faciale pour trouver et emprisonner sa population ouïghoure (Smellie 2019). Ce type de phénomène n'est pas sans rappeler la société de surveillance dont parlait Michel Foucault (1993), en évoquant le panoptique imaginé par Jeremy Bentham en 1791, qui implique que les prisonniers soient surveillés sans le savoir (Queffelec 2019; Ropert 2014).

Le système de crédit social chinois est également mis en question à plusieurs égards, par exemple au niveau de sa transparence dans les rétributions, bonnes ou mauvaises (Engelmann et al., 2019). La manipulation du comportement et, éventuellement, de la manière de penser, par le contrôle des actions, est une des critiques adressées à un tel système (Wright 2019). D'autres soutiennent que beaucoup de désinformation circule au sujet du crédit social chinois (ÓhÉigeartaigh et al. 2020, 8). C'est en définitive la démocratie qui souffrirait de la généralisation de certaines pratiques permises par les avancées en IA. Harari soutient que la démocratie pourrait

se dégrader en système dictatorial si « le pouvoir de traiter l'information et de prendre des décisions » continue de se centraliser entre quelques entreprises, dans les mains d'une élite, et grâce à l'IA (Harari 2018, §28-29, 38, 44).

Le portrait dressé est certes sinistre. Il importe de mentionner que si l'IA, mal employée, présente tous ces risques, elle peut être utilisée pour les détecter et les contrer. C'est le cas de technologies comme les « *deepfakes* », des fichiers « […] audio ou vidéo falsifiés ou fabriqués dans le but de tromper nos sens » [Traduction libre] (Manheim et Kaplan 2019, 148; voir aussi *Ibid*., 136; Abel 2019; Gibert 2019, §7; Business Wire 2019). L'agence américaine DARPA (Defense Advanced Research Project Agency) a par exemple mis de l'avant un réseau qui était tour à tour attaqué et défendu par des agents d'intelligence artificielle. Ce système fonctionnait de manière autonome (Calo 2017, 16). De plus, le DARPA a lancé le projet « Explainable Artificial Intelligence (XAI) », qui vise à ce qu'en apprentissage machine, les modèles produits soient plus facilement « explicables », par une forme de dialogue avec les usagers. Le programme vise également la constitution d'une boîte à outils pour favoriser l'explicabilité des SIA dans les domaines commercial et militaire (Turek s.d.).

La question politique qui sous-tend les cas de figure évoqués est celle de la place de l'État dans la régulation des SIA. D'une part, si l'État cherche à tirer avantage de ces technologies pour accumuler des données et influencer sa population dans une direction partisane, ou s'il se fait le partenaire commercial des grandes entreprises gérant nos transactions Web, il y a un problème de manipulation du comportement, voire de contrôle. D'autre part, face aux situations de monopole et de monopsone sur le marché de données (Hart 2019) d'aucuns appellent l'État à intervenir pour encadrer et réguler de manière efficace et non purement formelle ou symbolique. Cela dit, dans son intervention même, l'État doit veiller à ce qu'il n'encourage pas la judiciarisation du politique en laissant aux corps administratifs le soin de gérer ces enjeux sans que le gouvernement au pouvoir ne comprenne bien de quoi il en retourne et n'abandonne à la sphère juridique ce qui devrait être aussi une problématique politique (Crawford 2019, §8, 10).

g) Les enjeux économiques de l'automatisation et les répercussions sur l'emploi

Max Tegmark, co-fondateur du Future of Life Institute, suggère que dans un futur plus ou moins rapproché, les machines pourraient remplacer les humains sur le marché du travail (2017, 133). Ce type de considération apparaît lorsque l'on considère le nouveau printemps de l'IA comme la quatrième révolution industrielle (Schwab 2017). Alors que les révolutions industrielles précédentes ont touché l'automatisation du travail physique, c'est maintenant le travail mental qui serait pris en charge par des systèmes IA. La quatrième révolution industrielle serait caractérisée par

> […] une série de nouvelles technologies qui fusionnent les mondes physique, numérique et biologique, qui ont une incidence sur toutes les disciplines, les économies et les industries, et remettent même en question les idées sur ce que signifie être humain » [Traduction libre] (World Economic Forum s.d.e., § 3).

L'ingénieur et économiste Klaus Schwab, fondateur du Forum économique mondial, a spécifié que ce qui serait propre à la quatrième révolution industrielle serait sa vélocité, son ampleur et sa profondeur, en plus de sa portée (Schwab 2017, 2-3), constituant ainsi une véritable « transformation de l'humanité » (2017, 1). C'est un indubitable « changement de paradigme » qui est en voie de se produire et d'influencer, entre autres, le monde du travail (2017, 2). En effet, certains prévoient que de nouvelles compétences seront à développer puisqu'elles seront valorisées sur un marché du travail en transformation (Susskind et Susskind 2015, 1; Lamb et Doyle 2018 et McKinsey Global Institute 2017, dans Dutton 2018c). C'est dans ce sens qu'abonde Tegmark lorsqu'il pense aux enfants d'aujourd'hui, qui devraient selon lui choisir des champs de spécialisation où les machines ne sont pas performantes ou au sein desquels on risque moins d'automatiser les tâches humaines (Tegmark 2017, 121-122, 133; voir aussi Martin 2018, §12).

Si l'on suit le paradoxe énoncé par le professeur de robotique Hans Moravec dans les années 1980, les machines, quoique très rapides dans certaines fonctions de calcul ou de stratégies où elles dépassent l'humain, ont beaucoup de mal dans les opérations impliquant la mobilité dans l'espace, la perception et plusieurs compétences relevant du sens commun (dvanw 2010, s.p.; Piper 2018, §17). Néanmoins, même dans les tâches où l'IA peut être plus efficace que l'être humain,

comme dans la détection radiographique de tumeurs ou de cancers, ce dernier est tout de même appelé à entraîner ces systèmes (Martin 2018, §2). De même, l'IA se nourrissant de l'« intelligence collective », on pourrait facilement justifier le besoin de la participation active des humains dans leurs opérations (Hart 2019).

L'enjeu de la redistribution de la richesse créée par l'automatisation de certaines tâches est corollaire aux préoccupations sur l'IA et l'emploi (McAfee et Brynjolsson 2016; Tegmark 2017, 133; Martin 2018; Hart 2019). Cependant, si certains pensent que les SIA seront à l'origine d'inégalités accrues au plan économique, d'autres soutiennent que les institutions en place sont celles qui tendent à perpétuer de tels écarts (Paul 2018, dans Dutton 2018c). De plus, il ne faut pas oublier que les SIA ne fonctionnent pas actuellement de manière entièrement autonome. De nombreux êtres humains effectuent en réalité du travail que l'on croit accompli par des machines, tels des « Turcs mécaniques », sans pour autant que leur travail soit reconnu ou valorisé en tant que tel (Hart 2019; Sproule 2018, §21). Plus encore, de nouvelles combinaisons de travail automatisé et de travail humain sont effectuées, changeant selon le contexte, et peuvent constituer des modèles pour la transformation du monde de l'emploi dans les années à venir, selon les fluctuations du marché et les ressources disponibles (Shestakofsky 2017). En outre, la question du « risque » auquel font face les travailleurs est mise de l'avant par Friedemann Bieber et Jakob Moggia (2020), pour qui l'avènement de l'économie « gig » implique nécessairement une augmentation de la précarité de l'emploi.

La question de l'obsolescence de l'être humain constitue une inquiétude concernant le futur de l'IA (McAleer 2018; Harari 2018, §6, 15; Yampolskiy 2019, §11; Totschnig 2019, 918; RodriguezRamos 2020, §2). À titre d'exemple, en 2018, 73 % d'Américains affirmaient être convaincus que l'IA allait créer moins d'emploi qu'elle n'en ferait perdre (Martin 2018, §1), donnant l'impression que l'être humain devient inutile. Devant ce type de conjectures, de nombreux pays ont adopté des stratégies industrielles et économiques face au développement de l'IA, distinctes dans certains cas des positionnements éthiques officiels. Si les transformations du monde de l'emploi peuvent être régulées, jusqu'à un certain point, par des investissements et la valorisation de certaines tâches, il n'en demeure pas moins que la question même du lien entre l'humain et le travail est soulevée. Si l'IA en venait à automatiser la plupart des emplois, peut-être

le citoyen moyen pourrait-il donc se consacrer à une vie de loisirs, comprise au sens de la culture grecque ancienne (Totschnig 2019, 919). Floridi et al. soutiennent quant à eux que ce n'est pas l'obsolescence de certaines compétences qui pose véritablement problème, ni leur remplacement par d'autres, mais la rapidité de ce changement et les inégalités qui en découlent (2018a, 3-4). Harari estime pour sa part que le tableau pourrait s'avérer plus sombre qu'on le pense, puisque

> d'ici 2050, une classe inutile pourrait émerger, résultat non d'une pénurie d'emplois ou d'un manque d'éducation pertinente, mais aussi d'une endurance mentale qui soit insuffisante pour continuer à acquérir de nouvelles compétences [Traduction libre] (Harari 2018, §20).

Une autre question se poserait alors, à savoir celle du sens et de la valeur de l'humain, qu'il ne trouverait plus dans les occupations et les productions qui étaient son pain quotidien (Tegmark 2017, 128-129; Hart 2019). Ni Tegmark ni Totschnig ne pensent que ce scénario est problématique, puisque les mêmes états de bien-être et de sentiment d'autoréalisation pourraient générés par d'autres moyens (par exemple par la contemplation de l'avènement de ce nouveau monde) et le bien-être de tous garanti d'une manière ou d'une autre (Tegmark 2017, 129; Totschnig 2019, 919). Ce pourrait être, moyennant la garantie d'un revenu de base universel, ce qui permettrait à chaque citoyen de se consacrer à des occupations que l'on juge actuellement « inutiles » en termes d'efficacité et de rendement, comme la création artistique. Les fonds pour cette redistribution proviendraient des impôts que les compagnies, employant des robots au lieu d'êtres humains, paieraient en contrepartie des salaires non versés (Hart 2019). De leur côté, Bieber et Moggia estiment que leur « principe de couverture inversée » (*Principle of Inverse Coverage* [PIC]), qui comprend une taxe forçant les employeurs à compenser les effets délétères de leurs façons de fonctionner, qui elle financerait un fonds pour assurer socialement les travailleurs de l'économie « gig », serait préférable au revenu de base universel (2020, 18-20).

Ce qui revient fréquemment dans la littérature, tous thèmes confondus, est le désir que l'IA soit centrée sur l'humain : autrement dit, que la technologie bénéficie à l'être humain, sans que ce dernier devienne une ressource à exploiter, ou un instrument en vue de l'atteinte d'autres objectifs. La cohabitation entre humains et robots est une préoccupation qui apparaît souvent aux côtés des questions touchant à la potentielle supériorité des machines sur les humains. Il s'agit de deux sous-thèmes de l'éthique de l'intelligence artificielle que je compte explorer à présent.

h) Les interactions entre humains et machines dans une nouvelle société technologique

Bien qu'il soit difficile de prévoir dans quelle mesure les êtres humains continueront d'avoir à cohabiter avec des machines opérant grâce à des SIA, plusieurs penseurs se sont penchés sur des modalités de vivre ensemble dans une société transformée par la technologie. Il s'agit de se demander « […] quel rôle voulons-nous que ce type de technologie joue dans la société actuelle? » [Traduction libre] (Sproule 2018, §24). Des questions aussi concrètes que les soins de santé prodigués par des robots (Millar 2016; Coeckelbergh et al. 2018) peuvent exemplifier de telles réflexions. Abhishek Gupta, fondateur du Montreal AI Ethics Institute (MAIEI) en 2018, de concert avec Mirka Snyder Caron (2019), propose un contrat social pour l'adoption de l'IA par la société. L'IA doit être présentée comme ayant des « finalités socialement acceptables », tout en tenant compte du caractère mouvant des individus et des valeurs au sein d'une société donnée (2019, 2). Si l'élimination de tout risque est difficile lorsqu'il est question de développer une technologie, il faut voir jusqu'à quel point une société donnée est prête à faire face aux dangers qu'elle peut poser et quels compromis elle est disposée à accepter (Snyder-Caron et Gupta 2019, 2-3). C'est ainsi qu'ils proposent que l'IA combine une finalité socialement acceptable à une méthode responsable, des retombées bénéfiques au plan social, dans le contexte où la société est consciente des risques impliqués (2019, 1).

La philosophie politique fournit quelques outils pour penser ce vivre ensemble dans le contexte d'émergence de technologies de rupture, en tenant pour acquis que « la technologie n'est pas neutre » (Sproule 2018, § 2). On le voit avec une proposition de « contrat social », notion bien familière à la pensée politique. Dès 1980, le professeur en sciences sociales Langdon Winner suggérait que certaines infrastructures étaient politiques en elles-mêmes, d'autres selon le contexte (Winner 1980, 134-135). Il affirme en effet que

> comprendre quelles technologies et quels contextes sont importants pour nous, et pourquoi, est une entreprise qui doit impliquer à la fois l'étude de systèmes techniques spécifiques et de leur histoire, ainsi qu'une compréhension approfondie des concepts et controverses de la théorie politique. [Traduction libre] (Winner 1980, 135)

En suivant cette logique, on entre dans la sphère des régulations éthiques de l'IA par les décideurs politiques. Puisque la thèse porte directement sur cette question et présente une analyse d'un échantillon de documents de positionnement à ce sujet, je vais me contenter de mentionner quelques « projets de société » qui ont été mis de l'avant, dans le monde académique, pour une cohabitation positive entre l'humain et l'IA.

Les douze experts du Comité scientifique AI4People, présidé par Luciano Floridi, estiment qu'une « bonne société de l'IA » doit être consciente des potentialités et risques de l'IA (Floridi et al. 2018a, 1-2). C'est sur la base des principes traditionnels de bioéthique (bienfaisance, non-malfaisance, justice et respect de l'autonomie) (Beauchamp et Childress 2001) qu'ils proposent de bâtir un cadre éthique, en y ajoutant celui d'explicabilité, qui inclut à la fois l'intelligibilité et la responsabilité (Floridi et al. 2018a, 8,12). L'IA « pour le bien social » (« AI for Social Good [AI4SG] ») doit procéder à une forme d'équilibrage de ces cinq principes en eux-mêmes et entre eux, selon les modalités du contexte et l'inclusion de plusieurs points de vue (Floridi et al. 2020, 19, 21).

Cette sensibilité au contexte est très présente dans la proposition de Vallor, pour qui « un avenir qui vaut la peine d'être désiré » passe nécessairement par le développement et l'acquisition — par les êtres humains — de vertus « technomorales » et de la sagesse technomorale, qui impliquent une capacité d'adaptation aux diverses situations parfois imprévisibles (Vallor 2016, 99, 50, 35, 118, 27). On assisterait au phénomène de la « convergence technologique » de la nanotechnologie, la biotechnologie, la science cognitive et les technologies informatiques, qui impliquerait leur fusion « […] d'une manière qui amplifie considérablement leur portée et leur pouvoir de modifier les vies et les institutions, tout en amplifiant également la complexité et l'imprévisibilité de l'évolution technosociale » [Traduction libre] (Vallor 2016, 27). Les citoyens devraient conséquemment chercher à acquérir des vertus, pour vivre « la bonne vie », sans pour autant le faire indépendamment du bien des autres, ni estimer que les systèmes technologiques seuls seront porteurs d'une amélioration des conditions de vie (Vallor 2016, 19, 252). L'utilitarisme et l'éthique déontologique ont eu tendance à placer un poids indu sur les épaules des agents en leur demandant l'impartialité « […] en pesant les intérêts divergents des étrangers et de leurs proches.

De telles considérations amènent à conclure que ces récits s'éloignent trop des intuitions morales communes » [Traduction libre] (Vallor 2016, 24).

Si l'on veut optimiser les interactions entre l'humain et la technologie dans la société actuelle et à venir, il convient de ne pas concevoir l'éthique de l'IA comme une démarche manichéenne, fixe, mais plutôt comme une recherche ouverte qui se nourrit de possibilités (Vallor 2016, 28). Qui parle du futur et de possibilités dans le domaine de l'intelligence artificielle ne peut manquer de faire allusion au « risque existentiel », qui pour quelques penseurs est un enjeu très sérieux (Bostrom 2014, 256). Cependant, certains chercheurs sont, au mieux, sceptiques quant à la possibilité que l'IA puisse égaler, voire dépasser l'humain en raison de l'acquisition de la conscience (par exemple, voir l'exercice de la « chambre chinoise » de John Searle [1980], ou encore les objections de Hubert Dreyfus [1999, 2007]; Totschnig 2019, 908).

Le professeur de sciences cognitives à l'Université Carleton, Jim Davies, exhibe de tels doutes en critiquant « […] l'attention déplacée à la possibilité que cette technologie puisse développer la conscience », et qu'il est fort peu probable qu'« à un moment donné, un logiciel va "se réveiller", faire passer ses désirs avant les nôtres et menacer l'existence de l'humanité » [Traduction libre] (Davies 2016, s.p.; voir également Harari 2018, §22). À son avis, ce n'est pas la conscience machine qui générerait des risques existentiels : une machine superintelligente sans conscience et donc, sans empathie pour la freiner à faire le mal, serait bien plus effrayante qu'une machine consciente (Davies 2016, s.p.). Des chercheurs en IA ont, par ailleurs mis en garde contre l'idée selon laquelle une IAG serait « bienveillante par défaut » (Piper 2018, §82). D'autres ont fait exactement le contraire, en soutenant que de tenir pour acquis que l'IA sera malveillante est un mythe à déboulonner (Tegmark 2017, 41).

Quelques chercheurs estiment que ce genre de considérations nous éloigne des enjeux éthiques importants dans l'immédiat (Davies 2016; Calo 2017, 25; Maclure et Saint-Pierre 2018). Vallor suggère que

> ceux qui prédisent une « ascension imminente des robots » ou une « singularité de l'IA », dans laquelle des êtres artificiellement intelligents décident de se passer de l'humanité ou de nous asservir, constituent à [son] avis une distraction inutile face aux dangers bien plus plausibles, mais moins cinématographiques de l'intelligence

> artificielle. Il s'agit principalement d'interactions inattendues entre des personnes et des systèmes logiciels qui ne sont pas assez intelligents pour éviter de faire des ravages sur des institutions humaines complexes, plutôt que de robots dominateurs dotés d'une « superintelligence » qui éclipse la nôtre. [Traduction libre] (Vallor 2016, 250)

Carina Prunkl et Jess Whittlestone (2020) réagissent à cette diversité de points de vue sur la pertinence et l'urgence des problèmes éthiques en IA en examinant la distinction entre le court et le long terme (voir par exemple Müller 2020, §2.10.2). Si l'on voulait brosser un tableau très sommaire de l'éthique de l'IA, on pourrait diviser les chercheurs en deux communautés : les « présentistes » et les « futuristes », qui estiment chacun que leurs préoccupations sont les plus pressantes actuellement (Prunkl et Whittlestone 2020, 2). Néanmoins, cette façon de camper le débat est réductrice, et Prunkl et Whittlestone proposent de situer les enjeux autour de quatre axes : les capabilités, les effets, la certitude et l'extrémité (2020, 3). Ce dernier axe représente bien, en raison de sa préoccupation pour les incidences de l'IA à très grande échelle, la sous-section vers laquelle j'aimerais diriger mon lecteur à présent.

i) La singularité, le risque existentiel et les agents moraux artificiels

Le paradoxe de Fermi, qui remonte à 1950, indique que selon toute apparence, il n'existe pas d'autres êtres intelligents dans l'univers, alors que des planètes présentent des conditions suffisantes pour la vie. Une réponse hypothétique serait qu'une civilisation d'êtres intelligents se serait autodétruite en poussant trop loin la technologie (que ce soit avec des armes ou non) ou en gaspillant leurs ressources (Vallor 2016, 251; RodriguezRamos 2020, §2). Il s'agit d'un risque dont sont conscients les chercheurs et penseurs qui centrent leurs travaux sur les agents moraux artificiels et une possible « explosion de l'intelligence » artificielle.

Martin Gibert rapporte qu'il y aurait 50 % de chance que, d'ici 2050, une IA équivalente à l'intelligence humaine soit développée, du moins selon « de nombreux experts » (Gibert 2020, s.p.). Le « saint Graal » des chercheurs en IA, et ce, depuis les touts débuts, est certainement l'intelligence artificielle générale ou généralisée (IAG) (« Artificial General Intelligence (AGI) ») (Boden 2018, 18) ou encore l'IA à l'échelle humaine (« Human-Level AI (HLAI) ») (Russell et

Norvig 2010, 27)[8]. L'IAG serait « la vraie IA », selon Bostrom et Yudkowsky (2014, 318). Pendant les « hivers de l'IA », la recherche sur l'IAG subit une perte de popularité importante (Russell et Norvig 2010, 24). Toutefois, depuis le début du 21e siècle, on assisterait à une résurgence (Boden 2018, 19).

Il faut établir une distinction entre l'IA de niveau humain et celle qui pourrait dépasser l'humain; la « superintelligence », ou « explosion d'intelligence » selon l'expression proposée par Nick Bostrom (2014; voir aussi Totschnig 2019, 907). Il s'agit de la « vie 3.0 » (la vie technologique) qui, à la différence de la vie 1.0 et de la vie 2.0, possède la capacité de « concevoir son matériel et ses logiciels » (Tegmark 2017, 39). Cette explosion d'intelligence, Tegmark la conçoit comme une forme de singularité. La « singularité » (Kurzweil 2006) comme terme n'est pas reprise par Bostrom, en raison de ses connotations négatives (2014, 2). Il s'agit d'un concept emprunté aux mathématiques qui indique qu'une valeur donnée est infinie. La physique a repris le terme de singularité pour signifier une gravité « infinie » (ou presque infinie, puisque la physique quantique ne permet pas les valeurs infinies) : la meilleure illustration est celle d'un trou noir. Appliquée à l'intelligence artificielle toutefois, la singularité ne signifie pas l'infinité de l'intelligence, mais plutôt de « vastes niveaux » (qui nous sembleraient à nous, humains, infinis) (Kurzweil 2006, 487, 485, 486).

Ce qui est particulier à l'argumentaire de Kurzweil est sa croyance en la possibilité que l'humain puisse lui-même atteindre une forme de superintelligence en devenant graduellement un cyborg immortel qui aurait le contrôle de l'univers (Totschnig 2019, 918; Kurzweil 2006, 486-487; voir également Tegmark 2017, 48; RodriguezRamos 2020, §6). Kurzweil, directeur de l'ingénierie chez Google, a par ailleurs créé son propre mouvement idéologique, le « singularitarianisme » (RodriguezRamos 2020, §5), et soutient que la singularité pourrait survenir autour de 2045 (Cuthbertson 2020). La base de son argumentaire est la loi de Moore, selon laquelle, depuis les années 1970, la puissance de calcul des ordinateurs connaît une croissance exponentielle tous les

[8] Bostrom et Yudkowsky ne tracent pas une équivalence entre l'IAG et l'intelligence humaine, affirmant que cette adéquation est « discutable » (2014, 318), mais Tegmark entend l'intelligence humaine, l'IAG et l'IA dite « forte » comme des synonymes (2017, 39). Müller, quant à lui, indique qu'une IA forte et l'IAG ne sont pas la même chose (2020, §2.10). C'est Searle qui a traditionnellement distingué entre une IA « faible » ou « prudente », et une IA forte, qu'il entend comme équivalente à l'intelligence humaine (1980).

deux ans (Müller, 2020, §2.10; RodriguezRamos 2020, §7). Elon Musk, quant à lui, soutient que la singularité pourrait arriver vers 2025, et propose, en préparation, la puce Neuralink qui, une fois implantée dans le cerveau, le connecte aux IA de manière à ce que nous puissions devenir compétitifs avec elle (Cuthbertson 2020). Tout cela étant mentionné, une des choses que l'on reproche à Kurzweil est de tracer une adéquation entre les notions d'intelligence et de puissance de calcul (Müller 2020, §2.10).

À l'origine, l'idée d'explosion de l'intelligence a été proposée par un collègue d'Alan Turing lors du décryptage de codes pendant la Seconde Guerre mondiale, Irving John Good (1966). Good pensait qu'une « machine ultraintelligente » avait le potentiel de devenir réalité et de concevoir d'elle-même d'autres machines bien plus avancées que les êtres humains (Bostrom 2014, 4; Piper 2018, §31; Yampolskiy 2019, §8). Selon Bostrom (et Tegmark également), l'avènement de la superintelligence ne tardera pas beaucoup après celle de l'intelligence artificielle de niveau humain. Plus important encore, ses conséquences, explique-t-il, seront soit très heureuses, soit très malheureuses, et c'est ce qui constitue, en fin de compte, le risque existentiel pour l'humanité (Bostrom 2014, 20). En effet, explique Bostrom, nous sommes comme de petits enfants qui jouent avec une bombe à retardement et notre « priorité morale » est de nous pencher sur la mitigation de ce risque (2014, 259, 260). En effet, selon la « thèse de l'orthogonalité » mise de l'avant par Bostrom, l'IA pourrait logiquement ne chercher qu'à optimiser les tâches qu'on lui confie, sans se préoccuper des conséquences des mesures prises pour ce faire (Gibert 2020, s.p.). En effet, selon cette thèse, les sphères de la moralité et de la rationalité sont en réalité complètement distinctes l'une de l'autre (Müller 2020, §2.10.1).

Le désormais célèbre exemple des trombones, donné par Bostrom, peut venir à l'esprit (Bostrom, s.d. §13; Gibert 2020, s.p.; Totschnig 2019, 918; Müller 2020, §2.10.3), ou encore celui des fourmis installées dans un secteur à inonder pour y bâtir un barrage hydroélectrique, évoqué par Stephen Hawking (Piper 2018, §34). C'est la raison pour laquelle l'éthique de l'IA, si elle porte sur l'IAG, différera nécessairement des autres champs d'éthique « appliqués à des technologies non cognitives » (Bostrom et Yudkowsky 2014, 320). D'un autre avis, Yampolskiy estime quant à lui que la recherche visant l'intelligence artificielle générale « devrait être considérée comme contraire à l'éthique » [Traduction libre] (2019, §10-12), en raison des risques existentiels associés

à cette possibilité. Il est bien connu à présent que des figures comme Hawking ou encore Musk ont mis en garde contre les dangers que présente l'IA d'annihiler l'humanité, constituant ainsi « sa plus grande menace existentielle » (Piper 2018, §1). En effet, certains chercheurs redoutent qu'

> une IA intelligente pourrait prédire que nous voudrions l'éteindre si elle nous rendait nerveux. Elle s'efforcerait donc de ne pas nous rendre nerveux, car cela ne l'aiderait pas à atteindre ses objectifs. Si on lui demande quelles sont ses intentions, ou sur quoi elle travaille, elle essaiera d'évaluer quelles sont les réponses les moins susceptibles de la faire s'éteindre et répondra par ces réponses. Si elle n'était pas assez compétente pour le faire, elle pourrait prétendre être encore plus bête qu'elle ne l'est — anticipant que les chercheurs lui donneraient plus de temps, de ressources informatiques et de données de formation [Traduction libre] (Piper 2018, §59).

Il y aurait donc des risques importants à tenir un SIA connecté à Internet, en raison de l'abondance des données disponibles et c'est la raison pour laquelle Yampolskiy propose, par mesure de sûreté, de « confiner » les SIA, dans le but de les empêcher d'échanger de l'information avec un environnement qui leur serait extérieur, ou avec lequel ils ne seraient pas autorisés à le faire (Yampolskiy 2019, §4; Piper 2018, §62-63).

Yampolskiy pense qu'une façon d'empêcher de tels scénarios catastrophiques est, en plus de confiner les systèmes d'IA, de les protéger avec un protocole comprenant une question « sécuritaire ». À la manière du test de Turing, une telle question ne pourrait se voir répondre que par un humain (Yampolskiy 2019, §5). À ceux qui se préoccupent de comment « contrôler » une superintelligence, Wolhart Totschnig (2019) répond que la bonne question à se poser est celle de la cohabitation avec cette dernière, qui est essentiellement politique. Il n'est pas d'accord avec la supposition de Kurzweil, selon laquelle la cohabitation sera pacifique : les coûts pour obtenir les nouvelles technologies créeront de vives tensions, voire une élite qui pourra envisager une forme de vie éternelle, tandis que les autres ne le pourront pas (Totschnig 2019, 908, 918). Nick Bostrom, Allan Dafoe, Carrick Flynn et Matthew S Liao proposent, pour les décideurs politiques nationaux et internationaux qui auront à faire face à l'IAG, une approche inspirée des champs vectoriels, question d'orienter les politiques normatives (2020, 4). Les « qualités désirables » de ces politiques émergeront si l'on tient compte du processus, de la population, de la répartition (ils mentionnent l'idée d'un voile d'ignorance [11] et de l'efficacité [2020, 4-5]).

Tout un autre pan de la littérature concerne les agents moraux artificiels, ou, du moins, l'attribution d'un statut moral à un agent artificiel. Bostrom, qui admet que l'on puisse douter du statut moral de personnes humaines atteintes de démence sévère, pense que des machines pourraient se voir attribuer un statut moral (Bostrom et Yudkowsky 2014, 321). C'est sur la base de la « sapience », par rapport à la « sentience » que Bostrom et Yudkowsky proposent de distinguer les humains des animaux, même s'ils affirment par ailleurs que de jeunes enfants manquent de sapience et que de « grands singes » en possèdent quelques attributs. Une IA pourrait, en raison de la « sentience », se voir attribuer un statut moral partiel et, si elle possède une sapience similaire à celle d'un « être humain normal et adulte », alors elle aurait un statut moral plein et entier. Ne pas le reconnaître serait une faute équivalant au racisme, disent-ils (2014, 322).[9] En effet, « […] il n'y a aucune différence morale entre un être fait de silicium ou de carbone, ou si son cerveau utilise des semi-conducteurs ou des neurotransmetteurs » [Traduction libre] (Bostrom & Yudkowsky 2014, 323).

Ce qu'on appelle les « droits des robots » (« *robot rights* ») ne fait pas l'unanimité. Yampolskiy est tout à fait contre l'idée que les robots puissent se voir octroyer des droits, des privilèges, voire le statut de personne. Ce serait ouvrir la voie à leur domination sur nous (2019, §15). Si la conscience humaine pouvait être répliquée dans un agent moral artificiel, des obligations morales s'ensuivraient (Müller 2020, §2.10). En tous les cas, la question de « l'alignement des valeurs » (Müller 2020, §2.10.3) est cruciale lorsqu'il est question d'agents moraux artificiels et des potentiels risques que leur existence peut générer.

La dernière sous-section de cette revue de littérature recense un nouveau type de démarche, soit celle d'identifier et d'analyser les documents traitant de l'éthique de l'IA. Il s'agit déjà d'approches plus « méta », plutôt qu'appliquées à des enjeux précis. D'ores et déjà, le lecteur se rapproche de ce qui me préoccupe dans cette thèse, à savoir la métaéthique derrière les positionnements pour l'éthique de l'IA. À la suite de l'exposé de cette petite parcelle de la

[9] Comme dans le reste de la revue de littérature, je me contente de rapporter le propos des auteurs, quoique dans ce cas-ci en particulier, je le trouve très discutable à plusieurs niveaux, notamment la non-reconnaissance de la différence essentielle de potentiel entre un enfant et un grand singe.

littérature, je serai en mesure de présenter les questions et hypothèses qui ont orienté mon travail de recherche.

j) Les compilations de démarches éthiques

« Au début de l'année 2018, j'avais l'impression d'entendre parler d'un nouvel ensemble de principes d'IA tous les deux jours », a affirmé la professeure de droit à Harvard Jessica Fjeld (Berkham Klein Center 2020). Il est vrai que depuis la publication de la lettre ouverte du Future of Life Institute (FLI), « Research Priorities for Robust and Beneficial AI », en 2015, les directives traitant de l'éthique de l'IA se sont littéralement multipliées (Assemblée nationale et Sénat de France 2017, 11). On parle de prises de position éthiques qui émanent des entreprises du secteur privé, mais aussi d'acteurs gouvernementaux nationaux et internationaux, ainsi que divers groupes représentant la société civile. La littérature académique en éthique de l'IA contient de plus en plus d'analyses de ces types de positionnement qui, eux, n'émanent pas (pour la plupart) du monde universitaire en tant que tel.

Les chercheurs Jobin, Ienca et Vayena estiment en fait que la publication de directives éthiques pour l'IA a grimpé de 88 % après 2016 (2019, 3). Depuis, le Future of Life Institute a commencé à recenser systématiquement ces positionnements. Y étaient incluses les stratégies nationales de différents pays, de même que la variété de thèmes qui devraient être prioritaires pour les décideurs politiques (Future of Life Institute s.d.a). Le chercheur Tim Dutton a procédé à un exercice similaire de compilation des stratégies nationales en matière d'intelligence artificielle, provenant de près de trente pays (Dutton 2018b). Le professeur Alan Winfield (2019) compile également, sur son blogue personnel, les documents traitant de l'éthique de l'IA et de la robotique. Sur le Web, la plateforme « Algorithm Watch » (2020) met à jour un inventaire de toutes les directives sur l'éthique de l'IA qui ont été publiées. C'est d'ailleurs la raison pour laquelle Müller (2019) a cessé de mettre à jour sa propre compilation de directives éthiques, sur la plateforme « Philosophy & Theory of Artificial Intelligence ».

C'est une fois que la vague de publications de principes a déferlé que les chercheurs du monde académique se sont penchés sur le contenu de ces directives éthiques. Par exemple, Zeng, Lu et Huangfu (2018) ont analysé la récurrence de principes éthiques dans une cinquantaine de

directives. Hagendorff a quant à lui recensé les principes présents dans une quinzaine de documents de positionnement éthique, dans le but de vérifier ce qui était présent, ce qui manquait, tout en demeurant critique par rapport à leur potentiel d'efficacité normative (2019, 1-2). Fjeld et son équipe ont de leur côté cartographié visuellement une série de principes glanés dans des directives éthiques émanant des entreprises privées, de la société civile, des gouvernements et des multiples parties prenantes (Fjeld et al., 2019, 3-6). Un peu comme l'étude de Hagendorff, l'idée était de déterminer les recoupements et la portée normative des principes, dans le but de publier un livre blanc sur l'éthique de l'IA.

L'équipe de chercheurs précédemment mentionnée de Jobin, Ienca et Vayena a, quant à elle, analysé 84 directives éthiques émanant des sphères internationale, nationales et privées. L'équipe a par exemple relevé que la plupart des documents s'adressaient à des auditoires composés de « multiples parties prenantes » (2019, 3-6). Leur démarche leur a permis de recenser les valeurs et les principes qui se recoupaient le plus dans toutes ces directives. Elles sont au nombre de onze, à savoir : « […] la transparence, la justice et l'équité, la non-malfaisance, la responsabilité, le respect de la vie privée, la bienfaisance, la liberté et l'autonomie, la confiance, la dignité, la durabilité et la solidarité » [Traduction libre] (Jobin, Ienca et Vayena 2019, 6). Les principes autour desquels une certaine « convergence » semble s'orchestrer sont la transparence, la justice et l'équité, la non-malfaisance, la responsabilité et le respect de la vie privée (Jobin, Ienca et Vayena 2019, 14). Les chercheurs ont également relevé le fait que la proportion de documents en provenance des secteurs public et privé était presque équivalente, indiquant de ce fait « […] que les défis éthiques de l'IA concernant à la fois les entités publiques et les entreprises privées » [Traduction libre] (Jobin, Ienca et Vayena 2019, 13). Le constat est différent du côté de l'équipe de Schiff et al. qui, ayant analysé un échantillon de 88 directives éthiques concernant l'IA, estime que c'est le secteur privé qui a produit le plus de documents en lien avec l'intelligence artificielle (Schiff et al. 2020, 154).

Néanmoins, des problèmes se posent de manière plus concrète, notamment

> […] quant aux principes éthiques à privilégier, à la manière de résoudre les conflits entre les principes éthiques, à la personne chargée de la surveillance éthique de l'IA et à la manière dont les chercheurs et les institutions peuvent se conformer aux directives qui en découlent. Ces résultats suggèrent l'existence d'une lacune au niveau de la

> formulation des principes et de leur mise en œuvre dans la pratique qui peut difficilement être résolue par une expertise technique ou des approches « descendantes » (« *top-down* ») [Traduction libre] (Jobin, Ienca et Vayena, 2019, 14).

Puis, dans le but d'opérationnaliser les principes éthiques qui sont énoncés dans la majorité de ces documents, une équipe de chercheurs de Microsoft a mis au point une liste de contrôle de l'équité pour les développeurs en IA (Madaio et al. 2020). En suivant cette liste de contrôle, il deviendrait plus facile pour les développeurs de s'assurer que les systèmes mis au point seraient conformes au principe (récurrent) d'équité. Une autre équipe de chercheurs a développé un cadre permettant l'audit algorithmique des SIA selon les valeurs et les principes de l'entreprise qui les produit (Raji et al. 2020, 33). Puis, la fondatrice de l'organisme « AI for Peace », Branka Panic (2020), a publié un « guide non technique » sur l'IA à l'intention des décideurs politiques, dans lequel elle expose les bases de l'intelligence artificielle, les défis éthiques qu'elle pose, ainsi que le rôle des décideurs politiques face à ces enjeux. Elle y mentionne également quelques directives éthiques publiées lors des dernières années, comme celles du IEEE ou de DeepMind (Panic 2020, n° 18). En somme, si la publication de directives éthiques relatives à l'IA est encore une pratique répandue dans plusieurs sphères de la société, une nouvelle préoccupation a fait surface, soit celle de traduire ces principes et valeurs dans la pratique.

2. Les questions et les hypothèses qui sous-tendent cette thèse

De l'éventail d'enjeux dont il est question dans le champ de l'éthique de l'intelligence artificielle, le lecteur a pu constater que la grande majorité traite d'éthique appliquée, à des problématiques bien concrètes. Je l'ai mentionné d'entrée de jeu, mais cette thèse de philosophie politique porte surtout sur la métaéthique des directives concernant l'intelligence artificielle. De manière analogue aux auteurs de la dixième section de cette revue de littérature, je me suis moi aussi employée à recenser des directives éthiques sur l'IA, les catégorisant par la suite selon leur provenance. La première question de recherche qui a animé ma démarche était donc la suivante : *Quelle(s) tradition(s) ou approche(s) éthique(s) informent les directives éthiques sur l'intelligence artificielle parues entre janvier 2016 et janvier 2020?* À ma connaissance, une telle démarche n'a été effectuée que par Charles Ess (2019), qui a exploré la métaéthique de quelques directives sur l'IA dans le contexte de sa réflexion sur le respect de la vie privée à travers les cultures. Toutefois,

aucune analyse que je connaisse n'a inclus le pluralisme des valeurs, comme je le caractériserai au chapitre trois, dans les approches éthiques potentiellement à l'œuvre dans cette littérature. L'hypothèse que j'avais avant la compilation et l'analyse d'un nombre appréciable de ces documents était que l'éthique procédurale et l'éthique pluraliste des valeurs seraient majoritaires dans les documents.

Il m'apparaît que la métaéthique est un angle mort de la prolifération de réflexions portant sur l'éthique de l'IA — ou, plus précisément, une métaéthique « complète ». À mon sens, lorsque les chercheurs parlent des traditions éthiques à l'œuvre dans les esprits, non seulement, comme on l'a vu plus tôt, ils tendent à ne pas tenir compte de traditions éthiques autres qu'occidentales (à l'exception notable de Vallor), mais encore ils oublient de nommer le pluralisme des valeurs, alors qu'il semble omniprésent dans les démarches. Un tel constat appelle une seconde question de recherche. Alors que la première tend à produire une réponse plus descriptive, la seconde est normative : *Quelle approche éthique permet de favoriser un dialogue optimal des décideurs politiques en ce qui a trait à l'éthique de l'intelligence artificielle?* J'entends par « optimal » l'idée du « meilleur possible », qui n'est pas un meilleur absolu, mais « l'état le plus favorable » (Larousse, s.d.) selon les circonstances. (Il s'agit, bien évidemment, de répondre à cette question en incluant, dans la réflexion, le pluralisme des valeurs.) Mon hypothèse de départ était que l'éthique de la vertu, avec une influence de la philosophie herméneutique, constituerait la réponse. Le lecteur sera à même de voir l'évolution du propos dans les chapitres qui suivront. Je lui propose donc de me suivre à présent dans l'exploration des fondements des traditions éthiques à l'œuvre dans les documents, dans le but de les replacer sur un continuum s'étirant entre le monisme et le pluralisme.

SECTION 1 : FONDEMENTS

Chapitre 2 — Traditions monistes

« *Depuis que les Grecs ont inventé la logique et la géométrie, l'idée que tout raisonnement puisse être réduit à une sorte de calcul — afin que tous les arguments puissent être réglés une fois pour toutes — a fasciné la plupart des penseurs rigoureux de la tradition occidentale*. » — Hubert Dreyfus [Traduction libre] (1999, 67)

Introduction

Puisque l'objet de cette thèse est de proposer une approche des enjeux éthiques de l'IA aux décideurs politiques, il importe d'aborder, avec un certain degré de profondeur, l'éthique telle qu'elle a été et est appréhendée dans le monde occidental.[10] Le lecteur s'en doute : il ne suffit pas de choisir une théorie éthique et de l'appliquer à des enjeux, ou encore de confectionner un pot-pourri d'écoles éthiques selon les différents aspects du problème à traiter. Procéder de cette façon implique bien des angles morts. Je suggère qu'il faudrait plutôt creuser un peu plus profondément, question d'explorer les présupposés métaéthiques des approches théoriques et non théoriques à l'éthique. En effet, choisir une approche éthique, c'est aussi choisir ses prises de position métaéthiques. Il semble pertinent qu'au moment de conseiller les décideurs politiques, cette réalité soit prise en compte. Autrement, et selon les traditions invoquées, il serait possible de combiner sans le savoir des idées dont les fondements métaéthiques sont en contradiction complète. C'est la raison pour laquelle cette analyse me semble incontournable (à ce sujet, voir Ess 2009, 167-168).

Une façon de se pencher sur les postulats métaéthiques des traditions éthiques est par l'entremise de la question de l'Un et du Multiple. Dans cette optique, les différentes traditions éthiques relèveraient soit du monisme, soit du pluralisme (ou, plus rarement, d'un paradoxal mélange des deux) (Blattberg 2018, 150-151). Même s'il est plutôt récent, en philosophie morale et politique, de parler de « monisme » et de « pluralisme », les penseurs du monde occidental ont toujours été aux prises avec le problème de l'Un et du Multiple et ce, dès la Grèce antique (Apfel 2011, 2-5, 23). Dans ce chapitre, ce sont les fondements de trois traditions monistes en éthique qui

[10] Cette thèse ne traite que des traditions éthiques occidentales d'importance, en raison de leur influence dans les documents et positionnements éthiques analysés concernant l'intelligence artificielle. Il ne sera donc pas question des traditions éthiques orientales ou encore africaines, par exemple.

seront explorés, c'est-à-dire de l'éthique de la vertu, de l'utilitarisme et de l'éthique déontologique. Ces écoles ont exercé une influence importante dans le monde occidental, lors des dernières décennies (Hursthouse et Pettigrove 2018; Crisp 1996, 8). Dans le chapitre suivant, j'exposerai les fondements métaéthiques du pluralisme des valeurs. Il sera par la suite possible, dans les chapitres quatre et cinq, d'analyser des directives éthiques ciblées en vue de déterminer leurs ancrages métaéthiques.

L'éthique de la vertu, l'éthique déontologique (ou « déontologisme ») et le conséquentialisme sont des traditions éthiques qui relèvent du monisme, c'est-à-dire du postulat qu'il existe une unité fondamentale au réel, et que toutes choses peuvent être réunifiées, ramenées à leur état « normal » ou « préfragmentaire ». Il s'agirait d'une « antique croyance » selon laquelle « […] toutes les valeurs positives auxquelles les hommes sont attachés seraient finalement compatibles et peut-être même interdépendantes » (Berlin 1988, 213). En effet, pour les monistes, trois choses sont centrales :

> en premier lieu que, comme dans les sciences, toutes les vraies questions doivent avoir une seule vraie réponse, et une seulement — toutes les autres étant nécessairement des erreurs; en deuxième lieu, qu'il doit y avoir un chemin sûr vers la découverte de ces vérités; en troisième lieu, que les vraies réponses, lorsqu'elles sont trouvées, doivent nécessairement être compatibles entre elles et former un tout unique, car une vérité ne peut être incompatible avec une autre — que nous connaissons a priori. [Traduction libre] (Berlin 1990, 5-6)

Le monisme peut porter d'autres noms dans la littérature traitant de métaéthique. Par exemple, lorsque le professeur Charles Ess traite du pluralisme, il le comprend comme une sorte de synthèse entre ce qu'il appelle « le dogmatisme éthique » et le « relativisme éthique » (Ess 2009; Ess 2006). Le « dogmatisme éthique » renverrait en fin de compte à sa compréhension du monisme. Cependant, je ne partage pas cette interprétation. Si certains monistes peuvent présenter des traits absolutistes, voire dogmatiques, dans leur pensée, il n'en découle pas nécessairement que tout penseur moniste est dogmatique au sens péjoratif du terme, ou encore qu'il tend à l'usage de la force (Ess 2020, 556). Cela apparaît plus clairement quand on se penche sur la pensée d'un néo-Kantien comme John Rawls ou d'une néo-Aristotélicienne comme Shannon Vallor, pour qui la démarche moniste est une manière de tenir compte de la multiplicité des opinions en la réunissant sous l'égide d'une théorie somme toute assez englobante.

C'est souvent (mais pas toujours) au moyen d'une *théorie* éthique que les monistes expriment leur prise de position métaéthique et métaphysique qu'est l'unité du réel. Une manière de concevoir la théorie éthique serait comme « […] un ensemble de raisons et d'arguments interconnectés, explicitement et systématiquement articulés, avec un certain degré d'abstraction et de généralité, qui donne des orientations pour la pratique éthique » (Nussbaum 2000, 233-234). La théorie peut avoir tendance à se présenter comme « explicite, universelle, abstraite, systématique, complète et prédictive », et à être formulée indépendamment d'un contexte donné (Flyvbjerg 2001, 38-39). Dans cet ordre d'idées, une bonne théorie éthique serait donc claire et explicite, elle possèderait un certain degré d'abstraction et de généralité, elle serait « universalisable », elle fournirait des recommandations pour des problèmes d'ordre pratique, et elle systématiserait et élargirait les croyances en montrant comment en tester la validité (Nussbaum 2000, 234-236).

Avant de poursuivre, toutefois, une clarification s'impose. Martha Nussbaum, ici citée, fait volontiers appel à la théorie éthique : c'est pourquoi son propos est rapporté pour arriver à bien cerner le monisme. En revanche, l'ancrage métaéthique de sa pensée, prise dans son entièreté, n'est pas purement et simplement « moniste », à la différence des approches éthiques qui seront développées dans ce chapitre. Entre le monisme et le pluralisme s'étire un continuum riche sur lequel on peut se positionner de différentes manières. Certes, pour assurer la compréhension des catégories monistes et pluralistes telles que je les emploie dans cette thèse, j'en traiterai de manière assez compartimentée. Cependant, certains penseurs sont difficiles à camper et c'est le cas de Martha Nussbaum, qui peut être comprise comme une philosophe « moniste non orthodoxe ». Sa réflexion éthique, notamment dans son ouvrage *La fragilité du bien* (2016), est influencée par les tragédiens grecs anciens ainsi que par Bernard Williams, un philosophe pluraliste. Sa version du monisme est donc nuancée, en permettant de tenir compte des imperfections parfois irrémédiables du monde de la pratique et en faisant place à la tragédie, voire à la corruption, et ce, même chez une personne vertueuse (2016, 494-496, 505). Je reviendrai sur cette catégorie en détail au sixième chapitre.

On pressent déjà que, pour les monistes, la division, le conflit et les pertes sont des anormalités dans les questions éthiques et politiques. Il faut toutefois mentionner que plusieurs d'entre eux tolèrent des désaccords puisqu'ils sont inévitables. En règle générale, cependant, les

pertes sont à éviter ou encore à « réparer ». Ainsi, en éthique, plusieurs monistes seront partisans de théories éthiques qui parviennent à réconcilier, dans l'abstraction, les différends et à pallier les préjudices. Si elles n'y parviennent pas, les tenants des théories éthiques pensent tout de même, malgré les torts encourus et qu'ils reconnaissent, qu'ils ont « les mains propres », peu importe le mal qu'ils ont dû commettre ou causer en voulant atteindre une bonne finalité. Ils pensent ainsi, car ils auront agi de bonne foi et en suivant la prescription de la théorie. Ils n'auront, en fin de compte, « rien fait de mal » (Blattberg 2018, 152).

Dans cette sous-section seront présentées les trois traditions (monistes) qui sont, depuis environ trente-cinq ans, dominantes en éthique : l'éthique de la vertu, l'utilitarisme, une branche du conséquentialisme, et l'éthique déontologique. Les principaux aspects de chacune de ces théories seront mis de l'avant pour que le lecteur ait clairement en tête, lors de l'étude des positionnements éthiques concernant l'IA, quelles sont les assises de chaque approche éthique. Quelques critiques adressées à chaque tradition seront également passées en revue. Il importe de mentionner que, comme il ne s'agit pas d'une thèse de philosophie morale, mais de philosophie politique, le niveau de détail des débats entre les écoles éthiques n'est pas maximisé dans la discussion. L'idée générale est simplement de permettre au lecteur de mieux comprendre et distinguer les différents types de réponses aux défis éthiques de l'IA. Les choix éthiques que j'exposerai dans la suite de la thèse seront par conséquent plus faciles à suivre une fois ces clarifications faites.

1. L'éthique de la vertu

Le terme « vertu » tire son origine du latin « *virtus* », qui quant à lui, provient du grec « *arête* », signifiant « excellence » (Vallor 2016, 17). La vertu renvoie à « […] tout trait stable qui permet à son possesseur d'exceller dans l'accomplissement de sa fonction distinctive […] » [Traduction libre] (Vallor 2016, 17). Elle est « un excellent trait de caractère » [Traduction libre] (Hurtshouse et Pettigrove 2018, s.p.) et, en conséquence, « […] elle [fait] de quelque chose un excellent exemple de son genre » [Traduction libre] (Darwall 2003b, 2). Comme elle est une excellence, elle est valorisée en elle-même, et non seulement en raison de ses conséquences. Elle

est en quelque sorte une « capacité à faire le bien », enracinée dans la volonté. En effet, « […] si la santé et la force sont des excellences du corps, et la mémoire et la concentration de l'esprit, *c'est la volonté qui est bonne* chez un homme de vertu » [Traduction libre, je souligne] (Foot 1978, 3). Ce qui rend la vertu intéressante à s'efforcer d'acquérir, c'est qu' « […] une fois cultivée, elle mène à des choix délibérés, efficaces et raisonnés du bien » [Traduction libre] (Vallor 2016, 18). Dans son *Éthique à Nicomaque*, Aristote la définit comme un « habitus » (une bonne habitude) ou une disposition (Aristote 2014, Livre II, 4, 1105b25-30) — plus précisément des dispositions du *caractèr*e (Vallor 2016, 18).

Il est bien connu que la figure d'Aristote est fondamentale à l'éthique de la vertu. Comme le font Alasdair MacIntyre (1984, 146), Philippa Foot (1978, 1-2), Martha Nussbaum (1995) et bien d'autres, je lui accorde une place centrale dans cette section, bien que l'éthique de la vertu contemporaine ne soit pas nécessairement néo-aristotélicienne à tout coup (The Editors of Encylopaedia Britannica 2016, s.p.). En revanche, pour les penseurs néo-aristotéliciens,

> […] la théorie aristotélicienne de la vertu n'[est] pas une éthique moderne satisfaisante. […] Aucun éthicien contemporain de la vertu ne peut nier qu'il y a des problèmes, des ambiguïtés et des lacunes importantes dans le récit d'Aristote; la question de savoir s'il est possible de les modifier, de les clarifier et de les compléter sans détruire l'intégrité ou la valeur contemporaine de son cadre est un sujet de discussion permanent. [Traduction libre] (Vallor 2016, 21)

Il faut aussi mentionner que toute théorie traitant de la vertu ne mérite pas forcément l'épithète d'éthique de la vertu. En effet, le conséquentialisme et le déontologisme peuvent en traiter — ainsi que d'autres approches éthiques — sous le regard de leur approche théorique, sans pour autant constituer des exercices d'éthique de la vertu (Hursthouse et Pettigrove 2018, s.p.; Crisp 1996, 5). Ce qui distingue l'éthique de la vertu des autres théories éthiques, c'est qu'elle concerne moins ce qu'il faudrait *faire* (l'action, la procédure), que ce qu'il faudrait *être* (l'agent) (Darwall 2003b, 1; Hursthouse 2003, 185). Un autre point de divergence entre l'éthique de la vertu d'un côté, et le déontologisme et le conséquentialisme de l'autre, se trouve dans la distinction que relève Charles Taylor entre universalistes et communautariens; ou encore, entre une éthique des règles et une éthique du bien (Taylor 1994b, 23-24).

a) Le « nouveau souffle » de l'éthique de la vertu

Historiquement, le parcours de l'éthique de la vertu est singulier. Elle a vécu une sorte d'âge d'or depuis la redécouverte des œuvres d'Aristote et ce, jusqu'aux Lumières environ. Elle perd ensuite son lustre pour disparaître momentanément au 19e siècle (Hursthouse et Pettigrove 2018, s.p.). La raison est que, ayant été largement adoptée par des philosophes et théologiens chrétiens comme Saint Thomas d'Aquin, elle semblait entrer en conflit avec les idéaux des Lumières, qui cherchaient justement l'affranchissement de l'Église catholique et de ses enseignements pour se tourner vers les seules réponses de la science (Vallor 2016, 20).

Même si l'éthique de la vertu est reléguée aux oubliettes pendant quelque temps, le « langage de la moralité », lui, demeure (MacIntyre 1984, 5). C'est en 1958 qu'Elizabeth Anscombe redonne à la vertu ses lettres de noblesse, avec son essai « Modern Moral Philosophy », en critiquant du même coup l'emploi moderne de ce langage moral. Selon elle, il faut larguer les concepts d'« obligation » et de « devoir » moraux puisqu'ils appartiennent à une conception légale de l'éthique désormais caduque (Anscombe 1958, 1, 7). Procédant à une critique de l'éthique déontologique d'Emmanuel Kant comme de l'éthique conséquentialiste-utilitariste de Jeremy Bentham et John Stuart Mill, elle aborde aussi l'influence historique du christianisme dans la philosophie éthique ayant amené une conception légale que l'on trouvait aussi dans le stoïcisme ancien (Anscombe 1958, 2-5). MacIntyre s'engagera plus tard dans le sillage d'Anscombe, en suggérant que le langage moral moderne est en réalité « […] un masque pour l'expression de nos préférences personnelles […] » [Traduction libre] (MacIntyre 1984, 19), soit l'argumentaire derrière la théorie anglaise de l'émotivisme.

Cet abandon d'un langage juridique ou du « devoir » (*moral ought*) n'a toutefois pas été effectué par l'éthique conséquentialiste ni l'éthique déontologique. La raison pour laquelle Anscombe prône ce délaissement est simple : il est difficile de s'entendre sur la loi qu'il faudrait suivre pour bien agir, dans une société où les croyances sont multiples. C'est que l'éthique contemporaine doit nécessairement tenir compte d'une multiplicité des points de vue sur une question donnée. Une éthique de la vertu n'y serait pas exemptée. Pourtant, au lieu de chercher une procédure d'action dans des théories éthiques comme le déontologisme ou le conséquentialisme,

l'éthique de la vertu permet de mettre l'accent sur l'amélioration personnelle (Annas 2004, 66). De fait, il importe de savoir dépasser la simple obéissance à des règles, puisqu'il existe des occasions où la règle n'est justement pas ce qu'il faut faire (Nussbaum 2000, 238).

b) Les types de vertus

Si l'on désire saisir la conception aristotélicienne de la vertu, il est nécessaire de faire un petit détour par quelques aspects de sa philosophie anthropologique. Cette dernière a d'ailleurs été reprise et commentée par Thomas d'Aquin, plusieurs siècles plus tard. Cette courte digression permettra, je l'espère, de mieux camper l'éthique de la vertu dans ses origines. Aristote conçoit l'âme humaine comme étant divisée en deux parties, l'une rationnelle et l'autre irrationnelle. La première se subdivise de nouveau en deux sous-parties, « […] l'une par laquelle nous contemplons ces sortes d'êtres dont les principes ne peuvent être autrement qu'ils ne le sont, et l'autre par laquelle nous connaissons les choses contingentes […] » (Aristote 2014, Livre VI, 2, 1139a1-20). Autrement dit, il s'agit des « sciences théorétiques » et des « sciences poiétiques » (les sciences de la production) (Aristote 2014, Livre VI, 2, 1139a1-20). Aristote les appelle aussi les parties « scientifique » et « calculative » (2014, Livre VI, 2, 1139a1-20). Ces deux parties de l'intellect, selon le philosophe, « […] ont pour tâche la vérité » (Aristote 2014, Livre VI, 2, 1139b10-15). Elles sont aidées en cela par des vertus qui leur sont propres (Aquin s.d., question 56, article 1). Aristote établit qu'il y a cinq vertus de l'intelligence[11] : le *nous*, l'*episteme*, la *sophia*, la *phronesis* et la *techne*. Ces vertus caractérisent des applications ou fonctions particulières de l'esprit humain dans le but de « […] saisir la vérité intelligible » (Aquin 1928, 238).

Quand l'intelligence se penche sur les universels, elle fait appel à sa fonction théorique ou spéculative. Les vertus y étant associées sont le *nous* et l'*episteme*, dont la combinaison débouche sur la *sophia* ou sagesse métaphysique, la plus haute forme de sagesse selon Aristote. L'esprit humain appliqué, par contraste au spéculatif, à ce qui mène vers l'action, est subdivisé en deux

[11] J'entends « intelligence » et « raison » comme une seule et même chose, comme le fait Thomas d'Aquin : « Saint Augustin dit que le principe par lequel l'homme surpasse les animaux irrationnels c'est la raison, ou l'esprit, ou l'intelligence, ou comme on voudra l'appeler. Raison, esprit et intelligence ne sont donc qu'une même puissance » (Aquin 1928, 1a, question 79, article 8, 238).

autres catégories : ce qui est pratique, et auquel est associé la *phronesis*, et ce qui est productif, auquel est attachée la vertu de *techne*. La fonction spéculative de l'esprit humain fait que nous considérons la réalité comme un spectateur. Alors que la *techne* est « la raison droite dans les choses à produire », la *phronesis* est, selon Thomas d'Aquin, la « raison droite dans les choses à faire » (Aquin s.d., question 56, article 4). On l'a vu dans les fondements de l'éthique de la vertu, mais en ce qui a trait à l'action, dit Aristote, « […] ce qu'on fait est une fin au sens absolu, car la vie vertueuse est une fin, et le désir a cette fin pour objet » (Aristote 2014, Livre VI, 2, 1138b-1139b).

L'intelligence spéculative, de pair avec l'intelligence pratique, renvoient à une seule et même intelligence, qui diffèrent toutefois selon leurs finalités (Aquin 1928, 1a, question 79, article 11, 254-255). Ce qui caractérise l'intellect pratique est qu'il constitue « […] une faculté de mouvement, non en tant qu'il exécute, mais en tant qu'il dirige le mouvement » (Aquin 1928, 1a, question 79, article 11, 255). Son objet est « […] le bien qui peut être ordonné à l'action, considéré comme vrai. L'intellect pratique connaît effectivement la vérité, comme l'intellect spéculatif, mais il ordonne à l'action cette vérité connue » (Aquin 1928, 1a, question 79, article 11, 255-256). Suivant cela, la *phronesis* doit nécessairement être ordonnée à un bien. Aristote soutient en effet que « […] pour la partie de l'intellect pratique, son bon état consiste dans la vérité correspondant au désir, au désir correct » (Aristote 2014, Livre VI, 2, 1139a30). C'est la raison pour laquelle on parlera plus bas de la notion de « bien commun ».

On a vu qu'Aristote traite des vertus intellectuelles. Il propose aussi une autre série de vertus, qui sont morales. Il s'agit de la douceur, du courage, de la pudeur, de la tempérance, de l'indignation vertueuse, de la justice, de la libéralité, de la vérité, de l'amitié, de la dignité, de l'endurance, de la magnanimité, de la magnificence et de la sagesse (Aristote 2014, Livre II, 7, 1107 b). Alors que les vertus intellectuelles perfectionnent l'intelligence, les vertus morales confèrent une excellence au caractère. Elles s'acquièrent par ailleurs différemment :

> la vertu intellectuelle dépend dans une large mesure de l'enseignement reçu, aussi bien pour sa production que pour son accroissement; aussi a-t-elle besoin d'expérience et de temps. La vertu morale, au contraire, est le produit de l'habitude, d'où lui est venu aussi son nom […]. (Aristote 2014, Livre II, 1, 1103a15)

Les vertus morales ne sont donc pas innées. Plutôt, « […] c'est en pratiquant les actions justes que nous devenons justes, les actions modérées que nous devenons modérés, et les actions courageuses que nous devenons courageux » (Aristote 2014, Livre II, 1, 1103b1-5).

La vertu relève ainsi de l'habitude parce'elle a été forgée par des actions répétées, de manière à pouvoir être qualifiée de disposition qui engage toute la personne — sa volonté ainsi que toutes ses facultés. En effet, cette bonne habitude est « […] une disposition, bien ancrée dans son possesseur […] à remarquer, attendre, valoriser, sentir, désirer, choisir, agir et réagir de certaines manières caractéristiques » [Traduction libre] (Hursthouse et Pettigrove 2018, s.p.). Conséquemment, il existe une distinction entre la personne réellement vertueuse et la personne simplement continente, car

> il n'est pas facile de mettre ses émotions en harmonie avec la reconnaissance rationnelle de certaines raisons d'agir. […] Ceux qui sont pleinement vertueux font ce qu'ils doivent faire sans lutter contre des désirs contraires; le continent doit maîtriser un désir ou une tentation de faire autrement. [Traduction libre] (Hursthouse et Pettigrove 2018, s.p.)

McDowell abonde dans le même sens : « si quelqu'un a besoin de surmonter une inclination à agir autrement, en se faisant agir comme l'exige, disons, la tempérance du courage, alors il ne fait pas preuve de vertu, mais de (simple) continence » [Traduction libre] (2003, 125).

Dans un autre ordre d'idées, mentionnons qu'en plus de dresser une liste de vertus, Aristote établit également que certaines actions sont perverses en elles-mêmes. Il nomme par exemple la malveillance, le vol, l'adultère, l'envie et le meurtre. Dans de tels cas, il est absurde de parler de juste milieu dans la pratique (Aristote 2014, Livre II, 6, 1107a10) ou d'excellence. Ce « juste milieu » dans la pratique renvoie au fait que la vertu se situerait entre deux vices : celui de défaut et celui d'excès. Conséquemment, le « moyen dans la chose » (Aristote 2014, Livre II, 5, 1106a25-35-1106b1-5) est relatif au contexte de chaque personne et de chaque situation. On ne peut définir, une fois pour toutes et de manière abstraite, ce que sera le « moyen dans la chose » pour chaque situation, en raison des contingences qui lui sont uniques.

Pour illustrer cette réalité, Aristote donne l'exemple de la charge à soulever pour ceux qui s'entraînent au gymnase : ce qui sera considéré trop pour l'un pourrait être trop peu pour un autre

(Aristote 2014, Livre II, 5, 1106a25-35-1106b1-5). Cette sensibilité au contexte de chaque situation est aussi transposée par les néo-aristotéliciens dans des contextes plus larges. Par exemple, comme Nussbaum, Vallor soutient que la définition « épaisse » des vertus prend forme selon le contexte culturel ou politique, scientifique de chaque situation (Vallor 2016, 119). La détermination du juste milieu, dans chaque contexte, se fait au moyen de la raison pratique et, plus précisément, de la vertu de *phronesis*.

c) Le caractère moniste de l'éthique de la vertu

L'éthique de la vertu n'est évidemment pas présentée par Aristote comme une théorie au sens moderne du terme, inspirée par le modèle des théories des sciences de la nature. Néanmoins, son approche éthique peut être qualifiée de « moniste » en ce qu'elle met de l'avant la doctrine de l'unité des vertus. Cette doctrine se justifie par le fait que, comme la vertu forme le caractère de la personne vers l'excellence, la possession d'une seule vertu enrichit toute la personne. Plus encore, l'on pourrait dire que toutes les vertus « se tiennent par la main ». Dans une même situation, la combinaison de vertus pouvant être simultanément « exigées » par le contexte est infinie. Conséquemment,

> […] aucune vertu ne peut être pleinement possédée que par un possesseur de toutes les vertus, c'est-à-dire un possesseur de la vertu en général. […] Nous utilisons plutôt les concepts des vertus particulières pour marquer les similitudes et les dissemblances entre *les manifestations d'une même sensibilité qui est ce que la vertu, en général*, *est* : une capacité à reconnaître les exigences que les situations imposent à son comportement. C'est une *sensibilité unique* et complexe de ce genre que nous cherchons à inculquer lorsque nous voulons inspirer une perspective morale. [Traduction libre, je souligne] (McDowell 2003, 123-124; voir aussi Wolf 2007, 45)

On comprend donc qu'un tenant de l'éthique de la vertu présenterait la caractéristique moniste de croire que les dilemmes moraux peuvent être résolus « en principe ». De ce fait, il soutiendrait que « […] tout acte véritablement vertueux doit contribuer au bien-être de l'acteur ainsi qu'au bien commun de sa communauté politique » [Traduction libre] (Blattberg 2018, 155).

Un autre aspect moniste de l'éthique de la vertu se trouve dans son caractère téléologique, auquel souscrivent plusieurs de ses variantes — bien que pas toutes (Darwall 2003b, 2). Lorsqu'une

éthique de la vertu est dite téléologique, c'est qu'elle systématise une réflexion par rapport à une finalité identifiée. En découle une certaine unité dans la conception des vertus, qui sous-tend la doctrine de l'unité des vertus dont il a été question plus haut. Aristote affirme que « […] le Bien est ce à quoi toutes choses tendent » (2014, Livre I, 1, 1194a1-5). La vertu permet de tendre vers le bien, tout en rendant l'agent bon. En effet,

> […] toute « vertu », pour la chose dont elle est « vertu », a pour effet à la fois de mettre cette chose en *bon* état et de lui permettre de bien accomplir son œuvre propre : par exemple, la « vertu » de l'œil rend l'œil et sa fonction également parfaits, car c'est par la vertu de l'œil que la vision s'effectue en nous comme il faut. (Aristote 2014, Livre II, 5, 1106a15)

La téléologie dans l'éthique de la vertu aristotélicienne comporte la particularité que la fin n'est pas sujette à délibération : elle est plutôt « objet de souhait », tandis que « […] les moyens pour atteindre à la fin [sont] objets de délibération et de choix, [et] les actions concernant ces moyens seront faites par choix et seront volontaires […] » (Aristote 2014, Livre III, 7, 1113b1-10). La vertu concerne les moyens dans l'action. MacIntyre suit Aristote en affirmant que l'être humain — comme les êtres des autres espèces — poursuit un certain *telos*, une finalité. Ce bien consiste en ce qu'Aristote appelle l'*eudaimonia*, la bonne vie — ou encore bonheur, prospérité, béatitude. L'*eudaimonia* n'est pas une agrégation quantitative des biens de la vie ni des vertus : « la bonne vie met plutôt les biens partiels ensemble dans leur ordre propre, selon leur rang propre, pour ainsi dire » [Traduction libre] (Taylor 1994b, 25). Néanmoins, « […] quand Aristote donne pour la première fois ce nom au bien de l'homme, il laisse largement ouverte la question du contenu de l'*eudaimonia* » [Traduction libre] (MacIntyre 1984, 148). Malgré cette flexibilité dans le contenu, la conception aristotélicienne de la bonne vie implique une hiérarchisation des biens.

Deux nuances sont à prendre en compte lorsqu'il est question de vertu. Tout d'abord, la possession ou non d'une vertu n'est pas déterminée de façon manichéenne. Il serait plus exact de parler de degrés de possession de vertu (Hursthouse et Pettigrove 2018, s.p.). Ensuite, il est vrai que les vertus sont certes un chemin vers l'*eudaimonia*, mais elles n'en sont pas entièrement distinctes non plus. Il ne s'agit pas de « moyens » (au sens moderne) vers une fin. Aristote ne raisonne pas en séparant ces termes. Cette séparation des fins et des moyens est éminemment moderne et tend à renvoyer à la raison instrumentale. Il faut comprendre que les vertus, une fois acquises, font déjà partie de cette bonne vie qu'est l'*eudaimonia* (MacIntyre 1984, 148-149). C'est

ce qui permet à Vallor d'affirmer que « […] la vertu *est* l'activité de bien vivre » [Traduction libre] (2016, 19). Les vertus sont donc un chemin et presque déjà, en un sens, la destination qu'est l'*eudaimonia*.

d) La *phronesis* ou sagesse pratique

Un peu plus haut, il a été question d'une vertu en particulier qui est cruciale pour Aristote, lorsqu'il s'agit de l'exercice de la raison à une fin pratique, comme l'éthique et l'activité politique. Cette vertu intellectuelle est la *phronesis,* qui est, pour Aristote, une vertu intellectuelle. Avant de poursuivre, une précision s'impose. La *phronesis* aristotélicienne a été traduite tantôt par « prudence », tantôt par « sagacité ». Comme le lecteur le constatera dans la section normative de la thèse, je ferai référence à ma conception de la prudence comme dérivée de la *phronesis*, sans toutefois y être identique. La compréhension aristotélicienne de cette vertu renvoie à la capacité d'orienter sa flèche vers la cible du « Souverain Bien » (Aristote 2014, Livre I, 1, 1094a20-25 — 1094b-1095b5). Cette façon de voir est plutôt moniste, la cible étant unifiée et entièrement déterminée d'avance. Ma conception de la prudence est plus près de celle des pluralistes des valeurs, dont je traiterai au chapitre suivant. Pour le dire simplement, il m'apparaît qu'en éthique politique, la cible (le bien commun) vers laquelle l'archer fait pointer sa flèche n'est pas entièrement déterminée d'avance et requiert un certain dialogue pour en déterminer les modalités. C'est la raison pour laquelle lorsque j'évoquerai la *phronesis*, j'entendrai par là la vertu dans son acception entièrement aristotélicienne. De plus, ma compréhension spécifique de la « prudence » sera développée en détail au sixième chapitre. J'y exposerai ma proposition éthique alternative, qui se veut une démarche de sagesse pratique, à savoir une démarche éthique faisant appel à la raison pratique plutôt qu'à la raison spéculative ou théorique.

Un individu en possession de la *phronesis* sera caractérisé par la capacité

> […] de délibérer correctement sur ce qui est bon et avantageux pour lui-même, non pas sur un point partiel (comme par exemple quelles sortes de choses sont favorables à la santé ou à la vigueur du corps), mais d'une façon générale, quelles sortes de choses par exemple conduisent à la vie heureuse. (Aristote 2014, Livre VI, 5, 1140a25-35)

La *phronesis* est la capacité d'exercer son jugement dans des cas particuliers. Plus encore, elle est une vertu intellectuelle nécessaire à toute vertu morale (MacIntyre 1984, 154), ce qui n'est pas le cas des autres vertus intellectuelles. À cet égard, il est intéressant de noter que Thomas d'Aquin diffère un peu d'Aristote en concevant la prudence comme une vertu à la fois intellectuelle et morale. Selon Aquin,

> […] il appartient à la prudence […] d'appliquer la bonne raison à l'action, et cela ne se fait pas sans que la tendance soit bonne. La prudence a donc la nature de la vertu à la manière des vertus intellectuelles, mais aussi à la manière des vertus morales » [Traduction libre] (Aquin s.d., partie 1, question 47, article 4).

Ainsi, la conception chrétienne de cette vertu implique une sorte de rectitude de la volonté, une idée à laquelle Foot adhère également (1978, 4). Elle soutient qu'« […] il faut opposer la sagesse à l'astuce, car l'astuce est la capacité de prendre les bonnes mesures à n'importe quelle fin, alors que la sagesse n'est liée qu'aux bonnes fins, et à la vie humaine en général plutôt qu'aux fins de certains arts » [Traduction libre] (Foot 1978, 2). Conséquemment, la finalité doit être bonne, pour que l'on puisse parler de « sagesse » pratique (Foot 1978, 6; voir aussi MacIntyre 1984, 154).

e) Différentes approches à l'éthique de la vertu

L'éthique de la vertu peut prendre plusieurs formes. Hursthouse et Pettigrove en recensent quatre. La première, l'éthique de la vertu « eudaimoniste », pense la vertu en lien étroit avec une certaine conception de la bonne vie, de celle qui vaut la peine d'être vécue : en somme, de l'*eudaimonia* (2018, s.p.). La deuxième forme d'éthique de la vertu est celle qui est basée sur l'agent (« *agent-based* »), à savoir que la moralité dépend des motivations de l'agent (Hursthouse et Pettigrove 2018, s.p.). Cette approche se distancie un peu d'Aristote, en ce qu'elle voit chez l'individu, dans l'excellence de son caractère, la source de la vertu : ce qui qualifie la bonne action, c'est le fait qu'elle ait été choisie et exécutée par l'agent vertueux (Slote 2003, 203-204). L'accent, dans la qualification d'une action, est ainsi mis sur l'intentionnalité, beaucoup plus que sur l'objet (Slote 2003, 205-206). Une troisième forme place la finalité au centre de l'éthique de la vertu, ou encore sa cible : pour ses tenants, il importe de comprendre ce vers quoi la vertu tend. En quatrième et dernier lieu, une éthique de la vertu dite « platonicienne » tend quant à elle à être axée sur la contemplation du Bien, ce qui « […] fait place à de nouvelles habitudes de pensée qui se

concentrent plus facilement et plus honnêtement sur des choses autres que le soi » [Traduction libre] (Hursthouse et Pettigrove 2018, s.p.). On peut trouver des échos de cette dernière approche dans la pensée d'Augustin d'Hippone.

Évidemment, il existe d'autres d'approches que celle d'Aristote à l'éthique de la vertu. Le but de cette section n'est pas de les explorer en profondeur ni de connaître les détails des débats dans le sous-champ, mais d'en présenter un bref survol. Bien qu'Aristote soit la figure de proue associée à cette branche de l'éthique, il faut mentionner que d'autres penseurs ont mis de l'avant leurs propres conceptions d'une éthique de la vertu, comme le philosophe David Hume. L'on peut aussi penser à Francis Hutcheson, qui propose une distinction entre le bien et le mal moraux, d'un côté, et le bien et le mal naturels, de l'autre. Le bien moral — comme la vertu — suscite l'admiration, car il est une excellence; mais le bien naturel peut en revanche susciter la jalousie (2003, 50-53). Nous percevons, comme êtres humains, le bien moral par un « sens moral » octroyé par Dieu, qui nous fait vouloir nous préoccuper du bien des autres, et non seulement du nôtre propre (Hutcheson 2003, 56). Plus largement, il faut reconnaître que des auteurs cherchent à élargir les origines intellectuelles de l'éthique de la vertu à la grandeur du globe, en incluant dans la tradition des penseurs comme Mencius et Confucius dans l'Est, en plus de Platon et Aristote dans l'Ouest (Hursthouse et Pettigrove 2018; Vallor 2016; Ess 2009).

f) Critiques de l'éthique de la vertu

L'éthique de la vertu, comme toutes les traditions éthiques, a généré sa part de critiques et d'objections. Il suffira d'en mentionner quelques-unes ici. On a reproché à l'éthique de la vertu, par son insistance sur l'*être* plutôt que le *faire*, de ne pas fournir de principes clairs qui permettent d'orienter l'action (Hursthouse et Pettigrove 2018). Or, les éthiciens de la vertu soutiennent qu'au contraire, l'action peut être orientée par cette approche, sans avoir recours à des principes « codifiables » ou encore à un code de conduite. À cet égard, des penseurs soutiennent qu'il faut justement résister à la tentation de faire de l'éthique de la vertu une approche « codifiable ». C'est le cas de McDowell, qui explique qu'« occasion après occasion, on sait quoi faire, si on le fait, *non pas en appliquant des principes universels, mais en étant un certain type de personne* : celui qui voit les situations d'une manière distinctive » [Traduction libre, je souligne] (2003, 139).

Une autre objection est soulevée quant au fait qu'un agent vicieux puisse commettre une action vertueuse et vice-versa (Hursthouse et Pettigrove 2018, s.p.). Annas voit dans ce reproche une projection de la vision technique de l'éthique, à savoir qu'au lieu de suivre un manuel technique de l'action, on identifie la personne vertueuse à un expert technique, comme un technicien informatique qui aurait, au fond, internalisé une marche à suivre (Annas 2004, 68). Plus encore, elle remarque qu'une telle objection est aveugle à la distinction entre l'agent qui lutte pour devenir vertueux, et celui qui l'est déjà devenu (Annas 2004, 73). Au fond, cette vision traduit une conception manichéenne de la possession de la vertu qui, nous l'avons vu, devrait plutôt être conçue comme un processus progressif, du moins pour ceux qui s'en revendiquent.

On oppose aussi un certain relativisme culturel à l'éthique de la vertu, en ce sens que différentes cultures et époques peuvent avoir des conceptions changeantes de la vertu. Cela pourrait mettre à mal l'idée d'une tradition unifiée d'éthique de la vertu. Nussbaum (1990) et Vallor (2016) contournent cette objection en postulant une sorte d'unité dans la forme, mais avec des contenus qui diffèrent. Vallor parvient ainsi à comparer différentes traditions éthiques mondiales, de l'aristotélisme au confucianisme en passant par le bouddhisme, pour démontrer qu'une forme de « noyau conceptuel mince » est commun à toutes ces traditions. Les contenus respectifs, plus « épais », sont propres à chaque tradition. En fin de compte, « ces visions de la vertu humaine reposeraient sur des bases conceptuelles très similaires, mais divergeraient sensiblement dans leurs recommandations détaillées sur la façon de vivre » [Traduction libre] (Vallor 2016, 43). C'est dans ce sens que Vallor se dit en faveur d'une éthique de la vertu qui soit « pluraliste », dit-elle, soit « […] ouverte à plus d'un mode d'expression de l'épanouissement humain » [Traduction libre] (Vallor 2016, 44). Néanmoins, l'idée d'« unité conceptuelle » de la vertu (Vallor 2016, 45), même en version « mince », renvoie plutôt au monisme qu'au pluralisme. D'autres éthiciens de la vertu répliquent quant à eux, sans vraiment répondre à la question, que la difficulté de l'universalisation est propre à chaque théorie, incluant le déontologisme et le conséquentialisme. Malgré cela,

> une stratégie plus audacieuse consiste à prétendre que l'éthique de la vertu a moins de difficultés avec la relativité culturelle que les deux autres approches. On peut prétendre qu'un grand nombre de désaccords culturels découlent de la compréhension locale des vertus, mais les vertus elles-mêmes ne sont pas relatives à la culture. [Traduction libre] (Hursthouse et Pettigrove 2018, s.p.)

Enfin, un autre reproche soulevé est que l'éthique de la vertu — dans sa forme eudaimoniste, du moins — n'est qu'une expression d'égoïsme, en ce sens qu'être vertueux, c'est œuvrer à son propre bonheur, à son propre épanouissement. Pourtant, il faut admettre que dans l'éthique de la vertu, sont reconnus comme bons et vertueux des actes réalisés par des personnes, même s'ils ne concouraient pas directement à leur bien. On y trouve même des actes héroïques, ou certaines formes de sacrifice au profit des autres (Hursthouse et Pettigrove 2018, s.p.). Hutcheson affirme aussi que la bienveillance, soit de faire du bien pour les autres, n'apportera pas nécessairement du plaisir à l'agent qui la pratique. En fait, c'est souvent l'inverse qui se produit, selon lui (Hutcheson 2003, 58-59). Cette façon de voir entre en opposition frontale avec la prochaine tradition éthique dont nous traiterons, à savoir l'utilitarisme, en tant que branche du conséquentialisme.

2. L'utilitarisme comme théorie éthique conséquentialiste

L'utilitarisme est une théorie éthique issue du conséquentialisme, selon lequel la moralité d'une action (ou inaction) est déterminée sur la base des conséquences qu'elle génère. Les évaluations morales portent donc sur conséquences des actions ou des décisions de l'agent. Elles ne touchent pas son caractère (contrairement à l'éthique de la vertu), « la qualité morale intrinsèque de [ses] actes » (Taylor 1997a, 127) ou encore les principes formels qui l'animent (contrairement à l'éthique déontologique) (Darwall 2003b, 1). Je vais brièvement présenter les grandes lignes du conséquentialisme, pour ensuite traiter davantage de l'utilitarisme, sa variante la plus répandue.

a) Le conséquentialisme

Le philosophe irlandais Philip Pettit soutient que les théories morales peuvent être divisées en deux catégories, soit les théories conséquentialistes et non conséquentialistes, ou encore téléologiques et non téléologiques (ou encore déontologiques) (Pettit 1991, 230; voir aussi Sinnott-Armstrong 2019, s.p.). On entend par « téléologique », dans ce cas, des théories éthiques qui sont orientées vers une finalité particulière et déterminée. Le conséquentialisme, une théorie dite « téléologique », se veut « universalisable » et simple (Pettit 1991, 237-238). Dans une perspective

éthique conséquentialiste, les valeurs portées par les agents existent antérieurement et extérieurement à la moralité. Il en résulte qu'une relation instrumentale s'établit entre l'agent et les valeurs qui lui sont chères, et qu'il cherchera à maximiser. L'évaluation morale est « […] fondamentalement une évaluation de la valeur *instrumentale* ou *extrinsèque* » des actions [Traduction libre] (Darwall 2003a, 2). Cette évaluation extrinsèque est possible, car, contrairement à une approche déontologique, les principes moraux n'ont de la valeur que s'ils maximisent les conséquences positives, ou minimisent les retombées négatives. Cette maximisation se fait souvent de manière indirecte (Darwall 2003a, 6).

Souvent, les conséquentialistes font appel à des règles et des principes, mais leur valeur est instrumentale. Quant à leur véracité intrinsèque, elle est non-pertinente (Darwall 2003a, 6). Quand il veut prendre une décision de nature éthique, l'agent établit une hiérarchie de ses différentes options à l'aide de leurs pronostics respectifs (ou conséquences avérées), pronostics qu'il aura préalablement hiérarchisés entre eux (Pettit 1991, 232). Cela revient souvent à « parier une option » pour ensuite suivre la théorie décisionnelle, qui, elle, en calculera la valeur (Pettit 1991, 232).

Le conséquentialisme se décline en différents types. L'une des typologies du conséquentialisme est la différenciation entre le « conséquentialisme de l'acte » et le « conséquentialisme de la règle ». Le premier type soutient que la valeur morale de l'action est tributaire des conséquences de l'acte lui-même, tandis que le second affirme qu'elle dépend au contraire « […] des conséquences de l'acceptation sociale des règles qui soit exigent, soit interdisent, soit permettent l'acte, par rapport aux conséquences de l'acceptation d'autres règles possibles pour des actes et des circonstances de ce type » [Traduction libre] (Darwall 2003a, 2). Autrement dit, dans un conséquentialisme de la règle, on peut trouver des règles de type déontologique, mais désirées non pour elles-mêmes ou en vertu d'une appréhension rationnelle pure. Elles sont plutôt voulues pour servir l'utilité à long terme. Il faut aussi distinguer entre les conséquentialistes empiristes, comme Jeremy Bentham (1748-1832) et John Stuart Mill (1806-1873), et les conséquentialistes qui rejettent l'empirisme, tels que Henry Sidgwick (1838-1900) et George Edward (G.E.) Moore (1873-1958), pour qui « l'intuition rationnelle » (Darwall 2003a, 6) joue un rôle important dans l'évaluation des conséquences.

b) Postulats de l'utilitarisme

La branche la plus connue du conséquentialisme est l'utilitarisme (Lindbergh 2019, s.p.; Darwall 2003a, 3). Sans promouvoir une conception de la nature humaine ou encore de la « bonne vie », comme le fait par exemple Aristote, l'utilitarisme donne une certaine « direction » au conséquentialisme, soit celle — le nom l'indique — de l'utilité (Taylor 1997a, 128). Afin d'identifier le premier philosophe utilitariste, on pourrait remonter à quatre cents ans avant Jésus-Christ, au temps où le confucianisme était la doctrine dominante en Chine, avec la figure du penseur Mozi. Un siècle plus tard, en Grèce, c'est l'hédoniste Épicure qui peut y être associé (de Lazari-Radek et Singer 2017, 1-2). Toutefois, c'est aux penseurs occidentaux Jeremy Bentham et John Stuart Mill que l'on pense habituellement quand il est question d'utilitarisme. Systématisé par le premier dans sa forme dite « classique » (Darwall 2003a, 3), son idée de base est que « […] nous devrions faire du monde le meilleur endroit possible » [Traduction libre] (de Lazari-Radel et Singer 2017, 1).

Conçu d'abord comme une secte, l'utilitarisme se voulait, pour Bentham, un mouvement d'importance (de Lazari-Radek et Singer 2017, 4-5). Ce dernier proposait d'offrir une « science de la moralité » (Sandel 2016, 57), sur laquelle on puisse se baser de façon certaine pour appréhender la morale à la manière d'une science naturelle. En effet, soutient J. J. C. Smart, l'utilitarisme, « avec son attitude empirique à l'égard des moyens et des fins […] est adapté au tempérament scientifique et il jouit d'une grande flexibilité de traitement dans un monde changeant » (1997, 68). L'éthique procède désormais avec de nouvelles méthodes, celles inspirées des autres sciences, pour mener à bien sa recherche (Moore 2003, 89).

Bentham a fameusement déclaré que

> la nature a placé l'humanité sous la gouverne de deux souverains maîtres, la douleur et le plaisir. C'est à eux seuls qu'il appartient de nous indiquer ce que nous avons le devoir de faire […]. D'une part, le critère du bien et du mal, d'autre part la chaîne des causes et des effets, sont attachés à leur trône. Ils nous gouvernent dans tout ce que nous faisons, dans tout ce que nous disons, dans tout ce que nous pensons : tout effort que nous pouvons faire pour nous libérer de notre sujétion ne servira qu'à la démontrer et à la confirmer. [Traduction libre] (Bentham 2000, 11)

C'est devant ce constat que Bentham met de l'avant son « critère de l'utilité », placé au fondement de son système éthique. Ce dernier consiste en

> […] ce principe qui approuve ou désapprouve toute action, quelle qu'elle soit, selon la tendance qu'elle semble avoir à augmenter ou à diminuer le bonheur de la partie dont l'intérêt est en cause […]. Je dis de toute action quelle qu'elle soit; et donc non seulement de toute action d'un particulier, mais de toute mesure de gouvernement. [Traduction libre] (Bentham 2000, 11)

L'utilité, pour Bentham, est tout ce qui produit « […] bénéfice, avantage, plaisir, bien ou bonheur, (tout cela dans le cas présent revient au même) pour empêcher que se produisent des méfaits, des douleurs, des maux ou des malheurs à la partie dont l'intérêt est considéré […] » [Traduction libre] (Bentham 2000, 12). Ainsi, « […] la bonne action est, en toutes circonstances, celle qui produira le plus grand bonheur possible pour l'ensemble » [Traduction libre] (Sidgwick 2011, 597). En conséquence, « […] aucune action n'est jamais bonne ou mauvaise *en tant que telle* » [Traduction libre] (MacIntyre 1984, 15).

Pour le formuler sommairement, la moralité, dans une perspective utilitariste, peut se réduire en un calcul coût-bénéfice (Sandel 2016, 56). Il n'est plus question de loi naturelle, de vertu, ou d'impératifs catégoriques, mais d'une formule qui permet de calculer simplement et rationnellement, selon la raison instrumentale du moins (Taylor 1982, 129). C'est sans doute pour cette raison que l'utilitarisme est considéré comme étant contraire à la morale traditionnelle, qui serait de tendance plutôt déontologique (de Lazari-Radek et Singer 2017, 1). G.E. Moore soutient à cet effet que

> toutes les lois morales […] ne sont que des déclarations selon lesquelles certains types d'actions auront de bons effets. C'est le contraire qui prévaut généralement en éthique. […] le « juste » ne signifie et ne peut signifier que « la cause d'un bon résultat » et qu'il est donc identique à « utile »; il s'ensuit que la fin justifiera toujours les moyens et qu'aucune action qui n'est pas justifiée par ses résultats ne peut être juste. [Traduction libre] (2003, 89-90)

Le critère de l'utilité est le principe « du plus grand bonheur » (Bentham 2000, 320). Mill affirme par ailleurs que l'utilité n'est pas « […] quelque chose à distinguer du plaisir, mais le plaisir lui-même, ainsi que l'exemption de la douleur […] » [Traduction libre] (2003, 185). C'est le plaisir et la douleur qui confèrent un sens moral à nos expériences. Cette vision des choses n'est pas sans

lien avec l'hédonisme, selon lequel le bien et le mal « intrinsèques » sont le plaisir et la souffrance (Sinnott-Armstrong 2019, s.p.).

Le fait qu'un principe de maximisation de l'utilité se retrouve au cœur de l'utilitarisme pourrait suggérer qu'il s'agit d'un impératif et que, conséquemment, il y aurait certaines traces d'éthique déontologique dans la doctrine. Or, Smart soutient qu'au contraire, le « devoir » peut simplement renvoyer à une « recommandation » (1997, 18). Plus encore, on peut comprendre le principe de l'utilité comme une « […] orientation pour indiquer la voie la plus appropriée à suivre, en toute occasion, dans la vie publique comme dans la vie privée […] » [Traduction libre] (Bentham 2000, 320). Il reste toutefois que Bentham ne s'évertue pas à « prouver » le critère de l'utilité, à la base de son édifice éthique, puisque ce serait, de toute façon, impossible. Sidgwick est du même avis : il est sans objet de se demander pourquoi nous recherchions le plaisir et voudrions éviter la douleur (2011, 595). Dans la doctrine utilitariste, donc, le principe de l'utilité est considéré comme incontournable et indiscutable. Bentham affirme même que « lorsqu'un homme tente de combattre le principe d'utilité, c'est avec des raisons tirées, à son insu, de ce même principe » [Traduction libre] (Bentham 2000, 14). Pour le dire simplement, on n'échappe pas au principe de l'utilité. Malgré ce fait, des différences s'esquissent entre diverses formulations de cette même théorie qu'est l'utilitarisme.

c) Diverses conceptions de l'utilitarisme

On peut distinguer deux facettes de la théorie utilitariste. La facette « positive » cherchera à maximiser le plaisir, tandis que les tenants de l'utilitarisme « négatif » proposent de se centrer surtout sur la minimisation de la souffrance (Smart 1997, 30). De même, le contenu du mot « plaisir » (ou encore « utilité » ou « bien-être ») varie selon les penseurs utilitaristes. Bentham n'établissait pas de distinction entre les types de plaisirs, tandis que Moore, à la suite de Mill, en a proposé une hiérarchie (Smart 1997, 17). John Stuart Mill modifie de fait la théorie utilitariste de Jeremy Bentham en y introduisant une hiérarchie des plaisirs, en jugeant certains de ces derniers plus élevés que d'autres (Gordon 2007). Il lui semble en effet « absurde » que les plaisirs ne soient jugés que sur la base de la quantité, au lieu d'en inclure la qualité (Mill 2003, 34). Il s'agit d'ailleurs d'une faiblesse de l'être humain que de tendre vers le plaisir le plus immédiat, même si un plaisir

plus lointain était supérieur (Mill 2003, 36). G.E. Moore élargira la liste des biens pour y ajouter, en plus du plaisir, « l'appréciation de la beauté » et l'amitié, parmi d'autres biens (de Lazari-Radek et Singer 2017, 14-15).

Avec les ajouts des différents penseurs utilitaristes au cours des décennies, la théorie éthique se dote de contenu et de principes qu'elle intègre dans son système. Un exemple est le principe de la bienveillance généralisée. Ce sentiment consiste en une « […] disposition à rechercher le bonheur ou […] dans un sens ou dans un autre, des conséquences favorables pour toute l'humanité, voire même, pour tous les êtres vivants » (Smart 1997, 12). C'est dans un souci pour le bien-être du soi, mais de tous également que les utilitaristes développent leur théorie morale. À cet effet, Mill est on ne peut plus clair :

> […] personne dont l'opinion mérite un moment de réflexion ne peut douter que la plupart des grands maux positifs du monde sont en eux-mêmes éliminables […]. La pauvreté, en tout sens impliquant la souffrance, peut être complètement éteinte par la sagesse de la société combinée au bon sens et à la providence des individus […]. En bref, toutes les grandes sources de la souffrance humaine sont, dans une large mesure, et beaucoup sont presque entièrement, conquises par les soins et les efforts de l'homme […]. [Traduction libre] (Mill 2003, 40)

Cela dit, Mill affirme aussi que « la grande majorité des bonnes actions sont destinées non pas au bénéfice du monde, mais à celui des individus, dont le bien du monde est constitué […] » [Traduction libre] (Mill 2003, 43). On perçoit une unité théorique dans l'idée que le bien de chacun contribue au bien total, et ce, apparemment, sans heurts ou tragédies. L'utilitarisme serait, par conséquent, un système unifié. En revanche, peut-être y trouve-t-on un soupçon de naïveté quant à la réalité du potentiel égoïste de l'humain qui, en cherchant son bonheur, vient souvent compromettre celui d'autres personnes, et non y contribuer. Toutefois, dans la théorie utilitariste, il n'y a pas de conflit de ce genre — tant que l'on demeure, évidemment, dans le domaine théorique. C'est à cela que sert le principe de l'utilité, soit à « […] mettre en balance ces différentes utilités et délimiter la région dans laquelle l'une ou l'autre est prépondérante » [Traduction libre] (Mill 2003, 46).

d) Une éthique personnelle et politique

L'utilitarisme, tel que présenté par Bentham, se veut une éthique à la fois pour les prises de décisions personnelles et celles qui sont publiques ou politiques (Bentham 2000, 320-321). Ainsi, la formule décisionnelle est la même, qu'il soit question d'une personne ou encore d'une société. Dans le cas public, ce sera la somme de l'utilité, ou encore des plaisirs, qui sera mesurée (Bentham 2000, 124-125), puisque dans la pensée utilitariste, la communauté politique peut être comparée à « […] un "corps fictif", composé de la somme des individus qu'il comprend » (Sandel 2016, 56). Dans ce contexte, tous les membres du corps politique sont traités avec une égalité radicale :

> le principe que la plupart des utilitaristes ont tacitement ou expressément adopté est celui de l'égalité pure et simple : comme l'indique la formule de Bentham, « tout le monde compte pour un, et personne pour plus d'un ». Et ce principe est évidemment le plus simple, et le seul qui n'a pas besoin d'une justification particulière : car, comme nous l'avons vu, il doit être raisonnable de traiter un homme de la même manière qu'un autre, s'il n'y a aucune raison apparente de le traiter différemment. [Traduction libre] (Sidgwick 2011, 594)

Un décideur politique utilitariste devra se demander si la politique qu'il met de l'avant produit davantage de bonheur qu'elle ne crée de douleur, puisque l'objectif est de maximiser le bonheur de la communauté politique entière. Ainsi, la question que se poserait un décideur politique utilitariste est « […] si nous faisons la somme de tous les avantages d'une politique donnée et que nous en soustrayons tous les coûts, pouvons-nous dire qu'elle produit plus de bonheur que toute autre politique? » (Sandel 2016, 56). Il faut qu'il y ait une sorte de proportion ou, mieux encore, de déséquilibre entre les effets positifs et les effets potentiellement délétères d'une politique, en faveur des premiers. Comme l'éthique utilitariste s'inscrit dans une perspective positiviste selon laquelle les effets positifs et négatifs avérés sont mesurables, cette théorie peut être considérée comme « mathématique » — ou du moins, comme se prétendant telle. Étant donné que les valeurs en jeu peuvent être réduites quantitativement et mesurées dans le but d'établir une prédominance de l'utilité sur la souffrance, il est clair que l'utilitarisme considère les valeurs en jeu comme étant commensurables, ou réductibles à une seule.

e) La réductibilité des valeurs au principe d'utilité

D'un point de vue utilitariste, dans tous les cas de dilemmes éthiques, les compromis seront acceptables parce que, en optant pour la décision qui maximiserait l'utilité, l'agent garde pour ainsi dire « les mains propres » (Blattberg 2018, 156). Le principe de l'utilité est le seul principe, ou la seule valeur qui soit valide pour la gouverne politique : toutes les autres lui sont réductibles. Bentham soutient en effet que « […] tout principe qui s'en écarte doit nécessairement être un principe erroné » [Traduction libre] (Bentham 2000, 17). Il faut comprendre que les utilitaristes établissent des listes atomistes de plaisirs et de douleurs (Bentham 2000, 320-321), mais que dans l'éventualité où ces derniers entreraient en conflit, le principe de l'utilité ferait en sorte que ce conflit n'entache pas moralement les agents. De fait, le compromis, s'il sert à maximiser l'utilité, ne pose pas problème (Blattberg 2018, 156). Bentham serait ainsi un penseur moniste « non sophistiqué » ou « orthodoxe » (Blattberg 2018, 152), dans la mesure où, pour lui, toutes les valeurs sont réductibles en une seule, soit l'utilité. C'est pour cela que Mill peut soutenir que

> […] selon le Principe du Plus Grand Bonheur, […] *la fin ultime*, en référence et *au nom de laquelle toutes les autres choses sont souhaitables* (que nous considérions notre propre bien ou celui d'autres personnes), est une existence exempte autant que possible de douleur, et aussi riche que possible en plaisirs, tant sur le plan de la quantité que de la qualité […]. [Traduction libre, je souligne] (2003, 190)

On aura compris que l'utilité comme le bien-être est la seule finalité qui en vaille la peine. (Mill 2003, 33). Sur le plan méréologique, les tenants d'une approche ou d'une théorie holiste des valeurs éprouvent des difficultés particulières avec cet aspect de l'utilitarisme. On peut parler d'atomisme parce que la mesure à laquelle tout peut se réduire — l'utilité — est employée pour les individus, les sociétés ou encore la réalité matérielle dans son ensemble. Au fond, l'atomisme dans la doctrine utilitariste est tributaire du virage positiviste dans la modernité que Taylor retrace chez Hobbes. Dans cette compréhension, « tous les ensembles doivent être compris en fonction des parties qui les composent […] » [Traduction libre] : c'est le cas pour les sociétés et leurs individus (Taylor 1997a, 129). Toutefois, ce qui est perçu comme une faiblesse par des éthiciens holistes, les utilitaristes le voient au contraire comme une force.

f) Critiques de l'utilitarisme et du conséquentialisme

Comme les autres écoles éthiques, l'utilitarisme a été analysé, comparé, et contesté au cours des décennies. Tout d'abord, à ceux qui voudraient faire un rapprochement entre l'éthique de la vertu et l'utilitarisme, on pourrait répondre que cette démarche comporte quelques difficultés. Même s'il est question de « vertu » dans les écrits de Bentham, il ne s'agit pas de la conception aristotélicienne ou néo-aristotélicienne de cette dernière, mais bien du qualificatif d'une action menant à la maximisation de l'utilité. Inversement, le terme de « vice » sert d'adjectif à une action qui s'en éloignerait (Bentham 2000, 125-126). Il s'agit d'une divergence importante entre les deux traditions. Également, alors que les éthiciens de la vertu tendent à être favorables à une certaine forme d'héroïsme dans la pratique de la vertu, par l'entremise du sacrifice par exemple, les utilitaristes sont plutôt sceptiques. Certes, ils admettent que l'on puisse personnellement se sacrifier pour le bien d'autrui, mais « un sacrifice qui n'augmente pas […] la somme totale de bonheur, est considéré comme gaspillé » [Traduction libre] (Mill 2003, 194). Il ne s'agit donc pas d'une pratique encouragée par l'éthique utilitariste.

D'un autre point de vue, en tant que doctrine conséquentialiste, l'utilitarisme prête flanc à la critique de Bernard Williams selon laquelle la valeur d'une chose ne peut résider exclusivement dans ses conséquences. Si c'était le cas, « […] on remonterait sans fin de conséquences en conséquences et la régression serait évidemment désespérée » (1997, 78). Certaines actions sont accomplies non simplement en vue d'une finalité déterminée, mais pour elles-mêmes. Williams donne l'exemple de l'activité de voyager, qui peut être entreprise dans son propre intérêt (Williams 1997, 78).

Plus encore, il pose problème, selon Anscombe cette fois, qu'au sein du conséquentialisme, il n'y ait plus d'action qui soit essentiellement mauvaise. Plus problématique encore est le traitement que réserve Sidgwick à l'intention, qui doit prévoir toutes les conséquences des actions à poser. Les intentions prévues, mais non voulues ne seraient pas imputables à l'agent, puisqu'elles n'auraient constitué que des moyens vers une finalité qui, elle, était désirée (Anscombe 1958, 9). Dans cet ordre d'idées, et en poussant l'argumentaire à son extrême, un reproche adressé au conséquentialisme est de permettre les pires atrocités en raison d'un bien identifié (Sandel 2016,

51-55, 59-60). En effet, rien ne serait plus « impensable », dans un calcul conséquentialiste (Pettit 1991, 234). Ce à quoi des conséquentialistes répondraient qu'« il peut être terrible de penser à torturer quelqu'un, mais il doit être tout aussi terrible de penser à ne pas le faire et, par conséquent, de laisser, par exemple, une gigantesque bombe exploser dans un lieu public » [Traduction libre] (Pettit 1991, 234). C'est probablement ce qui pousse Smart à déclarer, dans son apologie de l'utilitarisme de l'acte, qu'il s'« […] adresse à des hommes bienveillants et compatissants, c'est-à-dire des hommes qui désirent le bonheur de l'humanité » (1997, 32). Certes, le risque de dérives est réel, mais, soutient-il, il faut accepter cette réalité comme un défi (Smart 1997, 65-68). Parallèlement, d'autres types d'abus pourraient être promus au nom de l'utilitarisme positif, mais aussi de sa formulation négative, à savoir la minimisation de la souffrance à tout prix. Smart soulève l'idée selon laquelle cette notion pourrait mener à ce que l'on désire exterminer l'espèce humaine, en voulant en finir la souffrance (1997, 30). Sans aller à cette extrémité, une expression contemporaine de ce type de raisonnement éthique est exemplifiée dans les débats entourant l'enjeu de l'euthanasie.

Vallor argumente, de son côté, que l'utilitarisme ne fournit pas les principes moraux dont nous avons besoin, tout spécialement dans une période de progrès technologique accéléré. Comme nous ne connaissons pas l'avenir — puisque nous sommes dans une situation radicale d'« opacité technosociale », il est difficile de prévoir les conséquences à long terme de nos actions. Le phénomène de la « convergence technologique » accentue la complexité du changement technologique. Vallor entend par cette convergence « […] des technologies distinctes qui se fusionnent en synergie de manière à amplifier considérablement leur portée et leur pouvoir de modifier les vies et les institutions […] » [Traduction libre] (Vallor 2016, 27). Les technologies les plus propres à cette synergie sont celles que l'on désigne par l'acronyme « NBIC » : la nanotechnologie, la biotechnologie, les technologies de l'information et la science cognitive (Vallor 2016, 27). L'utilitarisme, comme forme du conséquentialisme, est basé sur notre anticipation des retombées de ces avancées technologiques. Cependant, il est difficile de savoir si nous nous dirigeons, comme société, vers une augmentation ou une diminution du bonheur global (Vallor 2016, 27-28; Danaher et Vallor 2018, 17 : 15 min — 19 : 20 min). Non seulement l'avenir est-il voilé, ce qui nous pousse à des jugements moraux qui sont au mieux des paris, mais encore, la notion d'utilité n'est pas universellement comprise de la même façon. Anscombe le souligne

quant à la définition du plaisir : les Anciens eux-mêmes ne parvenaient pas à s'entendre sur une définition (Anscombe 1958, 2).

Par ailleurs, l'utilitarisme opère une distinction entre « […] la valeur morale des actes de la valeur morale des personnes […] » [Traduction libre], tandis que l'éthique déontologique kantienne place le respect de l'impératif catégorique au-dessus des liens qui unissent les êtres humains entre eux, déplore Vallor (2016, 23). Plus encore, ces deux théories éthiques sont trop exigeantes, « […] en demandant aux agents d'être impartiaux dans la pesée de leurs intérêts divergents des étrangers et de leurs proches. De telles considérations amènent à conclure que ces récits s'éloignent trop des intuitions morales communes » [Traduction libre] (Vallor 2016, 24). Cette critique de Vallor rappelle celle de Williams envers les théories morales et le principe de l'obligation incontournable qu'il appelle « *obligation out — obligation in* ». Les théories morales exigent que dans une situation d'urgence comme celle qui appelle à choisir qui sauver d'une catastrophe, entre sa femme et une autre personne, on doive suivre un raisonnement relevant de l'obligation morale. Pour Williams, il s'agit en réalité d'« une pensée de trop » (*a thought too many*) (1981, 18). Ces règles morales, de plus en plus abstraites, perdent de leur sens dans certains contextes. En quel cas, si l'on se console en se disant qu'il faut simplement permettre des exceptions à la règle, Vallor se demande si cela ne rendrait pas tout simplement les règles elles-mêmes obsolètes (Danaher et Vallor 2018, 19:20mn).

La théorie éthique conséquentialiste peut également recevoir le reproche d'être réductrice, en cherchant à produire « une théorie de l'action juste » (*a theory of right action*). Selon Julia Annas, le conséquentialisme

> […] isole un principe simple derrière les directives de notre discours éthique quotidien, puis nous indique comment formuler ce principe et l'appliquer pour nous dire, de manière systématique et spécifique, ce qu'il faut faire. Cette tâche est simple en principe, bien que difficile et technique en pratique. [Traduction libre] (Annas 2004, 63)

La théorie de l'action juste est analogue à un manuel technique. Cette critique d'Annas envers l'éthique comme recherche d'une procédure décisionnelle, si elle s'adresse au conséquentialisme, n'est pas moins pertinente pour le déontologisme. La force de ce type d'approche est son caractère universellement égalitaire; cependant, plusieurs problèmes se posent. D'une part, l'éthique comme

sagesse pratique est réduite à une technique, que quelqu'un de très peu expérimenté dans la vie, ou encore une personne avec de fort mauvaises mœurs pourrait maîtriser à la perfection, ce qui semble pour le moins contre-intuitif (Annas 2004, 63-64). D'autre part, une théorie de ce genre donne à penser que nous avons besoin, en tant qu'êtres humains, de nous faire dire quoi faire : il semble à Annas qu'il ne s'agisse pas là de ce que nous attendons *vraiment* d'une théorie éthique. En effet, cela dépersonnalise l'action de manière importante (Annas 2004, 64-65) et l'on pourrait supposer aussi que cela pourrait conduire à une forme subtile de déresponsabilisation personnelle. Elle poursuit :

> […] que la théorie soit représentée à l'extérieur de moi, comme un manuel, ou à l'intérieur de moi, comme un ensemble de directives sur la façon de penser, elle me dit toujours ce que je dois faire. Le fait demeure : ce que je devrais faire, c'est interpréter la théorie correctement. Cela s'apparenterait à essayer de prendre la peine d'internaliser des manuels techniques, comme les manuels des usagers pour la voiture ou l'ordinateur. [Traduction libre] (Annas 2004, 66)

Il devient clair qu'une théorie éthique déontologique ou conséquentialiste pose problème au regard de l'éthique de la vertu, puisqu'en dernière analyse, elle en fait un exercice de la raison technique (la *techne*) au lieu de la raison prudentielle (*phronesis*). Pour une éthique politique, l'emploi de la raison technique, ou encore celui de la raison instrumentale, tend à générer des pertes de sens. Cela dit, il faut reconnaître que même si les tenants de l'éthique de la vertu adressent ce reproche aux utilitaristes, ils sont parfois coupables de la même faute, notamment quand ils parlent d'une vertu comme d'une « compétence », possédant une structure intellectuelle particulière (Annas 1995, 233), ou d'une « boîte à outils » éthique visant l'exercice des vertus dans le design des technologies (Vallor 2018).

En somme, une éthique utilitariste ou plus largement conséquentialiste, pour l'encadrement de l'intelligence artificielle, est, malgré les critiques, imaginable, et portée par quelques chercheurs et organismes (par exemple, de Luca-Baratta 2019). Il faut reconnaître que l'utilitarisme présente certains attraits, dont son potentiel « mesurable », ainsi que celui du critère de l'utilité qui, au premier abord, peut sembler tout à fait intuitif, voire bienveillant (Taylor 1982, 129). Avant d'explorer les documents faisant appel à l'utilitarisme pour l'éthique de l'IA, il convient de se tourner vers les fondements de la théorie éthique déontologique.

3. L'éthique déontologique

Le déontologisme est une théorie morale axée sur le devoir formel, dont la figure emblématique est sans conteste le philosophe allemand Emmanuel Kant (1724-1804), qui s'est donné pour but de rechercher le « *principe suprême de la moralité* » (Kant 1848a, 11). Pour ce faire, il développe une « philosophie morale pure », c'est-à-dire entièrement détachée de l'empirie (Kant 1848a, 6). Le mot « déontologie » provient des termes grecs « *deon* » et « *logos* », qui signifient respectivement « devoir » et « étude de » (Alexander et Moore 2016, s.p.) (ou « raison », « langage »). Kant entend le devoir comme « […] *la nécessité de faire une action par respect pour la loi* » (Kant 1848a, 24). On peut aussi entendre le devoir non « […] comme ce qui a lieu, mais comme ce qui devrait avoir lieu, comme ce que la raison ordonne » (Kant 1848a, 36). On aura compris que la fondation de la moralité, pour Kant, est le devoir. De fait, soutient-il, « […] c'est ici précisément qu'éclate la valeur morale du caractère, la plus haute de toutes sans comparaison, celle qui vient de ce qu'on fait le bien, non par inclination, mais par devoir » (1848a, 21-22).

Une éthique déontologique défend le rôle central de la raison dans la connaissance et la différenciation du bien et du mal. Cette dernière nous aide à déterminer quelles actions sont immédiatement et universellement approuvées (Price 2003, 35). Conséquemment, la moralité ne dépend pas du point de vue ou de la perception : au contraire, elle est immuable. Si l'on refuse à la raison la capacité de distinguer le bien du mal moral, il faudrait alors lui refuser tout pouvoir d'appréhension du savoir. En effet, se demande le déontologue gallois Richard Price (1723-1791), « celui qui se méfie de sa raison dans un cas, pourquoi ne le ferait-il pas aussi dans l'autre? » [Traduction libre] (2003, 38-39).

L'éthicien écossais William David Ross (1977-1971) soutient qu'il y a des choses qui sont intrinsèquement bonnes, comme « […] la vertu, la connaissance et, avec certaines limites, le plaisir » [Traduction libre] (Ross 2003, 61). C'est dans la nature immuable de ces choses ou de ces actions d'être bonnes ou mauvaises : « le bien et le mal, semble-t-il, indiquent ce que *sont* les actions. Or, quelle que *soit* la chose, ce n'est ni par volonté, ni par décret, ni par pouvoir, mais par *nature* et par *nécessité* » [Traduction libre] (Price 2003, 36-37). Bien agir, c'est agir en conformité avec la « justesse intrinsèque » de certaines actions qui, elles, « […] ne dépend[ent] pas de [leurs]

conséquences, mais de [leur] nature » [Traduction libre] (Ross 2003, 79). Malgré cela, il n'est pas clair que les devoirs moraux vont de soi. Ross est plutôt d'avis qu'ils le deviennent, « […] tout comme les axiomes mathématiques » [Traduction libre] (2003, 68). La différence est que les mathématiques sont stables et statiques, tandis que les situations morales sont changeantes dans leurs caractéristiques (Ross 2003, 68-69). Une éthique déontologique, contrairement à l'éthique de la vertu, renvoie au *faire* plutôt qu'à l'*être*. Charles Taylor illustre le contraste entre une éthique déontologique et une éthique de la vertu, par exemple, en les opposant entre un type d'éthique procédurale (le déontologisme) et une éthique « substantielle » (« *substantive* » en anglais) (1993, 337). Il s'agit finalement d'une autre manière de les différencier. Je présenterai, dans un premier temps, la théorie éthique déontologique dans sa critique du conséquentialisme, qui me permettra de mettre de l'avant plusieurs de ses postulats, avant de passer à l'exposition des maximes et impératifs dans un second temps.

a) La critique déontologique du conséquentialisme

Le déontologisme est la théorie éthique la plus diamétralement opposée à celle du conséquentialisme.[12] C'est en opposition au conséquentialisme, entre autres, que la tradition éthique déontologique soutient que certaines actions sont justes, et d'autres injustes, en elles-mêmes (Lindbergh 2019, s.p.). Ce ne sont pas les conséquences à long terme d'une action qui importent, mais son motif (Sandel 2016, 155, 166). Autrement formulé, « […] certains choix ne peuvent être justifiés par leurs effets. Quelle que soit la qualité morale de leurs conséquences, certains choix sont moralement interdits » [Traduction libre] (Alexander et Moore 2016, s.p.). L'inverse est aussi vrai. C'est dans ce sens qu'il est souvent dit des déontologues que le juste prime sur le « bien », dans le sens où le respect des obligations importe davantage que de maximiser le « bien attendu » des conséquences. Pour un déontologue, quand un agent exécute une action en pensant aux bénéfices, il agit par intérêt personnel, et non par droiture ou « rectitude » (Ross 2003, 56).

[12] En fait, Kant a publié, en 1785, les *Fondements de la métaphysique des mœurs*, soit cinq ans après que Bentham ait publié ses *Principes de la morale et de la législation*. Sandel affirme à ce sujet que Kant a attaqué la pensée de ce dernier de manière « dévastatrice » (2016, 157).

Kant affirme que si le but ultime de l'être humain était effectivement son autoconservation et son autosatisfaction (par la maximisation du plus grand plaisir pour le plus grand nombre), la nature ne nous aurait pas dotés de raison, mais seulement de l'instinct, qui se serait chargé de nous y faire arriver (1848a, 16). Nous n'aurions, comme êtres humains, qu'à faire les choses en suivant nos tendances ou notre instinct. Il faut ajouter que, dans la perspective déontologique, il n'est pas garanti que ce qui génère du plaisir, ou encore ce qui constitue la préférence de la majorité, est nécessairement moral (Sandel 2016, 160).

Dans une perspective déontologiste, la recherche du bonheur comme finalité suprême est aussi problématique en ce que le bonheur est un but qui sera toujours conditionnel (Kant 1848a, 18), même s'il est vrai que tous les êtres humains désirent être heureux (Kant 1848b, 164). Par ailleurs, le bonheur et la bonté sont des choses différentes (Sandel 2016, 160). On l'a vu, le juste et le bien sont différents : ce qui est juste n'est pas toujours de maximiser le bien. Par exemple, il peut être préférable de tenir sa promesse, même si cela entraîne un peu de moins de bien « net » que de ne pas la tenir, car,

> après tout, une promesse est une promesse […]. Ce qu'est exactement une promesse n'est pas si facile à déterminer, mais nous sommes certainement d'accord sur le fait qu'elle constitue une sérieuse limitation morale à notre liberté d'action [Traduction libre] (Ross 2003, 69-70).

On ne peut pas non plus réduire la promesse à un instrument de maximisation du bien-être général (Ross 2003, 72-73).

Kant relève un autre problème aux théories morales utilitaristes, soit leur attachement excessif à la dimension de l'empirique et du contingent. Il est d'avis qu'« il ne faut […] jamais ériger en loi pratique un précepte pratique qui contient une condition matérielle (par conséquent empirique) » (1848 b, 180). Les lois morales kantiennes se veulent détachées de la causalité que l'on trouve dans le monde sensible (Kant 1848b, 156-157). Cette causalité fait appel à un usage inférieur de la raison, tandis que pour la morale, c'est la raison pratique *pure* qui est mobilisée, explique Kant. En effet,

> […] le moindre alliage compromet sa puissance et sa supériorité, tout comme le moindre élément empirique, introduit comme condition dans une démonstration mathématique,

> lui ôte toute valeur et toute vertu. La raison détermine immédiatement la volonté par une loi pratique, sans l'intervention d'aucun sentiment de plaisir ou de peine, même d'un plaisir lié à cette loi, et c'est cette faculté qu'elle a d'être pratique, en tant que raison pure, qui lui donne un caractère *législatif*. (Kant 1848b, 163-164)

Le niveau d'abstraction de la théorie éthique kantienne est considérable. En ce sens, on pourrait suggérer que par son détachement de tout contexte empirique ou matériel, elle s'oppose aussi à l'éthique de la vertu aristotélicienne dont la vertu de *phronesis* implique justement une prise en compte du contexte particulier dans lequel on se trouve. Plus encore, alors que plusieurs éthiciens de la vertu soutiennent que le meilleur moyen d'enseigner la vertu est d'en proposer des modèles, Kant soutient qu'il n'y a « […] rien de plus funeste à la moralité que de vouloir la tirer d'exemples » (1848a, 38). Le détachement entre la moralité est l'empirie est donc quasi complet dans l'approche kantienne. Les principes de la raison « […] sont tout à fait *a priori*, purs de tout élément empirique, et doivent être cherchés uniquement dans les concepts purs de la raison, et nulle part ailleurs, en quoi que ce soit » (Kant 1848a, 40).

Il est aisé de comprendre pourquoi toute moralité de l'« utilité » semble complètement erronée à Kant, puisque pour lui, le bien ne peut être confondu avec l'utile (Kant 1848b, 222). L'on ne peut se baser sur les désirs d'une majorité donnée à un moment donné ni sur des intérêts et des désirs changeants chez des individus — sinon, la dignité humaine elle-même serait sans fondement éthique solide (Sandel 2016, 160). Même si nous sommes des êtres sentients, ce ne sont pas le plaisir et la douleur qui devraient nous régir, mais la raison (Sandel 2016, 161-162, 175-176). Il faut éviter, argumente Kant, la confusion de notions entre, d'un côté, le bien et le mal et de l'autre, l'utilité (1848 b, 221-222). De plus, la réduction de la moralité à la promotion du plus grand bonheur pour tous est inadmissible pour les déontologues :

> […] il n'est pas concevable que la promotion du bonheur d'autrui comprenne l'ensemble de notre devoir, ou que la considération du bien public soit la seule, en toutes circonstances, qui puisse avoir un quelconque intérêt à déterminer ce qui est bien ou mal. [Traduction libre] (Price 2003, 45)

Il faut tout de même mentionner que, même si l'antinomie entre l'éthique déontologique kantienne et l'utilitarisme est réelle, il n'en demeure pas moins que certains auteurs semblent emprunter aux deux traditions à la fois. Si l'on prend l'exemple de « l'utilitarisme idéal », l'on est placés devant un devoir ultime, qui est celui de « produire du bien », « […] et que tous les "conflits de devoirs"

devraient être résolus en se demandant "par quelle action le plus de bien sera produit" » [Traduction libre] (Ross 2003, 57). En ce sens, on peut voir une sorte de cohabitation entre une éthique du devoir, et une autre fonctionnant selon la maximisation du plaisir, bien qu'il ne s'agisse pas à proprement parler d'une véritable éthique déontologique.

b) Maximes et impératifs

Dans sa théorie morale, Kant établit une distinction entre les maximes morales, que l'on peut se donner à soi-même, et qui seraient en quelque sorte comme de « bonnes résolutions », et les lois morales. Les premières ne sont pas « universalisables », les secondes, oui. Il explique à cet effet que

> des *principes* pratiques sont des propositions contenant l'idée d'une détermination générale de la volonté qui embrasse plusieurs règles pratiques. Ils sont subjectifs ou s'appellent des *maximes*, lorsque le sujet n'en considère la condition comme valable que pour sa propre volonté. (Kant 1848b, 153)

Les lois morales, en revanche, sont ces principes moraux « universalisables », c'est-à-dire dont on pourrait souhaiter qu'ils deviennent une loi universelle (Kant 1848b, 154). Les lois morales sont aussi appelées « impératifs ». Ces derniers « […] ont donc une valeur objective, et sont tout à fait distincts des maximes, qui sont des principes subjectifs » (Kant 1848b, 155). Les principes moraux sont évidents d'eux-mêmes, du moment que l'on est une créature rationnelle. C'est ce qu'affirme Price :

> les principes eux-mêmes, faut-il le rappeler, sont évidents, et conclure le contraire, ou affirmer qu'il n'y a pas de distinctions morales, en raison de l'obscurité qui entoure plusieurs cas où une concurrence s'installe entre les différents principes de la morale, est très déraisonnable. [Traduction libre] (2003, 52)

Ces principes sont également universels même si, dans la pratique, certains usages ont différé quelque peu (Price 2003, 53-54).

Dans la pensée kantienne, il existe des impératifs hypothétiques, qui sont conditionnels et font appel à la raison instrumentale (par exemple, assurer son propre bonheur). D'autres impératifs sont inconditionnels : ils représentent des choses que l'on ne désire pas en vue d'autre chose, mais

pour elles-mêmes. Ces impératifs sont catégoriques, et ils constituent des lois morales (Kant 1848b, 155). Ils font appel à la raison pure pratique. Une obligation enjoint tous les êtres humains à les pratiquer (Johnson et Cureton 2019, s.p.) : par exemple, « […] tenir ses promesses, payer ses dettes [et] dire la vérité » [Traduction libre] (Ross 2003, 57). La première formulation que propose Kant (Williams 2011, 70) de la « loi fondamentale de la raison pure pratique » (ou « impératif catégorique ») est la suivante : « Agis de telle sorte que la maxime de ta volonté puisse toujours être considérée comme un principe de législation universelle » (Kant 1848b, 174).

Ainsi,

> le seul principe qui dirige ici et doive diriger la volonté, si le devoir n'est pas un concept chimérique et un mot vide de sens, c'est donc cette simple conformité de l'action à une loi universelle (et non à une loi particulière applicable à certaines actions). (Kant 1848a, 27)

On doit comprendre que l'impératif catégorique tire son autorité de lui-même (Sandel 2016, 177-178). C'est ce qui amène Kant à affirmer que

> […] la règle n'est objective et n'a de valeur universelle, que quand elle est indépendante de toutes les conditions accidentelles et subjectives, qui distinguent un être raisonnable d'un autre. […] S'il arrive que cette règle soit pratiquement juste, c'est une loi, car c'est un impératif catégorique. (1848b, 156).

Plus simplement, le philosophe explique que « […] pour qu'une action soit moralement bonne, il ne suffit pas qu'elle soit *conforme* à la loi morale, mais il faut qu'elle soit faite *en vue de cette loi* […] » (Kant 1848a, 8). La volonté est « pure », autonome et donc emblématique de la liberté, lorsqu'elle guide les actions pour la loi morale elle-même, pour le devoir lui-même.

On pourrait résumer la loi morale kantienne avec l'impératif catégorique mentionné précédemment, soit l'universabilité des maximes. Une seconde formulation de l'impératif catégorique peut servir à condenser la moralité kantienne, soit de considérer « […] *l'humanité comme une fin en soi* » (Kant 1848a, 73).[13] L'une des raisons pour lesquelles il importe de se traiter comme des finalités, entre êtres humains, est la dignité intrinsèque à chaque personne, conséquence

[13] Il faut noter que cette seconde formulation de l'impératif catégorique est distincte de la « règle d'or », puisque cette dernière « […] dépend des faits contingents tenant à la façon dont les gens aimeraient qu'on les traite. L'impératif catégorique exige que nous fassions abstraction de telles contingences […] » (Sandel 2016, 185).

de sa rationalité et de sa volonté. En effet, chaque personne est le réceptacle du bien moral et, en conséquence, « […] il ne faut pas attendre de l'effet produit par son action » pour en juger la bonté (Kant 1848a, 26). Si on traite les êtres humains comme des fins, on s'assure d'éviter de favoriser des intérêts particuliers. Dans une perspective kantienne, tout le monde serait d'accord avec l'idée de ne jamais privilégier des intérêts particuliers au détriment d'autres : cela constituerait ainsi une maxime « universalisable ».

Dans le même ordre d'idées, le respect des droits des personnes ne repose pas sur des aspects particuliers de personnes humaines, mais sur leur dignité intrinsèque découlant de leur capacité pour la rationalité et l'autonomie (Sandel 2016, 161, 181, 183). Michael Sandel voit dans Emmanuel Kant l'instigateur théorique de droits humains universels (2016, 155). En raison de l'impératif catégorique de traiter les humains comme étant des fins en eux-mêmes, et non des moyens, les droits humains s'arriment bien à la pensée morale kantienne quand ils sont pensés en système unifié. Alors qu'un utilitariste pourrait être un tenant des droits humains « […] en faisant valoir qu'en les respectant on maximisera à long terme l'utilité » (Sandel 2016, 56, 155), un kantien respecterait les droits humains pour le principe moral que chaque personne est une finalité en raison de sa rationalité et de son autonomie (actuelles ou potentielles) (Sandel 2016, 157, 161, 183) : c'est ce qui fait sa dignité « intrinsèque ». C'est entre autres pour cette raison que l'éthique déontologique kantienne s'oppose à l'utilitarisme, car elle voit dans cette dernière une doctrine permettant l'utilisation des personnes en vue d'une fin autre qu'eux-mêmes, comme la maximisation du bonheur (Sandel 2016, 159-160, 180-183).

c) La primauté de l'autonomie

L'édifice kantien de la moralité est basé sur la volonté rationnelle, en adéquation avec la liberté. En effet, Kant trace « […] un lien étroit […] entre nos capacités à raisonner et à être libre » (Sandel 2016, 161). Conséquemment, agir en suivant notre instinct, nos désirs, nos envies ou encore nos sentiments de compassion pour les autres n'est pas une expression de notre liberté. Il s'agit plutôt d'une action qui se conforme à « une détermination donnée en dehors de nous » (Sandel 2016, 162-163). Il ne s'agit pas d'autonomie, mais au contraire d'« hétéronomie » (Kant 1848b, 130-131). La liberté est ainsi à comprendre comme autonomie ou auto-obéissance,

chez Kant. Autrement dit, nous sommes libres dans la mesure où notre volonté est déterminée de façon autonome (Sandel 2016, 162-163). C'est pour cela que Kant conçoit son système de moralité comme un tout qui ne peut se contredire. Pour lui, « la *volonté* est la causalité des êtres vivants, en tant qu'ils sont raisonnables, et la *liberté* serait la propriété qu'aurait cette causalité d'agir indépendamment de toute cause *déterminante* étrangère […] » (Kant 1848a, 98). En définitive, la volonté autonome est la liberté, et la liberté est la condition de la loi morale. L'autonomie de la volonté est la liberté, car la volonté n'est soumise qu'à elle-même et, si elle est bonne, elle guidera l'action en vue de la moralité. Ainsi, le seul moyen d'agir librement, avec autonomie — c'est-à-dire en s'obéissant à soi-même — c'est d'agir en suivant l'impératif catégorique.

Même si le système moral kantien fonctionne comme un tout unifié, des tensions ou des conflits peuvent survenir entre les devoirs — ou du moins, des apparences de conflit. Par exemple, une promesse faite auparavant peut sembler difficile à tenir actuellement.
Ross explique que

> lorsque nous nous estimons justifiés de rompre, et même moralement obligés de rompre, une promesse afin de soulager la détresse de quelqu'un, nous ne cessons pas un instant de reconnaître un devoir *prima facie* de tenir notre promesse, et cela nous porte à ressentir, non pas vraiment de la honte ou du repentir, mais certainement du scrupule [*compunction*], pour nous être comportés comme nous le faisons; nous reconnaissons, en outre, qu'il est de notre devoir de nous racheter d'une manière ou d'une autre auprès du promis pour la rupture de la promesse. [Traduction libre] (2003, 64-65)

Même si les données de la situation n'effacent pas le devoir de tenir sa promesse, il n'en demeure pas moins qu'un éthicien moniste, soutenant une approche déontologique de la moralité, pense qu'un certain inconfort, un regret, est de mise, sans pour autant verser dans la honte. On peut lire entre les lignes qu'il n'y a pas là faute morale véritable. Si la moralité fonctionne selon le cercle parfait de l'auto-obéissance, il n'y a pas de raison de voir de pertes irréparables ni de parler de culpabilité, pour autant que nous ayons agi conformément à la raison pure pratique.

d) Une éthique déontologique politique : le cas de John Rawls

Certaines approches d'éthique politique se sont inspirées de la moralité kantienne. On peut en déceler des traces dans la tentative de stipuler des principes de justice qui soient « neutres ». En

fin de compte, l'idée de primauté du juste sur le bien est présente dans le déontologisme kantien. Ainsi, deux camps se dessinent, quand il est question de penser l'éthique politique au prisme de la justice. D'un côté, on retrouve

> […] un groupe de penseurs dont John Rawls et Ronald Dworkin sont des figures clés contre un autre dont, par exemple, Michael Walzer et Michael Sandel sont des porte-parole importants. On considère qu'une théorie de la justice devrait être générale, s'appliquant à toutes les sociétés, ou du moins à toutes celles qui ont atteint un certain niveau de développement. L'autre estime que la justice est au moins en partie relative à la culture et à l'histoire particulières de chaque société. Appelons ces deux points de vue respectivement universaliste et communautarien. [Traduction libre] (Taylor 1993, 345)

La figure contemporaine dans le champ des idées politiques qui s'inscrit dans le sillage de Kant pour penser sa théorie de la justice est John Rawls. La motivation derrière sa *Théorie de la justice* (2009 [1971]) est de déterminer les fondements d'une société juste au moyen de la distribution des biens, en fournissant une alternative viable à une approche conséquentialiste (Simon 2002, 3-4). Pour lui, comme pour d'autres néo-kantiens et plusieurs libéraux, le juste doit primer sur le bien (Rawls 2009, 280). Rawls reprend à son compte la tradition de penser un contrat social et une sorte d'« état de nature », par l'entremise de ce qu'il appelle « la position originelle », dans le but d'élaborer des principes de liberté et d'égalité avec lesquelles l'entièreté de la société serait d'accord si elle avait été dans cette même situation hypothétique. Ce faisant, on peut dire qu'il fait appel à la première formulation de l'impératif catégorique pour fonder la légitimité de sa théorie. Avec l'exercice hypothétique du voile d'ignorance, Rawls « […] montre que, pour penser ce qu'est la justice, il faut identifier les principes auxquels nous consentirions si nous nous trouvions dans une position initiale d'égalité » (Rawls 1997 dans Sandel 2016, 208).

En ce sens, on peut voir la parenté intellectuelle avec Kant, pour qui la validité d'une maxime est son « universalibilité ». En effet, les porteurs de la primauté du juste sur le bien craignent qu'« […] une éthique du bien doit inévitablement renoncer à l'universalité, et donc à un point de vue critique envers toutes les formes culturelles » [Traduction libre] (Taylor 1993, 355). La primauté du juste sur le bien se porte ainsi garante de l'universabilité et ainsi, de la validité de cette théorie morale. Dans sa théorie de la justice en tant qu'équité (« *justice as fairness* »), Rawls soutient que

> […] les principes de la justice sont ceux qui seraient choisis dans la position originelle [et que] […] étant rationnelles, les personnes placées dans la position originelle reconnaissent qu'elles devraient examiner la priorité de ces principes. Car, si elles veulent établir des critères acceptés par tous pour arbitrer leurs revendications, elles auront besoin de principes de pondération. (Rawls 2009, 67-68)

Il faut faire appel à la raison dans la détermination et la hiérarchisation des principes éthiques, qui sous-tendront une société juste et équitable. Rawls affirme que « si nous ne pouvons pas expliquer la détermination de ces pondérations par des critères éthiques raisonnables, il n'y a plus moyen de poursuivre l'analyse rationnelle » (2009, 67). Cet appel à la raison n'est pas sans rappeler celui de Kant pour déterminer les fondements de la moralité. Dans la position originelle, les individus rationnels, ayant tous le même intérêt, établiront deux principes. Premièrement, le principe d'égale liberté, qui possède la priorité lexicale : il doit être respecté en premier lieu, car de lui découlent les autres. Le second, celui « qui gouverne les inégalités économiques et sociales » (Rawls 2009, 69), est divisé en deux, soit le principe de différence, qui stipule que la richesse des uns doit bénéficier aux autres, et le principe d'égalité des chances, qui garantit aux plus démunis la chance d'accéder aux postes qu'ils convoitent. Il reprend cette même catégorisation de principes au regard de l'attribution de liberté (Rawls 2009, 287).

L'objet de cette section sur l'éthique déontologique n'est évidemment pas d'exposer en détail la pensée de Rawls, mais simplement de la présenter comme un exemple parlant d'une approche éthique de type néo-kantien en philosophie politique. Certains aspects de sa théorie en font un bon exemple d'éthique politique déontologique. D'une part, on perçoit l'aspect moniste de son approche théorique, qui se veut un « tout » (Rawls 2009, 282) unifié et systématique dans lequel il ne devrait pas avoir de conflits sans solution consignée d'avance. D'autre part, un accent est placé sur la rationalité des individus, détachée de l'empirie dans la position originelle. Cela n'est pas sans rappeler la « raison pure pratique » que décrit Kant quand il est question d'identifier les impératifs moraux. L'idée d'ordonner les principes trouvés par la raison dans la position originelle en une sorte de hiérarchie consacre l'aspect systématique et unifié de sa vision, comme si l'organisation de la société pouvait se consigner dans un livre de règlements :

> une seconde possibilité serait que nous soyons capables de trouver des principes qui puissent être placés dans ce que j'appellerai *un ordre sériel ou lexical*. […] C'est un ordre qui demande que l'on satisfasse d'abord le principe classé premier avant de passer au second, le second avant de considérer le troisième, et ainsi de suite. On ne fait pas

> entrer en jeu un (nouveau) principe avant que ceux qui le précèdent aient été entièrement satisfaits ou bien reconnus inapplicables. *Un ordre lexical évite, donc, d'avoir jamais à mettre en balance des principes*. Ceux qui se trouvent placés plus tôt dans la série ont une valeur absolue, pour ainsi dire, par rapport à ceux qui viennent après, et *n'admettent pas d'exception*. [Je souligne] (Rawls 2009, 67)

L'idée est d'éviter que des principes ou des valeurs, lorsqu'ils entrent en conflit, soient soupesés entre eux, et éviter autant que faire se peut le recours aux jugements intuitifs en matière de justice (Rawls 2009, 70). Le Rawls de la *Théorie de la justice* pense que son propos est « universalisable » : étant donné qu'il « […] repose sur des présupposés faibles et largement reconnus, [il] peut gagner une approbation générale » (2009, 280).

e) Critiques de l'éthique déontologique

Plusieurs critiques sont adressées au déontologisme kantien. La présente section relève les plus « classiques », pour que le lecteur comprenne pourquoi je ne préconiserai pas, auprès des décideurs politiques, une approche déontologique pour l'éthique de l'IA. Il s'agit des reproches plus saillants, dont sont évidemment conscients les kantiens, sans pour autant constituer des arguments invalidant leur édifice théorique entier. Par exemple, un aspect critiqué de l'éthique déontologique est son niveau d'abstraction, qui ne convient pas à l'évaluation de l'éthique de l'action humaine. Si on prend la notion d'« humanité », dans la seconde formulation de l'impératif catégorique, on pourrait suggérer qu'elle est trop abstraite pour avoir une portée éthique réelle (Blattberg 2016, 10).

De même, les principes abstraits et rigides qu'énonce Kant, comme ses impératifs catégoriques, n'admettent aucune exception selon le contexte et nous frappent parfois par leur tension avec le sens commun. Par exemple, pour le philosophe allemand, il serait inadmissible de mentir pour protéger un innocent d'un sort malheureux, puisque mentir est toujours mauvais selon l'universabilité des maximes. Kant précise que l'on peut vouloir mentir pour se tirer d'affaire à un moment donné. Toutefois, on ne désire certainement pas en faire une loi universelle (Kant 1848a, 28-29). Dans le cas où l'on ne souhaiterait pas que la maxime de notre action donnée devienne une loi universellement applicable, il faut reconnaître que ce n'est pas

> […] parce qu'il en résulterait un dommage pour moi ou même pour d'autres [ce qui pourrait être compris comme un argument conséquentialiste], mais parce qu'elle ne peut pas entrer comme principe dans un système de législation universelle. (Kant 1848a, 29)

De son côté, MacIntyre note que, pour Kant, les maximes morales qui ont un caractère d'universalité sont celles qu'il a apprises de ses parents (1984, 44). La charge peut être injuste, à la lumière des écrits de Kant. Cependant, il n'en demeure pas moins que ce que Kant a identifié comme étant rationnel, il le croit universel :

> au cœur de la philosophie morale de Kant se trouvent deux thèses d'une simplicité trompeuse : si les règles de la morale sont rationnelles, elles doivent être les mêmes pour tous les autres rationnels, exactement comme le sont les règles de l'arithmétique; et si les règles de la morale *s'imposent* à tous les êtres rationnels, alors la capacité contingente de ces êtres à les appliquer doit être sans importance — ce qui est important, c'est leur volonté de les appliquer. [Traduction libre] (MacIntyre 1984, 43-44)

Une autre objection à l'éthique déontologique, surtout à l'échelle sociétale, est amenée par Charles Taylor. En s'appuyant notamment sur les travaux de MacIntyre, Taylor croit que l'on peut déceler, dans une éthique procédurale, un concept du bien qui « alimenterait secrètement » la conception du « juste » (1993, 337). MacIntyre relève en effet la distinction entre faits et valeurs, portée par Hume et Weber, et cette dernière débouche, soutient Taylor, sur une « vision mécaniste de l'univers », qui se veut neutre (Taylor 1993, 340). La conséquence est que

> […] la raison n'est plus définie *substantiellement*, en termes de vision de l'ordre cosmique, mais *formellement*, en termes de procédures que la pensée doit suivre, et en particulier celles impliquées dans l'adaptation des moyens aux fins, la raison instrumentale […]. […] La liberté prend donc un nouveau sens, ce qui implique de se détacher de toute autorité extérieure pour ne plus être régi que par ses propres procédures de raisonnement. (Traduction libre, je souligne) (Taylor 1993, 341)

Cependant, il n'est pas possible d'éviter l'enjeu du « bien », pour ne se centrer que sur le « juste ». Une vision de bien sous-tendra toujours une théorie du juste, comme celle que l'on retrouve dans des théories de la justice contemporaines : « l'ironie est que cette position est alimentée à l'origine par une vision du bien, celle d'une agentivité désengagée, libre et rationnelle, l'un des biens transcendants les plus importants et les plus formateurs de notre civilisation » [Traduction libre] (Taylor 1993, 357).

Conclusion

J'ai exposé ici quelques critiques de l'éthique déontologique, comme je l'ai fait pour l'éthique de la vertu ainsi que l'utilitarisme. Alors que ce chapitre était dédié à une exploration assez sommaire des fondements de ces approches éthiques monistes, le prochain, plus court, traitera des assises du pluralisme, à l'autre bout du continuum métaéthique. Ces deux chapitres ensemble forment la section « Fondements » de cette thèse, qui me semble incontournable pour ensuite dresser le portrait (la deuxième section) métaéthique des directives éthiques parues dans les dernières années. Une fois cette démarche accomplie, à l'aide d'un échantillon bien entendu, je proposerai au lecteur une critique et une proposition alternative pour approcher l'éthique de l'IA (la troisième et dernière section). C'est sur la base de mon approche éthique que je développerai un guide dialogique pour les décideurs politiques qui sont actuellement confrontés aux enjeux éthiques que peuvent poser les systèmes d'intelligence artificielle.

Chapitre 3 – Le pluralisme des valeurs

« *Ce qui est clair, c'est que les valeurs peuvent se heurter — c'est pourquoi les civilisations sont incompatibles. Elles peuvent être incompatibles entre les cultures, ou entre les groupes d'une même culture, ou entre vous et moi. […] Les valeurs peuvent facilement s'opposer au sein d'un même individu; et il ne s'ensuit pas que, si elles s'opposent, certaines doivent être vraies et d'autres fausses*. » — Isaiah Berlin [Traduction libre] (1990, 12)

Introduction

Au chapitre précédent, on a vu que pour les monistes, la question de l'unité est sous-jacente à leur pensée. En revanche, les penseurs pluralistes tiennent la position éthique et métaphysique opposée à celle des monistes. L'éthique en général, dans une perspective pluraliste, peut être appréhendée comme

> […] l'examen systématique des relations des êtres humains les uns avec les autres, des *conceptions, des intérêts et des idéaux* qui sont à la base des façons de se traiter mutuellement et des *systèmes de valeurs sur lesquels ces finalités de la vie sont fondées*. Ces croyances sur la manière dont la vie devrait être vécue, sur ce que les hommes et les femmes devraient être et faire, sont des objets d'enquête de la morale; et lorsqu'elles s'appliquent à des groupes et à des nations, et même à l'humanité dans son ensemble, on les appelle *philosophie politique*, qui est *une éthique appliquée à la société*. [Traduction libre, je souligne] (Berlin 1990, 1-2)

Il faut préciser d'entrée de jeu que le pluralisme auquel je me réfère ici renvoie à une prise de position métaéthique en opposition au monisme. Le mot « pluralisme » étant polysémique et fort populaire dans la littérature académique, notamment en éthique, j'aimerais faire quelques précisions de manière à situer le lecteur. Quand on invoque le pluralisme en éthique, c'est fréquemment pour illustrer le caractère multiple de quelque chose. Ainsi, la multiplicité ou la diversité pourraient être comprises comme ses synonymes. Le pluralisme peut même être invoqué pour renvoyer à l'incommensurabilité ou l'irréductibilité des valeurs entre elles (par exemple Ess 2020, 552). Néanmoins, ma manière de comprendre et d'avoir recours au monisme et au pluralisme dans cette thèse s'inscrit dans la démarche de Blattberg (2018) qui différencie

« incommensurabilité » et « incompatibilité ». L'incommensurabilité ou l'irréductibilité des valeurs en jeu serait un enjeu méréologique, tandis que la compatibilité (ou non) de ces dernières renverrait à la question de leur degré de connexion ou de cohésion (Blattberg 2018, 150-151). Autrement dit, dans ma conception du pluralisme, la multiplicité n'est pas automatiquement pluraliste : c'est le fractionnement ou la fragmentation des valeurs entre elles qui m'indiquera si nous faisons face à une démarche pluraliste ou non. L'exposé de ce chapitre devrait clarifier cette idée.

Contrairement à plusieurs éthiciens monistes, les penseurs pluralistes ne consacreront pas leurs énergies à l'élaboration d'une théorie éthique systématique et unifiée. En réalité, le point de départ de leur pensée est complètement différent. Les théories morales consistent en des « […] façons de mesurer des entités comme la "propriété" ou le "dommage"; elle […] pousse également à créer des règles générales qui peuvent résoudre les conflits entre les violations de la propriété et le dommage » [Traduction libre] (Leben 2018, 43). À l'opposé de cette manière de concevoir l'éthique, on retrouve Isaiah Berlin (1909-1997), Stuart Hampshire (1914-2004) et Bernard Williams (1929-2003), des figures de proue du pluralisme contemporain. Étant donné l'importance de ce courant dans les approches éthiques de l'intelligence artificielle, il m'apparaît judicieux d'explorer leur pensée et de relever leurs postulats principaux. L'un d'eux est qu'en raison de la pluralité de systèmes de valeurs et des valeurs elles-mêmes, il émergera nécessairement des tensions, voire des collisions entre ces valeurs. Selon les pluralistes, cela est inévitable (Apfel 2011, 11).

Puisque le compromis est entendu comme un incontournable, ce qui importe, pour les pluralistes, n'est pas de chercher à pallier ces inévitables pertes au moyen de la théorie éthique. En effet, le conflit de valeurs n'est pas un problème à éliminer : il s'agit plutôt d'un état de fait central à la condition humaine (Williams 1994, 282). Ainsi, chaque personne défend un ensemble de valeurs, dont certaines, telles la liberté et l'égalité, sont incommensurables. Cela signifie qu'il est souvent impensable de travailler au bénéfice de l'une de ces valeurs sans parler d'une forme de dommage à l'autre (Williams 1994, 288; Berlin 1990, 12-13). L'on fait alors face à une dynamique dont la somme est nulle.

Bernard Williams, comme Isaiah Berlin, soutient que

> […] le conflit de valeur n'a rien de nécessairement pathologique, qu'il fait au contraire partie intégrante des valeurs humaines, et qu'une compréhension correcte de ces valeurs doit le considérer comme central. (Williams 1994, 281-282)

Les penseurs pluralistes soulignent que les théories morales de l'Antiquité à la modernité tendent à promettre une sorte de « paradis terrestre » construit sur des idéaux vers lesquels l'humanité pourrait s'acheminer dès ici-bas. Pourtant, toutes les valeurs suprêmes que les êtres humains se donnent pour la pratique éthique ne sont pas forcément, à leur avis, compatibles ou réconciliables (Berlin 1990, 7-9). Ainsi, le monde présent est fragmenté. Il ne convient pas d'espérer une unité de la pluralité, et ce, même dans un monde qui serait à venir (Blattberg 2018, 158). Berlin affirme à cet effet que

> […] nous ne possédons aucune certitude *a priori* qu'il existe quelque part — peut-être dans un monde idéal qu'étant donné notre finitude nous ne pouvons même pas concevoir — une parfaite harmonie entre les vraies valeurs. (Berlin 1988, 214)

Par ailleurs, des conflits de valeurs peuvent éclore entre seulement deux, ou encore plusieurs personnes. Ces derniers peuvent toutefois survenir à l'intérieur d'une seule et même personne, laquelle se verrait, par exemple, confrontée à une opposition irréconciliable entre deux désirs (Williams 1994, 101, 104-105; Berlin 1990, 12). Sur un plan plus large, ces conflits de valeurs peuvent également surgir entre des sociétés et des pays. C'est pour cette raison que les pluralistes comme Hamsphire, Berlin et Williams sont appelés « pluralistes des valeurs » (Blattberg 2018, 158). Stuart Hampshire soutient qu'en plus du fait que ce type de collision est « […] à la fois inévitable et souhaitable »,

> il n'existera jamais de nation sans conflits, parce qu'il subsistera toujours, dans toute nation, non seulement des intérêts incompatibles, notamment économiques, mais encore des positions morales et des convictions inconciliables. (Hampshire 2011, 113, 97; voir aussi Hampshire 1989, 72-73)

Les pluralistes des valeurs affirment ainsi la diversité, l'incompatibilité et l'irréductibilité potentielles des valeurs — autrement dit, dans leur compréhension, leur incommensurabilité (Williams 1994, 281; Apfel 2011, 9-110).

Il importe de mentionner que je traite ici du pluralisme tel que systématisé par le cœur du groupe des « pessimistes d'Oxford » (Hall 2020, 13). Autrement dit, dans la thèse, lorsque je ferai mention du pluralisme, j'entendrai par là le pluralisme des valeurs tel que décrit dans ce chapitre. Il diffère du « pluralisme éthique » tel que l'entend Charles Ess (2019), par exemple. Ce dernier a mis de l'avant « […] les points communs culturels aux différences culturelles comme étant irréductibles, mais complémentaires les uns aux autres » [Traduction libre] (Ess 2019, 80). Or, le pluralisme des valeurs, tel que je le conçois, ne problématise pas seulement l'irréductibilité des valeurs, mais encore leur *incompatibilité* potentielle. De plus, cette incompatibilité tend à se manifester dans la pratique, la dimension éthique et politique par excellence. Ainsi, même s'il est possible de s'entendre sur une liste de valeurs ou encore une échelle de valeurs, leur respect dans les situations pratiques peut mener à des conflits dont l'issue n'est qu'un accommodement, qui implique forcément une forme de perte. Ainsi, dans cette thèse, le pluralisme n'est pas entendu comme synonyme de « multiplicité » ou de « diversité », mais comme une école d'éthique à part entière, en opposition aux écoles monistes, et dont la figure fondatrice contemporaine est le philosophe Isaiah Berlin.

C'est dans cette optique qu'au fil de ce chapitre, j'exposerai les fondements métaéthiques du pluralisme, sans pour autant prétendre à l'exhaustivité. Comme dans le chapitre précédent, le lecteur pourra trouver quelques critiques, adressées de la part des penseurs pluralistes à leurs collègues monistes, mais aussi, des critiques du pluralisme des valeurs lui-même. Une fois cette analyse complétée, les éléments nécessaires seront en place pour l'analyse d'une sélection de directives éthiques portant sur les enjeux de l'intelligence artificielle, à la section suivante.

a) Les dilemmes moraux et la tragédie

Chez les penseurs pluralistes, la reconnaissance de la centralité et de l'inévitabilité du conflit ne signifie pas qu'il existe une forme d'indifférence naïve par rapport à ce dernier. Plutôt, l'incommensurabilité des valeurs (entendue, pour eux, comme leur incompatibilité et leur irréductibilité) peut entraîner des situations dans lesquelles le choix effectué par le sujet agissant sera nécessairement mauvais, car il entraînera une forme de mauvaise action ou de mal moral (Williams 1994, 285). On l'a vu, pour un penseur moniste, si l'on peut se targuer d'avoir agi de

manière « moralement licite », l'on ne devrait pas en ressentir du remords ou de la culpabilité. Cela est en revanche compatible avec le fait de regretter une situation ou les conséquences d'une décision (Nielsen 2007). Il y aurait donc une distinction à opérer, chez les monistes, entre le regret et le remords, entre la déception et la culpabilité. Or, dans la tradition pluraliste, la réalité éthique est conçue bien différemment. En effet, il peut arriver que le sujet ait agi selon ce qui convenait dans la situation et que malgré tout, il éprouve du remords. Cela revient à dire que le sentiment subjectif de culpabilité n'est pas si aisé à balayer du revers de la main, dans une situation de dilemme éthique dont toutes les options impliquent une perte (par exemple, Nørskov et Rodogno s.d., 10).

Bien sûr, tous les pluralistes n'envisagent pas la fragmentation des valeurs avec le même degré de profondeur (Blattberg 2018, 158). Néanmoins, plus les valeurs qui se sont opposées dans un dilemme étaient riches au plan subjectif, plus le sentiment de perte est aigu. En conséquence, certains conflits moraux sont susceptibles d'arborer un caractère véritablement tragique, puisque les valeurs morales en jeu auraient été riches et profondes (Apfel 2011, 15). Des remords suivent la décision, et l'on peut affirmer que

> les souffrances qu'un homme éprouvera après avoir agi en pleine connaissance d'une telle situation ne viennent pas du fait qu'il doute sans cesse d'avoir choisi la meilleure solution, mais, par exemple, de la conviction claire qu'il n'a pas choisi la meilleure solution parce *qu'il n'y avait pas de meilleure solution*. [je souligne] (Williams 1994, 107)

Ainsi, pour les pluralistes, il est possible de faire face à un dilemme où les deux options impliquent nécessairement un mal moral. Plus encore, un dommage causé peut être non compensable, voire irréparable (Apfel 2011, 16). Les tragédies grecques en constituent de bons exemples : il suffit de penser à l'Antigone de Sophocle, ou encore à l'Hécube d'Euripide.

En éthique politique, si l'on trace un parallèle avec la littérature traitant du problème des « mains sales » (Blattberg 2018, 150), l'on pourrait affirmer que pour Williams ainsi que pour d'autres pluralistes des valeurs, dans des situations de conflits de valeurs, le sujet éthique aura souvent, sinon toujours, les mains sales (Williams 1994, 123). Une culpabilité quelconque est inévitable. Conséquemment, pour ces penseurs, au lieu de s'atteler à unifier le monde, la meilleure chose que l'on puisse faire est de limiter les dégâts. Cela se fera au moyen de la raison pratique.

L'importance accordée à la raison pratique est d'ailleurs un point commun entre les pluralistes des valeurs et les éthiciens de la vertu. C'est une des raisons pour lesquelles on peut localiser quelques traces de la pensée aristotélicienne dans les travaux des philosophes pluralistes.

b) La relation des pluralistes avec la pensée aristotélicienne

Certains parallèles philosophiques entre les pluralistes et Aristote peuvent être trouvés dans la pensée de Bernard Williams. En effet, l'importance du caractère, dans le jugement éthique, est capitale pour les deux philosophes. L'appel à la raison pratique pour les questions éthiques et politiques leur est aussi commun, à une différence près. Cette raison pratique, pour les pluralistes, n'est pas téléologique ou « orientée d'avance ». Si Aristote parle de « nature » ou de « finalité » (Aristote 2014, Livre I, 5, 1197a15-20), les pluralistes comme Williams s'en dissocient, rejetant de ce fait le caractère fixe et l'unité du bien comme *cible* des actions éthiques. La cible demeure, mais pour les pluralistes, elle n'est pas caractérisée par une unité ou un contenu prédéterminé. Aristote, pour décrire la cible des actions en éthique, emploie l'analogie d'un archer ayant devant les yeux le « Souverain Bien » :

> si donc il y a, de nos activités, quelque fin que nous souhaitons par elle-même, et *les autres seulement à cause d'elle*, et si nous ne choisissons pas indéfiniment une chose en vue d'une autre (car on procéderait ainsi à l'infini, de sorte que le désir serait futile et vain), il est clair que *cette fin-là ne saurait être que le bien, le Souverain Bien*. N'est-il pas vrai dès lors que, pour la conduite de la vie, la connaissance de ce bien est d'un grand poids, et que, *semblables à des archers qui ont une cible sous les yeux*, nous pourrons plus aisément atteindre le but qui convient? S'il en est ainsi, nous devons essayer d'embrasser, tout au moins dans ses grandes lignes, *la nature du Souverain Bien* […]. [je souligne] (Aristote 2014, Livre I, 1, 1094a20-25 — 1094b-1095b5)

Les pluralistes des valeurs divergent d'Aristote avec la notion de cible fixe, donnée par la nature ou essentielle à l'être humain en tant qu'humain. Certes, l'archer oriente sa flèche vers une cible : c'est là l'analogie de la raison pratique. Cependant, cette cible est mouvante et fragmentée pour les pluralistes des valeurs. Plus encore, non seulement « […] la cible est toujours en mouvement, mais [elle est] également en mouvement dans plusieurs dimensions » [Traduction libre] (Hamsphire 1989, 32), et ces dernières peuvent être en tension les unes avec les autres.

Ainsi, c'est la raison pratique qui guide l'orientation de la flèche vers la cible, chez Aristote et chez les pluralistes des valeurs; mais l'identification de la cible fait aussi l'objet d'un processus de raisonnement pratique chez ces derniers. Si les pluralistes font référence à la vertu de prudence (la *phronesis* chez Aristote) (Hampshire 2002, 645) lorsqu'ils mentionnent la raison pratique, il s'agit d'une prudence non téléologique (au sens de la nature), une prudence politique qui permet l'arbitrage des conflits. C'est la raison pour laquelle la conception de la prudence, chez un pluraliste et chez un éthicien de la vertu, diffère de façon importante. Les pluralistes des valeurs soutiennent que c'est dans l'identification de la fin ainsi que des moyens pour y parvenir que les valeurs entrent en conflit entre les individus et les sociétés. Face à cette réalité, il faut tenter de minimiser les pertes, autant que possible. Il incombe pour ce faire de négocier, de se réajuster, et de réorganiser le classement de ses valeurs dans chaque situation donnée. La négociation est centrale dans la sphère politique, et c'est dans les institutions qui permettent la négociation politique que Hampshire voit les fondements de l'éthique politique (Hampshire 1989, 13-14). J'y reviendrai un peu plus bas.

Conséquemment, ce n'est pas dans la nature ou dans une hiérarchie de l'âme ou de la cité qu'il faut chercher la fondation de l'éthique (Hampshire 2002, 635-636; 1989, 31). Moins encore dans une recherche de perfection d'un rôle, soit de politicien ou de philosophe, qui suppose des « potentialités essentielles », une « […] polyvalence, l'évitement de l'excentricité et de l'asymétrie […] » [Traduction libre] (Hampshire 1989, 27). Hampshire comprend que, dans la pensée éthique aristotélicienne, une certaine forme de complétude vertueuse est demandée, ce qui lui pose problème. Après tout, les personnes que l'on estime spécialement ont des vertus différentes qui les rendent admirables pour des raisons différentes, et il n'est pas nécessaire que tout le monde possède l'entièreté des excellences possibles. Il est clair que cette posture est en opposition à la doctrine moniste d'Aristote sur l'unité des vertus, pour qui la possession entière d'une vertu implique les autres. En tenant compte de cette variété des vertus, ou types d'excellence, demande Hamsphire, « […] y a-t-il quelque chose qui puisse qualifier la finalité incluant tout pour l'homme? » [Traduction libre] (Hampshire 1989, 28-29).

Dans le même ordre d'idées, une autre question peut être soulevée dans la comparaison entre l'éthique aristotélicienne et l'éthique pluraliste, à savoir, la place de la vertu, parfois mentionnée par les pluralistes des valeurs. Par exemple, Williams y fait allusion dans sa critique

des « systèmes de moralité » ou des théories éthiques modernes : les vertus ne sont visibles qu'en dehors de ces systèmes, dit-il (1985, 195). Si l'on considère les vertus comme une liste de valeurs pouvant entrer en conflit (comme les vertus abordées, une à la fois, dans *l'Éthique à Nicomaque*), on pourrait croire que, simplement, les pluralistes sont des « éthiciens de la vertu ». Ce serait trop rapide. D'abord, en raison de la conception différente de la prudence et du rejet de la téléologie dans la nature. Ensuite, même si « valeurs » et « vertus » affirment toutes deux des biens, il faut mentionner de prime abord que la notion de valeur est plus moderne que la notion de vertu. Cette dernière a des racines grecques anciennes; tandis que la première est surtout attribuée à Nietzsche en tant qu'entité créée (Anderson L. 2017, s.p.).

Plus encore, lorsqu'il est question de l'éthique de la vertu, on suppose la doctrine de l'unité des vertus — soit que les vertus ne peuvent entrer en conflit, mais se complémentent et se renforcent les unes les autres. En outre, contrairement aux valeurs, qui sont plus génériques, les vertus sont des qualifications de l'être humain qui les pratique. Elles impliquent l'acquisition d'une bonne habitude ainsi que d'une amélioration presque ontologique. En comparaison, la valeur pourrait être un attribut plus accidentel à l'être humain. Cela pourrait expliquer pourquoi certains tenants de l'éthique de la vertu conçoivent les différentes expressions de cette dernière comme étant universelles. Par exemple, Martha Nussbaum (1990) propose une liste universelle de vertus au fondement d'une démocratie sociale aristotélicienne. De son côté, Shannon Vallor établit des parallèles entre les vertus du monde philosophique occidental et celles des traditions de l'Asie du Sud-Est (2016, 22). Les mêmes vertus, les mêmes qualités ou bonnes habitudes peuvent ainsi être désignées par d'autres noms dans des cultures différentes, selon ces deux philosophes.

En définitive, même si les pluralistes des valeurs empruntent à Aristote, ils ne s'identifieraient pas entièrement à la pensée éthique proposée par ce dernier. L'usage de la raison pratique ne s'apparente pas totalement à cette vertu intellectuelle aristotélicienne qu'est la *phronesis*, puisqu'elle implique une finalité donnée par la nature. Les pluralistes des valeurs ne se considèrent pas comme des penseurs téléologiques et récusent l'emploi de tels concepts. C'est ce qu'Isaiah Berlin souligne dans son dialogue avec le philosophe canadien Charles Taylor :

> […] je crois en une multiplicité de valeurs, dont certaines sont conflictuelles ou incompatibles entre elles, poursuivies par des sociétés différentes, des individus

> différents et des cultures différentes, de sorte que la notion d'un seul monde, d'une seule humanité se déplaçant en une seule marche des fidèles, *laeti triumphantes*, est irréaliste. [Traduction libre] (Berlin dans Taylor 1994c, 3)

On aura compris que, dans la perspective pluraliste, non seulement les valeurs sont distinctes, mais également incompatibles entre elles, d'une société à l'autre (Berlin dans Taylor 1994c, 3). La question de l'arbitrage ou du choix entre ces valeurs se pose donc. J'exposerai brièvement en quoi on peut distinguer, chez les penseurs pluralistes, une branche plus décisionniste que d'autres.

c) Pluralisme et décisionnisme

Le terme « décisionnisme », inspiré de la philosophie de Søren Kierkegaard et de ses sauts de foi, aurait été proposé par Christian von Krockow, un sociologue allemand (Berger 2005, 91). Le concept renvoie à un acte « irréductible, purement volontariste » (Höffe 1993, 64), une prise de décision qui semble arbitraire. Pour Kierkegaard, et pour le dire assez crûment, « […] le décisionnisme est la thèse selon laquelle la valeur d'une décision se trouve dans le fait qu'on la prend, et non dans son contenu » [Traduction libre] (Lampert 2018, s.p.). Il est clair pour MacIntyre que, dans la modernité — surtout à la lumière de l'étude de l'ouvrage « *Enten-Eller* » de Kierkegaard, autorité et rationalité sont facilement scindées. Parfois, un choix éthique n'est fait « pour aucune raison » :

> dans notre propre culture, l'influence de la notion de choix radical apparaît dans nos dilemmes sur les principes éthiques à choisir. Nous sommes presque intolérablement conscients des possibilités morales rivales [Traduction libre] (MacIntyre 1984, 43).

Une autre façon de l'expliciter se trouve dans le fait qu'un décisionniste, lorsque confronté à un conflit de valeur incommensurable, tranchera sans y voir un appel à employer la raison pour sa décision. Pour lui, en effet, il n'y a pas de moyen de connaître quelle est la meilleure ou la moins mauvaise des options, puisque l'écart entre les valeurs empêche une comparaison sur le plan de la raison (Blattberg 2018, 158). Ce qui importe réellement est l'acte de trancher en prenant une décision. Il faut comprendre que tous les pluralistes des valeurs ne sont pas nécessairement décisionnistes, mais certaines le sont effectivement : c'est le cas, par exemple, de Max Weber ou de John Gray (Blattberg 2018, 158). De leur côté, Berlin et Williams se défendent bien d'être des

décisionnistes. Ensemble, ils ont étoffé une réponse aux voix critiques selon lesquelles le pluralisme se caractériserait par une dynamique à l'intérieur de laquelle

> […] il n'est pas nécessaire qu'il y ait une valeur qui, dans tous les cas, l'emporte sur l'autre; ou que, dans chaque cas particulier, la raison n'ait rien à dire (c'est-à-dire qu'il n'y ait rien de raisonnable à dire) à savoir laquelle de ces valeurs devrait prévaloir. Les pluralistes — nous, en tant que pluralistes, du moins — considèrent que la première de ces opinions avec manifestement vraie, et la seconde est manifestement fausse. [Traduction libre] (Berlin et Williams 1994, 308)

Ainsi, pour Berlin et Williams, même dans une situation d'incommensurabilité des valeurs, la raison doit s'exprimer. Elle s'insère tout à fait dans le processus de réflexion éthique. La raison pratique permet en effet l'exercice d'équilibrage des valeurs en conflit, sans que le différend en soit pour autant éliminé. Hampshire soutient à cet effet que

> en *pesant les pour et les contre* dans le cadre d'un conflit intérieur, nous parvenons à une *situation d'équilibre*, qui dure jusqu'à ce que nous ayons de nouveaux éléments à prendre en considération; il en va de même dans le domaine public et politique. [je souligne] (2011, 111-112)

Il s'agit au fond d'une démarche dialogique s'apparentant à la négociation, qui prend la forme d'un jeu à somme nulle. Autrement dit, pour régler les conflits qui se présentent dans la sphère politique, il importe de présenter les parties impliquées et d'entendre leurs revendications. C'est la façon qu'ont les pluralistes de répondre à l'incommensurabilité des valeurs en conflit. C'est ainsi que Hampshire renvoie à l'analogie de « la balance de la justice » [Traduction libre] (Hampshire 2002, 639).

Berlin abonde un peu dans le même sens lorsqu'il met de l'avant l'importance du compromis dans l'opposition des valeurs. Il récuse l'idée d'un État parfait, illustrant par des exemples historiques les aberrations et les atrocités auxquelles une telle idée a souvent mené, et soutient par ailleurs que l'obligation de l'État est d'abord et avant tout d'éviter les circonstances de souffrance extrême. Il est intéressant de noter que cette prise de position peut présenter quelques parallèles avec des formes d'utilitarisme en politique, chose que Berlin reconnaît lui-même (« les solutions utilitaristes sont parfois erronées, mais, je soupçonne, bénéfiques la plupart du temps » [Traduction libre] [Berlin 1990, 18]). Il souligne que

> nous devons donc nous engager dans ce qu'on appelle des *compromis* — les règles, les valeurs, les principes *doivent céder les uns aux autres* à des degrés divers, dans des situations spécifiques. […] Le mieux que l'on puisse faire, en règle générale, est de *maintenir un équilibre précaire* qui empêche l'apparition de situations désespérées, de choix intolérables — c'est la première exigence pour une société décente […]. Une certaine humilité dans ces matières est très nécessaire. [Traduction libre, je souligne] (1990, 18)

Cet exercice, cette pratique de la négociation acquiert conséquemment un relief particulier dans une éthique pluraliste. Elle constitue la forme de dialogue retenue pour l'arbitrage des conflits. La négociation est par ailleurs présentée, très souvent, comme ayant lieu entre deux « parties prenantes » (*stakeholders* en anglais). On verra, dans l'analyse des documents traitant de l'éthique de l'IA, que les appels à la négociation avec des parties prenantes peuvent s'avérer être des signes de sympathie ou de proximité avec la tradition éthique pluraliste.

d) La négociation et les « parties prenantes »

Dans une perspective pluraliste en éthique, la négociation entre parties prenantes tient pour acquis que les intérêts sont, d'entrée de jeu, fragmentés entre les diverses parties intéressées à un enjeu donné — ce qui est souvent le cas —, mais qu'une harmonisation parfaite de ces intérêts est difficile, voire impossible. C'est pourquoi la négociation est conçue par les pluralistes comme étant la solution dialogique pouvant exercer cette évaluation comparative des valeurs entre elles (Blattberg 2013, 8). Malgré les pertes inévitables, le processus de mise en balance des valeurs se fera de façon, somme toute, pacifique. Ce mode de fonctionnement est courant dans les entreprises, « […] lorsque les intérêts [des] différentes parties prenantes s'opposent à ceux des dirigeants d'entreprise et qu'ils sont appelés à tenter un équilibre en faisant des compromis adaptés aux circonstances » [Traduction libre] (Blattberg 2013, 2).

Ce qui est intéressant pour le propos, c'est que, comme nous allons voir, cette approche est l'une de celles qui sont prisées par des gouvernements ou des acteurs industriels quand il s'agit d'éthique de l'intelligence artificielle. Dans cette forme de dialogue, il ne s'agit pas d'éliminer le compromis, mais de le considérer comme un « investissement » : cela revient à abandonner un « profit » à court terme, dans le but d'en obtenir un à long terme (Blattberg 2013, 4). Une IA qui

soit éthique peut être perçue comme un but ou un bien commun, auquel tous participent et tous ont une *part*. En ce sens, les intérêts des parties sont dans le tout, et vice-versa. Néanmoins, dans le contexte d'une initiative pour le développement d'une approche éthique en intelligence artificielle, le fait de procéder d'emblée en identifiant les diverses parties prenantes, c'est opérer (ou reconnaître, selon la perspective éthique dans laquelle on se situe) une fragmentation dès le départ, sans notion d'un tout holiste qui puisse les rassembler dans leur multiplicité (Blattberg 2013, 9, 11). D'autres lignes de fractionnement peuvent nécessiter une démarche de négociation en éthique, par exemple le recours aux droits de la personne. Certes, comme je l'ai exposé au chapitre précédent, l'invocation des droits de la personne peut être traditionnellement associée à l'éthique déontologique de Kant — si ces derniers sont pensés en système formel, unifié. En revanche, si ce n'est pas le cas, peut-être que la fragmentation et la mise en balance des droits appellent une démarche éthique pluraliste.

e) Les pluralistes et les droits de la personne

Plutôt que de faire appel à un cadre formel dans lequel s'insèrent les différents droits de la personne, ou à une théorie les systématisant, les pluralistes auraient tendance à voir leur mention comme une occasion de négocier (Blattberg 2016, 9). Après tout, les droits représentent des valeurs encapsulées et listées, et sont souvent pesés les uns contre les autres. On pourrait donc suggérer qu'il y a deux grandes « écoles » quant aux droits de la personne : une moniste, théorique, déontologique; et une autre pluraliste qui prône la mise en balance et l'arbitrage.

Hampshire, comme pluraliste des valeurs, peut être associé à la seconde approche. En effet, il remarque que « […] les droits, dont dépend la liberté de l'individu, ne sont pas absolus, et qu'il est de leur nature *qu'ils doivent être mis en balance* avec d'autres revendications […] » et avec eux-mêmes [Traduction libre; je souligne] (Hampshire 1976, 89). La tension avec les principes absolus et les impératifs de Kant est manifeste ici. De même, dans les documents traitant de l'éthique de l'IA, quand il sera question de « droits absolus », on pourra conclure que cette conception des droits de la personne pourrait difficilement être pluraliste. Il faut comprendre que, pour un pluraliste, il ne peut y avoir de ciment pour l'entièreté des droits de la personne, qui les systématiserait sous une forme hiérarchique. Aucune conception de la nature humaine ou de la

bonne vie ne peut être partagée par tous et servir de critère à la réconciliation des droits entre eux. Même lorsqu'il n'est pas question de « nature humaine », mais plutôt d'« humanité », ce dernier concept peut facilement verser dans l'abstraction, dans le détachement du concret (Blattberg 2016, 10, 14-15).

Toutefois, à la différence de Hampshire, Charles Taylor (qui n'est pas un penseur pluraliste[14]) soutient quant à lui qu'une compréhension de la nature humaine — même si elle n'est pas nommée en tant que telle — sous-tend toujours la notion de droits de la personne modernes. En effet, « […] quoi qu'en dise le droit positif, l'être humain jouit de certains droits, du seul fait qu'il est humain » [Traduction libre] (Taylor 1994a, 18). En revanche, il est difficile de fonder rationnellement une seule conception de la nature humaine et, conséquemment, de la « bonne vie ». Taylor remarque qu'

> il y a clairement, d'une part, un *grand nombre de conceptions différentes de la bonne vie*, dont beaucoup méritent notre respect et notre admiration. D'autre part, il nous semble qu'il existe des *normes universelles claires* de ce qui est juste, ancrées dans un respect obligatoire des êtres humains et de leur égalité fondamentale, ce que nous ne pouvons regretter. *Ces deux-là sont régulièrement en conflit*, bien sûr. Certaines conceptions de la vie familiale restreignent gravement la vie des femmes, par exemple; le respect de la liberté peut être lié au déni de la liberté des adultes à leurs subordonnés; et ainsi de suite à travers une gamme pratiquement infinie de cas [Traduction libre, je souligne] (Taylor 1994a, 19)

Les réactions face à ces conflits sont diverses. Les pluralistes des valeurs (non-décisionnistes) rejettent le relativisme, et abordent donc ces conflits comme un jeu à somme nulle. Pour Taylor, cette dynamique de la perte est une occasion manquée de croître sur le plan humain, la croissance morale étant liée au conflit moral (Taylor 1994a, 19). Taylor est toutefois un penseur téléologique, pour qui « l'être humain est — oui, "par nature" — un animal langagier, un être de culture, un être qui invente et crée, mais l'invention et la création ne sont fécondes que lorsque l'inspiration vient au-delà du soi » [Traduction libre] (Taylor 1994a, 19). Il reconnaît donc la pluralité des valeurs, mais il se range tout de même plus près de la téléologie et, donc, du monisme, que de la fragmentation et du pluralisme. L'objection de Taylor pourrait donner à penser que le pluralisme

[14] Des précisions sur le positionnement de la pensée de Taylor seront apportées au début du chapitre 6.

tend facilement vers le relativisme, alors que le monisme éviterait cet écueil. La question a été posée et mérite d'être quelque peu explorée.

f) Pluralisme et relativisme

D'emblée, les pluralistes des valeurs non-décisionnistes, comme Berlin et Williams, se défendent d'être des relativistes. D'entrée de jeu, une précision s'impose. Le relativisme (et son possible antonyme l'absolutisme) sont des catégories distinctes du monisme et du pluralisme. En d'autres termes, le pluralisme et le relativisme ne sont pas synonymes, et constituent deux positions compatibles. En effet, le relativisme et l'absolutisme renvoient à des prises de position de nature *épistémologique*. Le monisme et le pluralisme nous informent plutôt sur le fractionnement ou non du réel. Comparer le relativisme et le pluralisme comme des éléments d'un même continuum, qui ne différeraient que par le degré, me semble donc être une erreur.

On l'a vu au chapitre dernier, Charles Ess, dans son ouvrage *Digital Media Ethics*, n'établit pas les mêmes distinctions que je viens d'esquisser. Selon ma compréhension, il trace une adéquation entre monisme et absolutisme (ou « dogmatisme ») en éthique. Le pluralisme serait donc une sorte de position mitoyenne entre l'absolutisme et le relativisme (Ess 2009, 167).[15] Plus récemment, Ess a spécifié que le pluralisme éthique axé sur les différences (*differences only*) présente le risque du relativisme (2020, 554), ce qui ouvre la porte à ce qu'un penseur pluraliste puisse être relativiste également. Il s'agit d'un avis que je partage. Ce relativisme « […] nie activement l'existence ou la possibilité "de normes et de standards éthiques qui peuvent être contraignants et légitimes pour plus que l'individu et/ou un groupe ethnique spécifique" » [Traduction libre] (Ess 2020, 554). Autrement dit, il se peut très bien qu'un pluraliste soit relativiste même si, à mon sens, cette proposition ne peut être généralisée à l'entièreté de la tradition éthique.

Bernard Williams n'est pas tendre à l'endroit du relativisme, le décrivant comme « […] l'hérésie de l'anthropologue, peut-être le point de vue le plus absurde qui ait été avancé, même en

[15] Cela dit, comme je l'ai mentionné au début de ce chapitre, la conception même du pluralisme est entendue de manière légèrement différente pour Ess.

philosophie morale » [Traduction libre] (1993, 20). À mon sens, ce qui pourrait expliquer l'aversion de Williams pour le relativisme est précisément l'importance qu'il accorde aux pertes et au remords. La tragédie est une possibilité, dans le cas de conflits de valeurs, parce que les valeurs ne sont pas purement relatives. Si les valeurs défendues par les gens étaient, somme toute, relatives, il deviendrait logiquement possible de *relativiser* les pertes. Or, si la tragédie, les pertes, le remords et la honte sont non seulement possibles, mais potentiellement déchirants, c'est parce que les pertes sont réelles et, en un certain sens, « absolues ».

Pour certains penseurs pluralistes, les relativistes confondent le fait que les sociétés aient entre elles des valeurs qui diffèrent, avec le principe selon lequel rien ne peut être dit ou fait à cet égard (Williams 1993, 23). Si on peut trouver une certaine relativité des valeurs chez les pluralistes, elle dépend, affirment-ils, de plusieurs choses : cette relativité n'est pas absolue. C'est précisément ce qu'ils rejettent de cette position métaphysique (Hampshire 1989, 62-63). Autrement, tout devient une question d'opinion et il n'est plus nécessaire d'argumenter, de négocier entre les tenants de valeurs différentes et conflictuelles. Ou encore, étant donné l'état de fragmentation des valeurs, la violence pourrait devenir un moyen de les asseoir (Ess 2020, 555-556). De plus, le relativisme commet une faute logique bien connue, soit celle de tenir pour absolument vrai son énoncé selon lequel rien n'est absolument vrai. Plus encore, dans le contexte de la relativité des valeurs selon les cultures, en découle la nécessité de la tolérance comme valeur universelle. Cependant, l'une des prémisses de l'éthique relativiste est que justement, il n'y a pas de normes universelles. Conséquemment, la tolérance devrait être relative, elle aussi (Ess 2009, 184).

Les pluralistes soutiennent au contraire que le fait de reconnaître la pluralité des cultures et des modes de vie ne renvoie pas automatiquement au relativisme culturel. Berlin le présente ainsi :

> je préfère le café, tu préfères le champagne. Nous avons des goûts différents, il n'y a plus rien à dire. C'est du relativisme. Mais le point de vue de Herder et de Vico, n'est pas cela : c'est ce que je devrais décrire comme le pluralisme — c'est-à-dire *la conception qu'il y a beaucoup de buts différents que les hommes cherchent tout en demeurant pleinement rationnels, pleinement hommes, capables de se comprendre et de sympathiser et de s'éclairer mutuellement*, comme nous le faisons en lisant Platon ou encore les romans du Japon médiéval — un monde, des perspectives qui sont très loin des nôtres. [Traduction libre, je souligne] (1990, 11).

Une manière d'interpréter ce que Berlin argumente ici est que le relativisme évacue la nécessité de dialoguer, voire d'apprendre les uns des autres. Dans une perspective purement relativiste, discuter de ses préférences morales pourrait constituer, au mieux, une perte de temps.

Plus encore, il y a chez les pluralistes un élément tenu pour commun avec le reste des humains, comme « un contenu universel à la moralité » (Apfel 2011, 18). Cela étant dit, les valeurs de ce « contenu universel » ne sont pas unifiées ni figées une fois pour toutes :

> ces valeurs, concepts et théories ne sont pas immuables, mais évoluent et changent graduellement avec le temps. C'est la raison pour laquelle nous ne pouvons les décrire que comme *presque* universelles : il demeure toujours la possibilité que notre expérience prouve le contraire, et c'est une possibilité qui doit toujours être prise en considération. [Traduction libre] (Apfel 2011, 18-19)

Si les concepts et les théories changent avec le temps, leurs noyaux sont toutefois immuables. Il est vrai que cette façon de voir n'est pas loin de certaines propositions d'éthique de la vertu contemporaines, que nous avons mentionnées précédemment, comme celles de Nussbaum (1990) et de Vallor (2016) — tout en maintenant les différences déjà exposées entre l'éthique de la vertu et le pluralisme. Tout de même, Berlin affirme qu'il existe « un monde de valeurs objectives », puisque la pluralité de principes moraux n'est pas infinie (1990, 11-12). Sur la base de la rationalité des êtres humains (une prise de position inspirée d'Aristote et de Kant), il soutient que

> […] la reconnaissance de certaines valeurs — aussi générales et aussi peu nombreuses soient-elles — entre dans la définition normale de ce qui constitue un être humain sain. [Traduction libre] (Berlin 1964, 222)

La situation du pluralisme des valeurs par rapport au relativisme est donc clairement différente, mais tout de même ambiguë. Ce qui contribue peut-être à brouiller les cartes, c'est la réalité de la pratique. Ainsi, dans la pratique éthique, on assistera seulement à « des consensus occasionnels » (Hampshire 2011, 115-116). Ces consensus sont généralement des accommodements, et non de réels rapprochements. Ils ne sont pas durables : ils sont toujours en équilibre précaire, menacé de s'effondrer, en grand besoin d'ajustements (Berlin 1990, 20).

Enfin, on pourrait aussi soutenir que les pluralistes se distinguent des relativistes par l'importance accordée au choix entre les valeurs qui sont en conflit. L'engagement de la raison

(pratique) marque donc une différence importante entre les deux — du moins en ce qui concerne les pluralistes non décisionnistes. Pour ces derniers, même le fait de s'abstenir de choisir constitue une forme de choix : c'est-à-dire qu'il n'est pas possible de rester passif ou paralysé devant le conflit, même s'il est tragique (Apfel 2011, 20-21). De là l'appel à la raison pratique, qui contribue à choisir, dans le contexte, la « meilleure » option — ou, pour employer une terminologie plus typiquement pluraliste — la moins mauvaise. Cet attachement à la raison pratique est l'une des différences importantes que l'on retrouve entre le pluralisme des valeurs, d'un côté, et les théoriques éthiques conséquentialistes et déontologiques de l'autre. Ici, le monisme et le pluralisme sont clairement en porte à faux et c'est dans cette direction que mon propos aimerait entraîner le lecteur.

g) La critique du pluralisme envers le conséquentialisme et l'éthique déontologique

Le constat de la pluralité des valeurs et des tensions entre elles fait que les pluralistes sont pour le moins insatisfaits par les théories éthiques contemporaines que sont le conséquentialisme et le déontologisme (ou encore, à ce titre, par la doctrine de l'unité des vertus). Ces approches sont problématiques pour plusieurs raisons — leur monisme principalement —, mais aussi parce que, dans la perspective pluraliste, elles ne possèdent pas les ressources qui conviennent au monde réel. En fait, les pluralistes cherchent à montrer qu'il est vrai que, dans un document, ou dans un raisonnement « abstrait », « théorique », il est difficile d'appréhender les collisions potentielles des valeurs ou principes moraux. C'est plutôt au cœur de l'action que les conflits émergent, dans le temps et l'espace (Williams 1994, 118). Formulé autrement, il peut sembler clair « sur papier » à des théoriciens éthiques que les valeurs et les vertus s'harmonisent, mais les choses vont souvent différemment dans la réalité de la vie pratique.

Williams déplore le fait que les théories morales comme le conséquentialisme et le déontologisme participent d'une philosophie morale qui soit « […] gouvernée par le rêve d'une communauté de raison trop éloignée […] de la réalité sociale et historique, ainsi que de tout sens concret d'une vie éthique particulière » [Traduction libre] (Williams 1985, 197). Cette « communauté de raison » ne tient pas compte de l'aspect parfois irrationnel de la moralité, soutient

Williams, qui fait que nous éprouvions des remords alors que, « rationnellement parlant », nous ne devrions pas, étant donné que nous avons fait tout ce qui était en notre pouvoir pour faire du bien ou éviter du mal. Conséquemment, ces mêmes théories morales ne feraient pas une place suffisante au sentiment de culpabilité et traiteraient le conflit moral comme quelque chose de contingent, et non de nécessaire à la vie humaine éthique (Williams 1994, 109-110). Puisque ces théories présentent les conflits comme étant contingents, la conséquence logique serait qu'ils seraient surmontables. Au contraire, d'affirmer Williams, « on ne peut dire des conflits moraux qu'ils puissent être systématiquement évités, ni qu'ils puissent être entièrement résolus sans laisser de traces » (1994, 107). Si ces théories monistes reconnaissent effectivement la possibilité du conflit, elles postulent toutefois que l'agent peut garder « les mains propres », moyennant le respect de la procédure décisionnelle.

Parallèlement, les théories éthiques modernes auraient sous-estimé l'importance des attachements, des projets et des liens qui existent entre les personnes, ainsi que celle de la fortune (au sens de chance), sur laquelle personne n'a de contrôle. La malchance peut — on le voit dans plusieurs tragédies grecques, sur lesquelles se sont penchés Williams et Nussbaum (2016) — générer des situations de collisions entre des valeurs (Nussbaum 2009, 213). Ce n'est pas en ignorant les conflits de valeurs qu'ils disparaîtront, car ils sont impossibles à éliminer. Le recadrage de la situation ou de notre positionnement par rapport à la situation conflictuelle ne peut pas non plus nous débarrasser complètement de ces conflits (Williams 1994, 115).

Les traditions éthiques déontologique et conséquentialiste, en raison de leur monisme, hiérarchisent les valeurs entre elles, selon un ordre interne. Or, soutient Berlin,

> supposer que toutes les valeurs peuvent s'ordonner sur une seule échelle, de sorte qu'il ne s'agirait plus que de déterminer laquelle est la plus haute, me semble aller à l'encontre de ce que nous savons : les hommes sont des agents libres; c'est aussi se représenter la décision morale comme une opération qu'une simple règle à calcul [ou un algorithme] pourrait accomplir. (1988, 217-218)

Ce reproche sied particulièrement au conséquentialisme, au cœur duquel les notions de hiérarchies formelles de valeurs et de calcul sont centrales. En vérité, ni le conséquentialisme ni l'éthique déontologique kantienne ne rendent compte du déchirement que vit souvent chaque être humain dans les dilemmes moraux. La complexité du réel ne peut être saisie exhaustivement par une

théorie. La multiplicité des valeurs, la « situation » du sujet pensant et les contextes particuliers à chaque individu ne pourront jamais être complètement englobés par une théorie morale. Ces théories sont, en quelque sorte, « plaquées » sur le réel : elles n'y collent pas vraiment, ou encore elles sont « procrustiennes ». La réalité est tissée de contrastes que les théories ne pourront jamais complètement rendre, argumente Williams :

> […] presque toute vie humaine digne d'intérêt se situe entre les extrêmes que nous offre la moralité. Elle met fortement l'accent sur une série de contrastes : entre force et raison, persuasion et conviction rationnelle, honte et culpabilité, aversion et désapprobation, davantage de rejet et de reproches. [Traduction libre] (Williams 1985, 194)

La moralité, elle, ne relève pas de la croyance, elle n'est pas « une invention des philosophes », mais le « point de vue » qu'ont la plupart des personnes sur des enjeux éthiques (Williams 1985, 172). Les questions morales sont complexes. Les valeurs peuvent entrer en conflit et l'on pourrait dire que l'utilitarisme, avec son attention au résultat, ainsi que le déontologisme, avec son insistance sur les moyens, ont sous-estimé ces conflits. On verra un peu plus loin que les pluralistes, quant à eux, ont tendance à les surestimer.

Une autre critique des pluralistes à l'endroit des conséquentialistes et des déontologistes concerne non seulement le recours à une théorie éthique, mais son élaboration même. Le théoricien, ce faisant, transposerait les conflits de valeurs (une réalité du monde de la pratique) au plan théorique. Or,

> c'est une erreur à plus d'un titre. Si le conflit entre nos valeurs n'est pas nécessairement pathologique et si, même lorsque la situation est viciée comme dans certains conflits d'obligation, *le conflit n'est pas dû aux troubles logiques de notre pensée*, ce doit être une erreur que de considérer le besoin d'éliminer le conflit comme une exigence purement rationnelle, une de ces exigences qui est justifiable d'un système théorique. [je souligne] (Williams 1994, 294)

Ainsi, on pourrait presque trouver chez les pluralistes des valeurs un reproche de positivisme mal placé à l'endroit des théories éthiques de la modernité, doublé d'une importance exagérée accordée aux croyances (Williams 1985, 93). Les théories éthiques sont perçues comme étant simplificatrices et idéalistes par les pluralistes des valeurs. À cet égard, Isaiah Berlin soutient que

> la notion de l'ensemble parfait, de la solution ultime, dans laquelle toutes les bonnes choses coexistent, [lui] semble non seulement irréalisable — c'est un truisme —, mais

> incohérente sur le plan conceptuel; [il] ne sai[t] pas ce qu'on entend par une harmonie de ce genre. [Traduction libre] (Berlin 1990, 14)

Plus encore, une recherche mal placée d'une forme de perfection inatteignable peut être dissimulée derrière des théories morales, comme celle d'Aristote, que récuse Hampshire, ou encore celle de Kant. Il faudrait plutôt dire, avec Berlin, que « nous ne pouvons faire que ce que nous pouvons : mais cela, nous devons le faire, malgré les difficultés » [Traduction libre] (Berlin 1990, 19-20).

h) Critiques adressées au pluralisme des valeurs

Comme toutes les traditions éthiques abordées dans cette thèse, le pluralisme des valeurs peut également se voir adresser des critiques. Certes, la courte analyse ici présentée est loin d'être exhaustive, mais elle donne une idée de points de faiblesse qui ont été relevés dans cette approche à l'éthique. On a vu que la centralité du conflit implique la centralité de la perte. Pourtant, cette place à « la somme nulle » peut être critiquée comme ouvrant la porte à une certaine exagération des différences ou de la distance entre les parties impliquées (Blattberg 2016, 15-17). Ce point de départ peut rendre le dialogue — peu importe la forme qu'il prend — plus difficile entre les interlocuteurs, puisqu'il est amorcé en plaçant l'accent sur l'irréductibilité des valeurs en tension, plutôt que sur ce qui est partagé, commun. Ainsi, on pourrait reprocher aux pluralistes une sorte de pessimisme tenace, un sens de la tragédie quelque peu moussé, qui trouve son expression dans des propos comme ceux de Berlin : « nous sommes *condamnés à choisir* et chaque choix peut entraîner une *perte irréparable* » [Traduction libre, je souligne] (Berlin 1990, 14). Il faut toutefois noter que le même philosophe ajoute un peu plus loin qu'étant donné les larges terrains d'entente entre les gens sur des questions morales, l'incompatibilité des valeurs ne doit pas être dramatisée (Berlin 1990, 19). Cette affirmation tempère quelque peu les propos, mais le point de départ de la réflexion demeure la fragmentation et une insistance est bel et bien mise sur le conflit et la perte.

Alors que Martha Nussbaum partage l'intérêt de Williams pour les dilemmes éthiques tels qu'ils sont exposés dans les tragédies grecques, elle critique ce dernier en ce que, sans prêcher la résignation à la tragédie, il s'en approche beaucoup par son silence sur des réponses possibles à nos problèmes, lorsque nous faisons face au monde tel qu'il est (Nussbaum 2009, 213-214, 217,

220). Les tragédies grecques n'étaient pas écrites dans le but de rendre leur public pessimiste ou désabusé face à la vie, mais plutôt de l'outiller à s'examiner et à réfléchir davantage, selon Nussbaum. Elle affirme donc que « […] son histoire du mal est trop belle pour être vraie » [Traduction libre] (Nussbaum 2009, 232). Nussbaum se demande où est la place de l'optimisme en éthique, sans pour autant tomber dans une sorte d'idéologie du progrès téléologique (2009, 237) :

> [elle] croit qu'il est excitant de penser à la façon dont nous pourrions créer un espace d'espoir au milieu de la laideur. En effet, c'est la contemplation de l'inaltérable qu'[elle] trouve plus proche de l'ennui, car qu'est-ce que la vie humaine, de toute façon, mais faire de son mieux avec les options dont on dispose réellement? [Traduction libre] (2009, 235)

En définitive, le pluralisme des valeurs peut se voir reprocher une forme de pessimisme devant les différences de valeurs, une apathie quant à la possibilité de les surmonter, ainsi qu'une résignation au scénario « le moins mauvais » : à se contenter, en somme, de « réparer les dégâts ».

Conclusion

Le pluralisme des valeurs est une approche à l'éthique qui présente plusieurs possibilités aux décideurs contemporains, en ce qu'il permet justement une cohabitation particulière de la diversité des valeurs, ainsi qu'une flexibilité dans sa démarche, qui ne relève pas d'une procédure formelle. On verra dans la prochaine section qu'effectivement, les penseurs de l'éthique y ont recours avec une réelle fréquence. Cependant, il n'est que rarement nommé. Méconnu dans la pensée éthique contemporaine, son influence est à mon sens réel, et mérite d'être documentée. C'est pour cette raison que dans la deuxième section de la thèse, je propose au lecteur une interprétation analytique — ou une analyse interprétative — d'une sélection de directives éthiques concernant l'intelligence artificielle. Parues lors des dernières années — depuis 2016 —, le regard sur ces documents permettra un « portrait sociologique » de l'éthique de l'IA dans un certain pan de la littérature, qu'il convient à présent d'explorer.

SECTION 2 : PORTRAIT

Chapitre 4 — Démarches monistes

> « *Le problème n'est pas seulement une question de pouvoir; il concerne aussi le bien : le bien pour les individus et le bien pour la société. Nos idées actuelles sur le bien de la vie et de la société, si nous pouvons les articuler toutes, pourraient bien avoir besoin d'une discussion beaucoup plus critique*. » — Mark Coeckelbergh [Traduction libre] (2020a, 177)

Introduction

Dans cette deuxième section de la thèse, je propose au lecteur une analyse d'une vingtaine de directives éthiques concernant les enjeux de l'intelligence artificielle. Alors que la première section esquissait sommairement les fondements du monisme et du pluralisme, celle-ci propose de repérer ces tendances à l'œuvre dans une sélection de démarches éthiques publiées du milieu à la fin des années 2010. À ma connaissance, le portrait que je dresse n'a pas encore été amené. Plus encore, il relève d'une démarche de philosophie politique, et se démarque ainsi, tant dans le fond que dans la forme, des autres analyses de directives éthiques telles que présentées dans la revue de littérature, au premier chapitre. En revanche, il faut préciser que ce portrait n'est pas exhaustif, puisque l'échantillon est assez limité. Il est fait dans l'optique de présenter, aux décideurs politiques qui devront légiférer sur l'éthique de l'IA, un tour d'horizon des écoles éthiques récurrentes dans les directives qui sont publiées par des organisations de divers secteurs de la société. La proposition éthique et politique que je mets de l'avant dans la troisième section de la thèse ne dépend pas de cette analyse du contenu métaéthique des directives, mais elle se comprend mieux une fois ce travail de défrichage effectué.

Dans ce chapitre, je présenterai en premier lieu l'échantillon de directives éthiques que j'ai sélectionnées pour cette analyse métaéthique, ainsi que le justificatif l'accompagnant. Je jetterai un regard comparatif sur les destinataires de ces démarches. En second lieu, je présenterai, tour à tour, les démarches éthiques informées par l'éthique de la vertu, l'utilitarisme et l'éthique déontologique. Pour chacune de ces traditions monistes, je procéderai en deux temps. D'abord, j'effectuerai une esquisse de ce dont une démarche éthique informée par l'école en question pourrait avoir l'air, ou

se présente dans la littérature existante. Ensuite, j'analyserai les passages des directives me semblant attachés à la tradition éthique dont il est question. Dans le cas du déontologisme, la section sera plus longue, car j'y ajoute la question incontournable, en éthique de l'IA, du dilemme du tramway, que j'ai décidé de traiter dans cette sous-section. J'ai également développé plus amplement l'esquisse d'une éthique déontologique face à l'IA selon son auditoire, ce qui me semblait nécessaire dans le cas de l'éthique déontologique spécifiquement, à des fins de clarté.

1. Présentation des directives sélectionnées

À la suite de la lecture de plus d'une centaine de documents traitant de l'éthique de l'intelligence artificielle, de provenance industrielle, académique, civile, gouvernementale, nationale et internationale, j'ai tiré un échantillon aux fins de cette analyse. Il est constitué d'une vingtaine de documents éthiques, parmi les plus mentionnés dans la littérature sur l'éthique de l'IA, dans les documents qui « s'entre-citent » ou encore provenant d'acteurs influents dans le domaine. Il s'agit par ailleurs de démarches qui sont citées comme étant emblématiques, par les équipes de chercheurs ayant fait des analyses de contenu des directives éthiques en IA.[16] La sélection est séparée en trois catégories. Ces dernières sont inspirées de la méthode de classification de directives éthiques que propose le projet « Principled Artificial Intelligence » abrité par l'Université Harvard. On y trouve des documents et directives émanant 1) des entreprises privées, 2) d'instances gouvernementales, intergouvernementales ou internationales et 3) de la société civile, ainsi que d'organisations regroupant de multiples partenaires[17] (Fjeld et al. 2019). Ces initiatives éthiques s'étendent sur la période de 2016 à 2020. L'année 2016 est importante pour l'éthique de l'IA, en ce qu'elle marque une éclosion de directives, qui prennent des formes aussi diverses que « […] des codes d'éthique, des principes, des lignes directrices, des cadres et des

[16] Par exemple, dans leur analyse de 88 directives, Schiff et al. (2020, 154-156) nomment l'UE, l'OCDE, le G20, Microsoft, Google, IBM, le IEEE, le FLI, le PAI, les principes d'Asilomar, ainsi que les Déclarations de Montréal et de Toronto, que j'ai tous inclus dans mon échantillon.

[17] La classification de Fjeld et al. parle de « parties prenantes » (« *stakeholders* ») plutôt que de « partenaires ». Cependant, étant donné la connotation pluraliste de l'expression « partie prenante », et dans le désir de ne pas fausser l'analyse d'emblée avec un type de vocabulaire, j'ai choisi de le traduire de cette façon. De leur côté, Schiff et al. (2020) divisent les documents entre le secteur privé, le secteur public et le secteur des organisations non gouvernementales (ONG). Plusieurs façons de diviser les documents existent et possèdent toutes certainement une part d'arbitraire — c'est donc aussi le cas de la mienne.

stratégies politiques » [Traduction libre] (Schiff et al. 2020, 153). La liste des documents sélectionnés, incluant leur année de publication, destinataires et type de document, figure au Tableau 1 en annexe de la thèse.

Il est important de souligner que j'ai retenu les documents d'abord pour leur pertinence, et non parce que je cherchais un nombre fixe de directives par catégorie. En réalité, la catégorisation des vingt documents en trois catégories s'est faite à la suite de l'échantillonnage, aux fins d'ordre et de compréhension. Par ailleurs, le but de cette section « portrait » n'est pas de recenser, de manière aboutie, les différences entre les secteurs privé et public, bien que l'analyse qui suit en suggérera quelques idées. Cela dit, la ligne entre le public et le privé est souvent difficile à tracer, surtout quand il est question de groupes composés de multiples partenaires.

À cet égard, l'organisation Partnership on AI [PAI] est plutôt particulier. Il s'agit d'une organisation de multiples partenaires, émanant du secteur privé. Autrement dit, ce sont des chercheurs en intelligence artificielle qui ont mis sur pied, en 2016, cette organisation qui, aujourd'hui, regroupe de multiples partenaires de toutes sortes d'horizons[18]. À son origine, toutefois, ce sont des chercheurs de DeepMind, Google, Apple, Microsoft, IBM, Amazon et Facebook qui l'ont fondée (Partnership on AI [PAI] s.d.a, s.p.). Sa genèse se retrace donc dans le secteur privé, mais je la place, aux fins de cette analyse, dans la catégorie des organisations à multiples partenaires. J'ai fait ce choix en raison de son activité et de sa définition d'elle-même, bien que les autres documents de cette catégorie soient associés au secteur public. On voit bien les limites de toute catégorisation lorsque des documents se retrouvent entre deux groupes, ou plus. De même, comme l'IA est un enjeu à la jonction entre le secteur privé et le secteur public, ce genre de chevauchement n'est pas surprenant. Il est même souhaité par une grande quantité de rédacteurs de directives éthiques.

[18] Il est toutefois intéressant de mentionner qu'en octobre 2020, l'organisation Access Now a quitté le PAI en lui reprochant un manque d'espace pour l'implication de la société civile en son sein, ainsi qu'une difficulté à prendre position sur des questions pressantes comme le rejet des technologies de reconnaissance faciale, en raison de « […] l'absence de consensus et [d]es divergences de vues radicales entre les parties prenantes […] » [Traduction libre] (Access Now 2020, §2, 1).

Ainsi, mon échantillon se divise en six directives associées à la société civile ou aux organisations « multipartenaires », six tombant dans la catégorie de gouvernance internationale ou de relations intergouvernementales internationales, et huit relavant des entreprises privées ou du monde industriel. La pertinence des documents a été évaluée au regard de leur importance potentielle pour des décideurs politiques, de quelque pays qu'ils soient, qui sont et seront appelés à légiférer en éthique de l'IA. D'une part, les orientations éthiques des entreprises privées, responsables de la conception et du déploiement des SIA, constituent un incontournable pour les décideurs politiques. La raison est que les experts techniques de l'IA se trouvent principalement dans la sphère privée, en plus des cercles académiques. Ce sont eux qui détiennent la clé du fonctionnement et la compréhension (relative, nous l'avons vu) des systèmes d'intelligence artificielle. Plus encore, les entreprises du secteur privé ont une part considérable de pouvoir et d'influence dans la conception, le développement et le déploiement des SIA.

Il faut ajouter que les stratégies politiques, industrielles ou économiques nationales, c'est-à-dire propres à un pays donné, n'ont pas été retenues pour mon analyse, bien que je les ai recensées dans le cadre de ma recherche préliminaire. Il en existait plus d'une trentaine au moment d'écrire ces lignes,[19] en provenance des quatre coins du monde. Étant spécifiques à un contexte politique et économique donné, elles présentent un intérêt réel, mais limité, parce que leur portée est restreinte. En outre, il difficile de justifier quel(s) pays choisir pour analyser les initiatives nationales dans une étude comme celle-ci. Dans cet ordre d'idées, je n'ai pas non plus retenu les codes d'éthique attachés à une profession particulière, qui ne traiteraient pas principalement de l'IA (par exemple, celui de l'Association for Computing Machinery [ACM] 2018). Les documents portant tout spécialement sur la robotique (European Parliament 2016), par exemple sur l'utilisation des drones (van Wynsberghe et al. 2018), n'ont pas été sélectionnés, sauf dans le cas d'un rapport des Nations Unies traitant spécifiquement de l'éthique de la robotique, et non seulement de lois à ce sujet.

Parallèlement, les rapports traitant d'enjeux très spécifiques comme l'usage de données massives (*big data*) (United Nations Development Group [UNDG] 2017) ou encore seulement de l'intelligence artificielle généralisée (« *AGI* ») (OpenAI 2018a) ne font pas non plus partie de la sélection. Il en va de même pour les études d'instituts universitaires (Université Stanford 2016;

[19] C'est-à-dire lors de l'année académique 2019-2020.

Campolo et al. 2017) ou de chercheurs externes au profit d'institutions internationales (Mantelero 2018), les réflexions de particuliers concernant l'éthique de l'IA (Copeland 2018), ainsi que les études ou projets menés par des organisations internationales portant sur des enjeux très précis de l'intelligence artificielle (Forum économique mondial [FEM] 2019b[20]), ou qui ne sont pas directement en lien avec l'éthique de l'IA tout en y étant pertinents (Turek s.d.).

De la même manière, des textes de loi comme le Règlement général sur la protection des données de l'Union européenne (RGPD) (European Union 2018a) ne font pas non plus partie de l'échantillon, puisqu'ils ne consistent pas en des directives éthiques à proprement parler — même si ces dernières, on l'a vu, prennent des formes très diversifiées. Des instituts pour un développement souhaitable de l'IA n'ont été retenus que dans la mesure où ils avaient produit des projets ou études sur l'éthique de l'IA, comme le Partnership on AI [PAI], et qu'ils semblaient actifs et à jour.[21] Enfin, la Commission européenne ayant publié plusieurs documents sur l'IA, je n'ai retenu que ceux qui étaient le plus cités dans la littérature grise et académique, et laissé de côté d'autres prises de position (par exemple, European Commission 2018b; 2018a; 2018c; 2019a)[22]. De même, j'ai inclus un document des Nations Unies sur l'IA, mais je n'ai pas sélectionné les rapports de conférences annuelles sur l'IA (ITU & the XPRIZE Foundation 2017; 2018).

En revanche, les études et chartes d'entreprises comme Google, Microsoft et IBM sont à l'étude, ainsi que les positionnements éthiques de centres de recherche en IA comme OpenAI, DeepMind ou encore ElementAI au Canada.[23] Les intérêts de ces acteurs du marché ont une

[20] C'est aussi le cas des divers projets du Forum économique mondial comme le « Centre for the Fourth Industrial Revolution » (s.d.a), « Teaching AI Ethics [TAIE] » (s.d.d), « Project on Artificial Intelligence and Machine Learning » (s.d.c) ou encore « Project Empowering AI Leadership » (s.d.b), pour n'en nommer que quelques-uns.

[21] C'est la raison pour laquelle je n'ai pas gardé le projet de AI Global 2017, par exemple. Il faut aussi noter que, depuis l'écriture de ces lignes, le PAI a lancé son projet de « Publication Norms for Responsible AI », que je n'ai pas analysé pour cette thèse.

[22] On notera aussi que, en ce qui concerne le gouvernement de l'Europe, le Parlement européen a voté en octobre 2020 sur la réglementation de l'IA, « […] pour que l'UE devienne un chef de file mondial dans le développement de l'IA » (Gerlat 2020, §3). La Commission européenne serait donc chargée de présenter « […] un nouveau cadre juridique définissant les principes éthiques et les obligations juridiques à respecter lors du développement, du déploiement et de l'utilisation de l'IA […] » (*Ibid*., §4). La Commission a par ailleurs publié un rapport sur l'IA dans les services publics en juillet 2020, que je n'ai pas non plus analysé aux fins de cette recherche.

[23] ElementAI a une certaine notoriété en raison d'un de ses fondateurs, le professeur Yoshua Bengio. Une petite mise à jour s'impose depuis le choix de l'échantillon et son analyse. La compagnie ElementAI a été vendue à la firme américaine ServiceNow en automne 2020, et a fait l'objet de quelques critiques concernant la quantité de financement québécois qui lui avait été octroyée dans le but d'œuvrer pour la collectivité québécoise (Bordeleau 2020; Gingras et Colleret 2020).

incidence sur les politiques gouvernementales de plusieurs pays donnés. En effet, des entreprises comme Microsoft sont les premières à demander un dialogue accru entre les décideurs politiques du secteur public et les experts en IA du secteur privé (Microsoft 2018b). C'est donc dans l'optique où ces chartes traitaient directement d'IA, et ne constituaient pas seulement un code général de conduite pour les employés de l'entreprise qu'elles ont été sélectionnées.

Les directives éthiques émanant des sphères publiques sont légion. À mon sens, un décideur politique s'intéressant à l'IA, à un endroit donné dans le monde, pourrait fort bien être interpellé par des initiatives d'organismes gouvernementaux, intergouvernementaux ou internationaux. C'est pour cette raison que l'Organisation des Nations Unies pour l'Éducation, la Science et la Culture (UNESCO) — la branche de l'Organisation des Nations Unies étant déléguée pour se pencher sur l'éthique robotique et des SIA — fait partie de l'échantillon, ainsi que la Commission européenne et son groupe d'experts de haut niveau sur l'IA, constitué en juin 2018. Les pays du G20, le Forum Économique Mondial (FEM) et l'Organisation de coopération et de développement économiques (OCDE) figurent aussi dans la sélection, en raison de leur influence et de leur résonnance sur les gouvernements nationaux dans leur contexte propre. Il est particulièrement intéressant de tenir compte du fait que le FEM constitue une sorte de trait d'union entre le monde gouvernemental et industriel.

Enfin, des projets de directives éthiques émanant soit de la société civile, soit de partenaires multiples (c'est-à-dire en provenance de différentes sphères de la société), sont aussi analysés à cause de leur notoriété dans le domaine de l'éthique de l'IA. Ce degré d'incidence peut résider dans le fait que ces documents ont souvent été élaborés de manière collaborative, mobilisant, grâce à eux, une grande quantité de personnes d'horizons différents (Schiff et al. 2020, 155). Le projet de l'année 2020 du Partnership on AI (PAI), pour une IA éthique, les deux études préliminaires du Institute of Electrical and Electronics Engineers (IEEE) sur l'alignement éthique dans la conception des systèmes autonomes et intelligents,[24] les vingt-trois principes d'Asilomar du Future of Life Institute (FLI), la Déclaration de Toronto et la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle forment la dernière catégorie de l'échantillon à l'étude.

[24] Depuis la parution des deux études du IEEE *Ethically Aligned Design*, en 2016 et 2017 respectivement, une première édition d'un rapport complet a été publiée après révisions (IEEE 2020 s.p.; IEEE 2019).

Comme il fallait s'y attendre, dans cette catégorie, on trouve une combinaison de documents des secteurs publics et privés.

a) Les destinataires des directives éthiques

L'auditoire des directives éthiques recensées diffère à quelques égards, mais tend à se recouper avec une certaine fréquence. On y interpelle fréquemment les concepteurs des SIA, de même que les gouvernements et la société civile. Parfois, les destinataires ne sont pas spécifiquement mentionnés dans la directive (Schiff et al. 2020, 155 et Jobin, Ienca et Vayena 2019, 2 émettent le même constat), comme dans le cas du positionnement d'ElementAI ou des principes publiés par Microsoft (Gagné 2019, s.p.; Microsoft 2018a). Cela dit, on peut tenir pour acquis que ces prises de position éthiques s'adresseraient non seulement au grand public, mais aux clients de l'entreprise qui les adopte. Dans l'échantillon, une étude de Google cible les gouvernements et la société civile (Google 2019a); c'est le cas aussi du positionnement éthique d'IBM sur l'IA (International Business Machines Corporation [IBM] 2019). La charte de DeepMind interpelle les concepteurs et développeurs en IA, afin qu'ils travaillent dans l'optique de créer de bons impacts avec leurs technologies (DeepMind 2017). Microsoft, dans son étude approfondie du rôle de l'IA dans la société, interpelle directement « les chefs d'entreprises, les décideurs politiques, les chercheurs, les universitaires et les représentants de groupes non gouvernementaux […] » [Traduction libre] (Microsoft 2018b, 57). Un écho comparable se remarque dans la démarche de la coalition Partnership on AI, qui s'adresse aux « […] acteurs du changement, aux militants et aux décideurs politiques qui travaillent au développement et à la mise en place d'une IA responsable » [Traduction libre] (Cavello 2020, s.p.).

Du côté des organisations internationales et de celles regroupant plusieurs partenaires, on retrouve un pot-pourri semblable de destinataires. Par exemple, les deux études du IEEE visent les « technologistes », définis largement comme « […] toute personne impliquée dans la recherche, la conception, la fabrication ou la diffusion de messages autour de l'IA/AS, y compris les universités, les organisations et les entreprises qui font de ces technologies une réalité pour la société » [Traduction libre] (Institute of Electrical and Electronics Engineers, Incorporated (IEEE) 2016, 4). Sans mentionner précisément les décideurs politiques, on peut toutefois les inclure dans le groupe

de personnes impliquées dans la diffusion de messages autour de l'IA et des systèmes autonomes. Les rédacteurs de la Déclaration de Montréal destinent aussi leur propos à un auditoire très large, soit

> toute personne, toute organisation de la société civile et toute compagnie désireuses de participer au développement de l'intelligence artificielle de manière responsable, que ce soit pour y contribuer scientifiquement et technologiquement, pour développer des projets sociaux, pour élaborer des règles (règlements, codes) qui s'y appliquent, pour pouvoir en contester les orientations mauvaises ou imprudentes, ou encore pour être en mesure de lancer des alertes à l'opinion publique quand cela est nécessaire. Elle s'adresse également aux responsables politiques, élus ou nommés, dont les citoyens attendent qu'ils prennent la mesure des changements sociaux en gestation, qu'ils mettent en place rapidement les cadres permettant la transition numérique pour le bien de tous, et qu'ils anticipent les risques sérieux que présente le développement de l'IA. (Comité d'élaboration de la Déclaration de Montréal IA Responsable 2018a, 6)

Puis, les principes d'Asilomar ont pour objectif d'interpeller les chercheurs en robotique et la société civile plus largement (Future of Life Institute (FLI) 2017), tout comme le rapport de l'UNESCO sur la robotique et l'IA (Commission mondiale d'éthique des connaissances scientifiques et des technologies 2017) et celui du Groupe d'experts de haut niveau de l'Union européenne (GEHN IA) (2019). La Déclaration de Toronto, à l'instar de l'étude du Forum économique mondial (FEM) sur la gouvernance de l'IA, s'adresse à la fois aux développeurs et aux décideurs politiques (Bacciarelli et al. 2018; World Economic Forum [WEF] 2019a). Enfin, le positionnement de la Commission européenne (European Commission 2018c), les lignes directrices du FEM concernant les marchés publics en IA (2019b) ainsi que les principes de l'OCDE (OECD 2019) visent principalement les décideurs politiques.

Une raison pouvant contribuer à expliquer cette panoplie de destinataires des documents apparaît plus clairement au fil de la lecture de plusieurs directives éthiques. Rares sont les documents qui comportent une précision exacte de ce qui est entendu clairement comme « éthique de l'intelligence artificielle ». En effet, le champ d'études est en plein développement. Comme je l'ai exposé d'entrée de jeu dans le premier chapitre, si l'on admet que l'éthique peut être 1) celle des principes programmés dans les machines ou SIA, 2) celle des concepteurs et développeurs, 3) celle des utilisateurs de SIA ou 4) celle pour les décideurs politiques qui encadrent la conception, le déploiement et l'utilisation de telles technologies, alors on remarque souvent que ces quatre groupes de destinataires sont quelque peu entremêlés dans les documents.

Hagendorff voit dans le caractère abstrait des principes mis de l'avant une façon de rejoindre les divers secteurs de la société concernés par l'éthique de l'intelligence artificielle :

> en général, les lignes directrices en matière d'éthique postulent des principes très généraux et primordiaux qui sont ensuite censés être mis en œuvre dans un ensemble très diversifié de pratiques scientifiques, techniques et économiques, et dans des groupes parfois géographiquement dispersés de chercheurs et de développeurs ayant des priorités, des tâches et des responsabilités fragmentées toutes différentes. L'éthique opère donc à une distance maximale des pratiques qu'elle cherche effectivement à régir. [Traduction libre] (Hagendorff 2019, 8)

Lorsqu'on se penche sur le positionnement éthique de la jeune entreprise montréalaise ElementAI, on remarque qu'elle s'aligne sur les principes éthiques mis de l'avant par le groupe d'experts de la Commission européenne (Gagné 2019). Ici encore, donc, se dessine clairement le recoupement entre le secteur privé et le secteur public. Ce chevauchement est presque constant et, dans ce sens, il est difficile de délimiter absolument ce que l'on entend par « éthique de l'IA » pour les décideurs politiques. Une autre conséquence de la richesse de ce sujet est la diversité de formes et de contenus des directives éthiques, qui ne sont somme toute pas toutes conçues de la même façon.

b) L'analyse et la catégorisation des directives éthiques

La « méthode d'analyse » du contenu des directives formant l'échantillon à l'étude est qualitative. Autrement dit, aucun logiciel d'analyse de contenu n'a été employé pour ce travail. La lecture attentive des principes éthiques et de leurs passages explicatifs, s'il y a lieu, a permis d'identifier les traces des traditions éthiques potentiellement à l'œuvre. S'il fallait parler d'une « grille de lecture » des documents, on pourrait souligner que, par exemple, l'emploi de mots-clés comme « parties prenantes » ou « conflit » pouvait être un signe d'allégeance pluraliste, bien qu'il ne s'agisse évidemment pas d'un automatisme. Il en va de même pour les mots « système » ou « règle » pour le déontologisme, « vertu » ou « épanouissement » pour l'éthique de la vertu ou encore « maximisation » et « bonheur » pour l'utilitarisme, pour n'en nommer que quelques-uns. La lecture de vocabulaire employé constituait ainsi une piste de départ. La manière d'organiser les concepts entre eux achevait de compléter le tableau.

Plus simplement, il s'agissait de voir comment on se positionnait, dans les directives, par rapport à la question du conflit de valeurs. Parfois, il fallait découvrir comment était conçue l'origine de potentielles hiérarchies de valeurs (théoriques, ou pratiques). De même, la question de savoir si les valeurs étaient entendues comme commensurables ou, au contraire, incommensurables était un autre indice de la tradition éthique à l'œuvre. La section précédente, composée des chapitres deux et trois présentait les fondements des traditions éthiques dont il est question dans cette thèse. Cette section m'a outillée pour mener cette analyse, et sa lecture permettra maintenant au lecteur de me suivre dans son exposé.

2. Portrait des démarches éthiques monistes

a) L'éthique de la vertu

1. Esquisse d'une éthique de la vertu face à l'intelligence artificielle

Penser une éthique de la vertu pour l'intelligence artificielle — ou pour n'importe quelle autre nouvelle technologie — a déjà été fait par quelques penseurs ou contributeurs aux débats sur l'éthique de l'IA. Par exemple, le chercheur au « Montreal AI Ethics Institute » (MAIEI) Ryan Khurana propose d'adopter l'éthique de la vertu et, plus spécifiquement, les quatre vertus cardinales que sont la prudence, la tempérance, la justice et la force pour aborder le développement technologique, en évitant de se limiter à des listes de contrôle (2020, §9, 17). Le professeur de sciences cognitives et d'IA Travis J. Wiltshire avait, quelques années auparavant, suggéré que l'on programme des agents moraux artificiels qui se comportent de manière héroïque, en s'inspirant des vertus aristotéliciennes (2015). Le philosophe Martin Gibert (2020) a quant à lui mis de l'avant une « éthique des algorithmes » visant à rendre les robots vertueux. Il puise pour ce faire non seulement à l'éthique de la vertu, mais au déontologisme que l'on retrouve dans les propos de John Rawls. Gibert estime que chaque citoyen pourrait désigner entre une et trois personnes qu'il juge vertueuses, pour que l'on puisse agréger les réponses de ces dernières à des dilemmes moraux de programmation. Les réponses morales fournies se justifieraient par le fait de s'accorder sur des vérités lorsque l'on est sous le voile d'ignorance tel que le suggère Rawls. À son avis,

> sortir son voile d'ignorance, c'est viser un point de vue impartial qui donne des raisons morales d'agir. C'est viser le point de vue éthique. […] Où trouver les principes pour paramétrer les bons algorithmes? En regardant, lorsque c'est possible, sous le voile d'ignorance. (Gibert 2020, s.p.)

Une analyse plus approfondie a été menée par la philosophe de la technologie Shannon Vallor. Quelques bribes de son propos ont déjà été rapportées plus haut. On a vu que selon elle, le meilleur moyen de se préparer, comme individu et comme société à la « convergence technologique » (Vallor 2016, 27), est de cultiver les « vertus technomorales ». Ces vertus seraient nécessaires à notre épanouissement dans un monde en rapide transformation technologique. Les douze vertus technomorales que Vallor propose — bien qu'elles indiquées sous réserve de modifications — sont l'honnêteté, la maîtrise de soi, l'humilité, la justice, le courage, l'empathie, le « care », la civilité, la flexibilité, la « hauteur de vue » (ou capacité de perspective), la magnanimité, ainsi que la sagesse technomorale. Cette dernière vertu semble relever de la pétition de principe, soit participant d'un raisonnement circulaire. En effet, on peut se demander si la personne sage, au plan technomoral, n'exhibe pas déjà les onze autres vertus qu'elle mentionne. Peut-être la séparation des éléments est-elle proposée à des fins de clarté, et que la compréhension de ces vertus serait plus holiste qu'atomiste.

Une éthique de la vertu telle que la propose Vallor semble s'adresser aux citoyens, mais peut également être valide pour les décideurs politiques. En effet, ces derniers sont appelés à légiférer en tenant compte du contexte propre des gens qu'ils représentent. Une force de l'éthique de la vertu est qu'au lieu de se tourner vers des règles morales de plus en plus abstraites, comme celles que l'on retrouve dans les théories déontologique et utilitariste, elle est justement bâtie sur la prise en compte de ce contexte. À titre d'illustration, Vallor soutient que les attitudes envers les robots donneurs de soin sont très différentes en Occident et en Asie. Elle suggère aussi qu'il est possible

> […] que les Européens de tradition libérale accordent une valeur à la vie privée d'une manière qui, sur le plan théorique, est tout à fait différente de la manière dont les valeurs de la vie privée sont encadrées dans la société chinoise traditionnelle, où les frontières entre soi et la communauté sont beaucoup moins nettes. Pourtant, si Google vise à nous connecter tous, et si nous voulons que Google agisse de manière éthique en ce qui concerne nos préoccupations en matière de vie privée, alors il ne peut pas être vrai, en termes pratiques, que ces valeurs distinctes en matière de vie privée n'ont rien à voir les unes avec les autres […]. [Traduction libre] (Vallor 2016, 22-23)

En plus de prendre en compte les éléments qui forment le contexte (dont la culture), l'éthique de la vertu n'est pas une éthique procédurale. Être vertueux, on l'a vu au chapitre deux, ne signifie pas de suivre un code, des règles ou une procédure particulière. En fait, il ne sert de rien de savoir suivre une règle parfaitement, s'il manque à l'agent le jugement pratique pour évaluer chaque circonstance (Vallor 2016, 17). Ce qui importe bien davantage, c'est de travailler à devenir des « penseurs moraux efficaces » (Danaher et Vallor 2018, 30:00mn). Conséquemment,

> […] la personne vertueuse doit donner un sens moral à chaque situation concrète rencontrée, et y apporter une réponse appropriée. Une réponse morale réussie se distingue d'une réponse échouée ou inappropriée dans la pratique, et les raisons du succès de cette réponse peuvent toujours être articulées après coup. Mais la différence entre le succès moral et l'échec moral peut rarement, voire jamais, être déduite à l'avance des principes *a priori*. [Traduction libre] (Vallor 2016, 24-25)

L'éthique de la vertu présente donc ces avantages en comparaison aux autres approches monistes dont il a été question plus haut, quand vient le temps de la mise en application pour l'adoption de nouvelles technologies. Toutefois, comme éthique politique, elle me semble poser quelques difficultés. D'abord, l'éthique de la vertu, avec l'accent qu'elle place sur le développement de l'excellence du caractère de l'agent, est éminemment personnelle. Autrement dit, il est impossible qu'on puisse devenir vertueux par contrainte. Même si c'était possible, ce ne serait pas souhaitable que les décideurs politiques constituent une classe qui veille au perfectionnement moral de leurs citoyens. Il y a là une forme de contrôle qui déplaît à la sensibilité moderne. À moins d'enfreindre la liberté personnelle de chaque citoyen, cette approche serait difficilement viable. Malgré cela, il faut mentionner que Vallor met de l'avant la vertu de « civilité » (Ess 2020, 554, 566), qui consiste en

> […] une disposition sincère à bien vivre avec ses concitoyens dans une société de l'information en réseau mondial : à délibérer collectivement et avec sagesse sur des questions de politique et d'action politique locales, nationales et mondiales; à communiquer, à entretenir et à défendre nos conceptions distinctes du bien vivre; et à travailler en coopération pour les biens de la vie technosociale que nous cherchons et espérons partager avec les autres. C'est une disposition à « faire cause commune » avec tous ceux avec qui nos destins sont désormais liés sur le plan technosocial. [Traduction libre] (Vallor 2016, 141)

La vertu de civilité a évidemment une portée sociale : il ne s'agit pas d'une vertu « individuelle ». Elle se rapproche de ce qu'Aristote entendait par « l'amitié civique », d'expliquer Vallor (2016, 141-142). Elle toucherait au caractère moral public des individus, et non à leur moralité privée

(Vallor 2016, 143). On peut toujours objecter, dans une perspective pluraliste des valeurs, que le contenu de la moralité publique changera selon les convictions des personnes chargées de le définir, générant de ce fait des compréhensions parfois incompatibles de ce qu'elle signifie.

Cela dit, Mark Coeckelbergh souligne à son tour que l'éthique de la vertu doit être comprise comme traitant d'agents compris comme étant des êtres relationnels (2020b, 3). Plus encore, insiste-t-il, « […] nous avons besoin d'une version moins moderne-individualiste d'Aristote, [une version] qui place l'éthique de la vertu dans un contexte social » [Traduction libre] (2020b, 4).[25] Pour la penser, Coeckelbergh s'inspire de la notion de « pratique » chez MacIntyre. Ainsi,

> […] la vertu et le vice deviennent […] « socialisés » en ce sens qu'ils ne sont plus seulement une question de caractère individuel, mais une caractéristique de toute une pratique et de l'histoire de cette pratique. La vertu peut donc être basée sur l'agent, mais elle est toujours liée et intégrée dans ce contexte social pratique plus large. [Traduction libre] (Coeckelbergh 2020b, 5)

Toutefois, force est d'admettre que MacIntyre conceptualise l'éthique de vertu avec une compréhension téléologique de la nature tout aristotélicienne (une sorte de « biologie métaphysique » [Taylor 1994b, 17; MacIntyre 1984, 58, 158, 162-163, 196-197]), qui entre justement en conflit avec certains aspects de la liberté moderne. Face à ce dilemme, Charles Ess propose une notion qui combine l'aspect relationnel de l'agent éthique, qui permet de placer l'éthique de la vertu dans un contexte sociétal, avec la notion d'autonomie que l'on retrouve chez Kant (2019, 73). Dans cette compréhension des choses, « […] l'autonomie implique à la fois l'indépendance et les relations avec les autres. [Elle] exige une forme irréductible de dialogue, de réflexion et de réceptivité à l'égard d'autrui » [Traduction libre] (Ess 2019, 78).

En outre, comme l'éthique de la vertu plonge ses racines dans la pensée aristotélicienne, il ne faut pas oublier que dans cette dernière, on trouve une adéquation complète entre l'éthique personnelle et l'éthique politique. La liberté personnelle par rapport à l'État est une notion inexistante dans la pensée d'Aristote. Ce dernier est on ne peut plus clair : « […] il y a identité entre le bien de l'individu et le bien de la cité […] » (Aristote 2014, Livre I, 1, 1094b10). Cependant, avec la conception moderne de la politique est apparue la notion de liberté individuelle

[25] J'exposerai, au chapitre 6, en quoi la pensée politique d'Aristote est justement non moderne, par définition, et donc en tension avec la notion (moderne) de liberté individuelle.

dont on ne voudrait pas se départir aujourd'hui. Nous sommes très attachés à cette idée et à ses nombreux bénéfices, et

> notre réticence à donner à la politique un *telos* ou une fin déterminée traduit notre souci de respecter la liberté individuelle. Nous considérons la politique comme une procédure permettant aux personnes de choisir elles-mêmes leurs propres fins. Aristote ne voit pas les choses de cette manière. […] Pour Aristote, la politique a une visée plus haute. Elle a pour objet de nous enseigner comment vivre une vie bonne. La finalité de la politique ne consiste en rien de moins que de permettre aux gens de développer leurs capacités et leurs vertus proprement humaines — de délibérer à propos du bien commun, de former leur jugement pratique, de prendre part au gouvernement autonome, de se soucier du sort de la communauté considérée comme un tout. (Sandel 2016, 283-284)

Cela dit, je ne pense pas que l'éthique de la vertu soit à rejeter entièrement en ce qui a trait à la sphère politique. Certains de ses aspects peuvent être repris par les décideurs politiques dans leurs délibérations sur l'intelligence artificielle. Par exemple, sans forcer les citoyens à la pratique des vertus, un gouvernement peut veiller à ce que des espaces et des conditions soient aménagés qui eux, sont propices à ce que les personnes acquièrent des vertus, ou développent des « capabilités » (Nussbaum 1990).

C'est un peu dans ce sens qu'abonde Coeckelbergh lorsqu'il pense la vertu en lien avec les travaux de Pierre Bourdieu et de son concept de « habitus », en plus de la notion de « pratique » d'Alasdair MacIntyre. Son but est de créer un point de rencontre entre la philosophie de la technologie phénoménologique, qui traite abondamment de la « corporéité » (*embodiment*), avec des figures telles que Edmund Husserl, Martin Heidegger et Maurice Merleau-Ponty (Coeckelbergh 2020b, 5). Coeckelbergh explique qu'

> alors que Vallor — dans la lignée d'Aristote et d'une grande partie de la pensée moderne occidentale — met l'accent sur les raisons, les motivations et les « états » de l'esprit, Bourdieu souligne la dimension du savoir-faire implicite et du comportement sans but. Nous sommes régis sans obéissance aux règles. [Traduction libre] (2020b, 6)

Le caractère social de l'individu, l'acquisition des habitudes et l'importance du corps et de la temporalité marquent le caractère particulier de l'éthique de la vertu que propose Coeckelbergh. Dans cette optique, il importe aux agents d'être vertueux et d'éviter, par exemple, de maltraiter un robot. Ce n'est pas pour le robot lui-même, mais pour se préserver du vice, qui constitue « […] un

échec moral individuel […], mais [qui] est en même temps un échec de l'environnement social à organiser l'exercice de la vertu » [Traduction libre] (2020b, 7).

Néanmoins, on pourrait objecter qu'avec l'arrivée de l'intelligence artificielle et des possibilités qu'elle offre, c'est davantage l'efficacité que la vertu que les gouvernements semblent souhaiter maximiser. À titre d'exemple, la ville de Montréal mise sur les technologies employant l'IA qui pourraient générer des améliorations comme « améliorer le temps de réponse des pompiers [,] [a] dapter automatiquement les feux de circulation en fonction de la congestion [,] [et a] méliorer l'efficacité du colmatage des nids-de-poule » (Normandin 2019, s.p.). Le gouvernement fédéral canadien, au début de l'année 2019, était déjà en train de tester l'usage de l'intelligence artificielle, chez Emploi et Développement social Canada. L'objectif était et se veut toujours qu'éventuellement, des *bots* puissent offrir des services que des humains offrent actuellement, notamment par l'entremise de services de messagerie instantanée en ligne (The Canadian Press 2019, s.p.). Ces technologies sont sans conteste des améliorations d'efficacité, mais il est plutôt ardu de se représenter un gouvernement qui placerait la pratique de la vertu par ses citoyens en priorité par rapport au développement de l'IA, en ce qui a trait à ses investissements en technologie.

Plus encore, les nouveaux systèmes employant l'intelligence artificielle pourraient même, indirectement, rendre la pratique de la vertu moins impérative : par exemple, moyennant un algorithme qui arriverait à déceler le vrai du faux dans les nouvelles (Business Wire News 2019, s.p.). Dans un tel cas, il devient moins pressant de s'attaquer au problème du potentiel vice derrière la dissémination de fausses nouvelles. Toutefois, il est vrai que le gouvernement encourage déjà une certaine forme de « bonne vie », même sans souscrire explicitement à une éthique de la vertu eudaimoniste. Force est d'admettre, soutient Nussbaum, que

> les gouvernements ne se tiennent pas […] complètement à l'écart de la question de choisir de soutenir certaines fonctions humaines plutôt que d'autres. Aucun État moderne ne se contente de mettre des revenus et des richesses dans les poches de ses citoyens; au contraire, les programmes sont conçus pour soutenir certains domaines de la vie — santé, éducation, défense, etc. […] Même pour répondre à la question « Quelles sont les ressources utiles et utilisables que nous avons sous la main », il faut une conception implicite du bien et du bon fonctionnement humain. […] En bref, pour répondre à toutes les questions politiques intéressantes et actuelles sur les ressources et leur affectation par l'entremise de programmes et d'institutions, nous devons prendre position, et prendre position tout le temps, sur la question aristotélicienne : « Quelles

> sont les fonctions humaines importantes? Qu'est-ce qu'une bonne vie exige? [Traduction libre] (Nussbaum 1990, 212)

Au fond, il s'agit de la question de la cible (le bien commun) vers laquelle les politiques sont orientées. Nussbaum soutient — comme Taylor le faisait, on l'a vu, dans sa critique de la supposée « neutralité » d'une éthique procédurale — qu'on ne peut échapper à la question du bien comme cible. Elle est soit implicite, soit explicite, mais elle est présente. La question qu'il faut se poser est plutôt de savoir si cette cible est donnée par la nature, ou déterminée par la raison pratique, comme les moyens pour l'atteindre. C'est ce qui marque la différence entre l'éthique de la vertu et l'éthique pluraliste, comme je l'ai explicité dans la section précédente.

Un tenant de l'éthique de la vertu *néo*-aristotélicien pourrait soutenir que l'État pourrait participer à la croissance des vertus intellectuelles des citoyens, puisqu'elles sont davantage le résultat de l'enseignement reçu que de pratiques et d'habitudes. En revanche, un risque demeure : celui de l'empiétement de l'État sur la vie morale des citoyens, peu importe s'il s'agit de la promotion de vertus intellectuelles ou morales. Un risque qui se profile est celui de l'ingénierie sociale. Le régime démocratique qui caractérise le Canada et plusieurs autres États se préoccupant de l'éthique de l'intelligence artificielle n'est absolument pas le même que celui de la Chine. D'aucuns s'inquiètent du fait que l'administration de cette puissance exerce un type de contrôle social sur ses citoyens pour leur enseigner une forme de « vertu », par l'entremise de son système de crédit social (Wright 2019, s.p.).

J'ai exposé qu'il est somme toute ardu de penser une éthique de la vertu pour les décideurs politiques, hors du langage de l'obligation légale ou morale, pour entourer le développement et l'utilisation de l'intelligence artificielle. À savoir si certains éléments contenus dans l'éthique de la vertu seraient désirables pour un « guide conversationnel » pouvant servir d'outil aux décideurs politiques, il me semble que la réponse est affirmative. Ces éléments comprendraient la prudence comme vertu politique, ainsi que la culture de la sensibilité au contexte — deux notions avec lesquelles des philosophes pluralistes pourraient être en accord. C'est ce qui constituera le cœur de ma proposition éthique alternative, que j'exposerai en détail au chapitre 6.

2. L'éthique de la vertu dans les démarches recensées

Des démarches analysées dans l'échantillon retenu, aucune ne consiste en un exercice d'éthique de la vertu. Ni les entreprises privées, ni les groupes combinant plusieurs partenaires, ni les organisations ou instances internationales ne présentent de directives entièrement inspirées de l'éthique de la vertu. On l'a vu, cette dernière n'est pas une éthique du faire ou de la règle, mais bien une éthique de l'être. En ce sens, Hagendorff soutient qu'il s'agit d'une éthique qui concerne les « technologistes », les développeurs, mais non le produit technologique en tant que tel (Hagendorff 2019, 9). Plus encore, selon lui, c'est l'éthique déontologique qui domine en éthique de l'IA actuellement.[26] Malgré tout, il exprime sa sympathie à l'éthique de la vertu en disant que

> […] l'approche prédominante de l'éthique déontologique de l'IA devrait être complétée par une approche orientée vers l'éthique de la vertu visant les valeurs et les dispositions du caractère. […] Lorsqu'elle suit la voie de l'éthique de la vertu, l'éthique en tant que discipline scientifique doit s'abstenir de vouloir limiter, contrôler ou diriger. Très souvent, l'éthique ou les directives éthiques sont perçues comme quelque chose dont le but est d'arrêter ou d'interdire une activité, d'entraver des recherches et des efforts économiques de valeur. Je veux renoncer à cette notion négative de l'éthique. L'éthique ne devrait pas avoir pour objectif d'étouffer l'activité, mais de faire exactement le contraire, c'est-à-dire élargir le champ d'action, de découvrir les angles morts, de promouvoir l'autonomie et la liberté, et de favoriser la responsabilité personnelle. [Traduction libre] (Hagendorff 2019, 9)

L'éthique de la vertu présente ainsi des forces indéniables, notamment son caractère positif et éloigné des interdits moraux. Cela dit, comme mentionné plus haut, penser un cadre d'éthique de la vertu pour les décideurs politiques, concernant l'IA, n'est pas si simple.

En dehors de l'échantillon à l'étude pour ce portrait, un document a été développé pour les ingénieurs en s'inspirant de l'éthique de la vertu. Shannon Vallor a mis au point une « boîte à outils éthique » (mentionnée au premier chapitre) qui présente « […] des moyens concrets de mettre en œuvre une réflexion, une délibération et un jugement éthiques dans les processus d'ingénierie et de conception de l'industrie technologique » [Traduction libre] (Vallor 2018, 2). Le document regroupant les sept outils fonctionne surtout par des questions précises, associées à chaque outil et adressées aux ingénieurs. L'emploi des outils est toujours illustré par un exemple concret qui sert

[26] Il faut préciser que sa catégorisation ne présente que des approches monistes.

de modèle. Les sept outils sont 1) « le balayage des risques éthiques », 2) « pré et post-mortem éthiques », 3) « l'élargissement du cercle éthique », 4) « l'analyse basée sur des cas », 5) « se souvenir des avantages éthiques du travail créatif », 6) « penser aux "gens terribles" » et 7) « boucler la boucle : la rétroaction éthique et l'itération » [Traduction libre] (Vallor 2018, 3). La terminologie qu'emploie Vallor est révélatrice. Avec le terme « outils », on pourrait se demander si Vallor confond l'art (*techne*) et la prudence (*phronesis*). Aristote entend par l'art la « production », et par la prudence « l'action ». Il soutient que « […] la disposition à agir accompagnée de règle est différente de la disposition à produire accompagnée de règle » (Aristote 2014, Livre VI, 4, 1140a1-5). Parler de « boîte à outils » pour mobiliser une vertu de la raison pratique peut mener à ce brouillement entre la production et l'action.

Plus largement, l'éthique de la vertu face à l'intelligence artificielle semble demeurer cantonnée dans les cercles académiques — on le voit avec la proposition de Martin Gibert (2020), visant à « faire la morale aux robots » de façon à ce qu'ils se comportent de manière vertueuse. Cela étant dit, des mentions à la vertu ainsi qu'à l'*eudaimonia* sont faites dans certaines démarches qui seront analysées en détail quand il sera question des directives exhibant une certaine tension métaéthique, dans le chapitre suivant. Ces démarches, sans être entièrement monistes ni entièrement pluralistes, présentent — paradoxalement — des aspects des deux positions métaéthiques et sont parfois le résultat de combinaisons de notions éthiques un peu disparates. Il est intéressant de relever que ce sont les démarches éthiques en tension (au plan métaéthique) qui apparaissent les plus populaires actuellement. Avant de les aborder, il convient de se pencher sur les autres approches monistes à l'étude dans cette thèse et leur usage dans les directives : l'utilitarisme et l'éthique déontologique.

b) L'utilitarisme

1. Esquisse d'une éthique utilitariste face à l'intelligence artificielle

J'aimerais esquisser sommairement le portrait des caractéristiques d'une éthique pour l'IA qui serait entièrement utilitariste. Tout d'abord, elle pourrait prendre la forme d'une théorie éthique systématisée et unifiée. Son caractère théorique rappellerait les théories des sciences de la nature,

avec une attention particulière au mesurable, au quantifiable et à la précision presque mathématique des principes et des conséquences. Ce type de théorie serait applicable soit aux machines, soit aux programmeurs et concepteurs qui les créent, ou encore aux les décideurs politiques qui encadrent leur production et leur utilisation (voire les conditions de leur conception).

Si l'on prend l'exemple d'un algorithme qui puisse réglementer la circulation des automobiles dans une ville donnée, il semble qu'une telle technologie que permet l'intelligence artificielle soit d'intérêt pour les décideurs politiques de même que pour la population, dont la condition des infrastructures routières revêt une certaine importance. Une étude menée à l'Université de Californie à Berkeley suggère que l'intelligence artificielle, dans le contrôle des feux de circulation ainsi que dans sa fluidité, participerait à éliminer les situations d'embouteillage (Manuguerra-Gagné 2018, s.p.). L'IA pourrait contribuer à ce que les usagers de la route maintiennent la distance de sécurité appropriée avec les autres véhicules. Le fait de respecter cet écart permet d'éviter d'accélérer et de décélérer à répétition, et conséquemment d'améliorer le débit de la circulation. La course de chaque voiture n'est donc pas arrêtée de manière répétitive. Il ne suffirait pas d'ajouter des voies, car l'étude en question suggère que même en enlevant des voies, la fluidité de la circulation automobile peut s'améliorer en misant sur le respect des marges de sécurité entre les véhicules (Manuguerra-Gagné 2018, s.p.).

Une approche utilitariste voulant régulariser l'utilisation de cet algorithme d'intelligence artificielle pourrait exhiber les caractéristiques suivantes. Tout d'abord, la technologie serait évaluée de manière à déterminer si elle produit réellement une maximisation de l'utilité (à comprendre comme bonheur ou plaisir) chez les usagers routiers. Un autre angle de l'étude serait l'envers de la maximisation du plaisir, soit la minimisation de la souffrance. On peut facilement s'imaginer qu'une telle technologie — dans l'éventualité où on ne connaît pas encore ses « effets secondaires », serait conforme à une éthique utilitariste. Le système d'IA ne revêtirait plus un caractère éthique acceptable dans l'éventualité où ce ne serait plus le cas, suivant la logique de Bentham et Mill. Au plan politique, donc, ce qui rend une technologie IA recevable, c'est sa maximisation du plaisir de chaque usager de la route, ce qui, mathématiquement parlant, assure une croissance quantitative du plaisir généralisé d'une société donnée.

Au risque de verser dans la caricature, il faudrait toutefois spécifier qu'une telle technologie, pour être encadrée éthiquement de manière entièrement utilitariste, ferait de cette maximisation de l'utilité la valeur à laquelle toutes les autres valeurs seraient commensurables. Quelques difficultés concrètes commencent à se profiler. Si l'on pense à la valeur de la sécurité, il faut demander si cette dernière sera absorbée et « dissoute », dans le raisonnement éthique, par l'utilité, ou si l'utilité, au contraire, ne sera complète que si elle englobe la sécurité des usagers. Il me semble que la seconde option est plus fidèle à l'esprit de l'utilitarisme, mais aussi à la logique selon laquelle il faut être en vie pour pouvoir maximiser son plaisir. Les choses se corsent quand vient le temps de penser à la façon de mesurer cette commensurabilité des valeurs, et cette maximisation de l'utilité. Les éthiciens utilitaristes qui auraient à cœur un tel projet de fluidité de la circulation automobile, par exemple, devraient élaborer une certaine procédure d'arbitrage entre les valeurs qui pourraient entrer en conflit. Le seul fait de suivre cette procédure garantirait aux concepteurs d'éviter un conflit de valeurs pouvant se solder par une tragédie.

On peut tout de suite voir un potentiel conflit de valeurs dans une telle technologie. En effet, pour maintenir une distance de sécurité entre les véhicules, et ainsi éviter les freinages et accélérations à répétition, une certaine soumission à des limites de vitesse est requise. Cette dernière, par définition, empiète sur une compréhension négative de la liberté individuelle de chaque citoyen. Cela revient à dire qu'un certain sacrifice de la liberté individuelle de chacun est nécessaire pour améliorer le bien-être généralisé de la population empruntant les voies routières d'une ville donnée. De tels compromis existent déjà dans les sociétés contemporaines et semblent fonctionner (par exemple, la loi entourant le port de la ceinture de sécurité) — jusqu'à ce que des utilisateurs décident d'enfreindre les restrictions mises dans le but de maximiser l'utilité de tout un chacun. Une solution, déjà existante, serait d'adopter une approche punitive à l'endroit de ceux qui enfreignent les limites de vitesse et du coup, gâchent l'efficacité de la technologie visant à améliorer le flux de la circulation.

D'autres exemples pourraient être donnés, et permettent de soulever la question de la définition et la mesure de l'utilité, question qui fait actuellement problème chez les penseurs utilitaristes. Cette notion de l'utilité pourrait prendre différents visages dans les réflexions entourant les innovations technologiques. Le gouvernement canadien, comme d'autres instances

industrielles ailleurs dans le monde, désire incorporer des algorithmes d'apprentissage profond pour certains niveaux de prise de décision dans la fonction publique (Gouvernement du Canada, Conseil du Trésor 2019). Si la motivation derrière ce type d'innovation repose sur la maximisation d'une valeur jugée centrale, soit l'efficacité, il faut se demander si, dans une éthique utilitariste pour l'intelligence artificielle, utilité et efficacité pourraient devenir synonymes. Si c'était le cas, il importerait également de se demander à quel prix. L'efficacité peut effectivement créer une augmentation de plaisir, ou tout au moins une diminution de la frustration qui peut généralement être ressentie par un agent devant une façon de fonctionner particulièrement inefficace. L'efficacité peut être perçue comme une volonté de « réduire la friction » qui ralentit les procédures d'exécution (Selinger 2019, s.p.). Comme n'importe quel « contenu » que l'on pourrait attribuer à l'utilité, dans une théorie éthique conséquentialiste, il est crucial de se pencher sur les implications de cette valeur. À cet égard,

> […] l'efficacité n'est pas toujours neutre sur le plan de la valeur. Placer l'efficacité au-dessus d'autres valeurs peut être une erreur — un manque de jugement éthique, politique, personnel ou professionnel. Certaines interactions humaines ou civiques se développent lorsqu'elles sont délibérées et s'érodent lorsqu'elles sont accélérées. [Traduction libre] (Selinger 2019, s.p.)

En définitive, une éthique entièrement utilitariste, pensée pour l'intelligence artificielle, paraît intuitivement intéressante. La maximisation de l'utilité, des bénéfices, et la minimisation des dommages ou de la souffrance sont des principes qui semblent évidents. C'est vers le contenu de l'« utilité » ou des « bénéfices », ainsi que sur les répercussions de la commensurabilité de toutes les valeurs à cette dernière qu'il importe de se poser des questions. Le caractère systématique, mesurable et, en dernière analyse, profondément moniste d'une telle approche pourrait impliquer d'importantes pertes qui seraient difficilement quantifiables et qui, par conséquent, tomberaient dans l'angle mort d'une éthique utilitariste.

2. L'utilitarisme dans les démarches recensées

Au premier regard, une grande majorité de réflexions développées pour l'éthique de l'intelligence artificielle peuvent sembler être conséquentialistes, ou encore utilitaristes. Si l'on ne jette qu'un regard superficiel aux énoncés d'intention de plusieurs des études et chartes

promulguées dans les dernières années, on pourrait penser qu'avec l'omniprésence de l'idée d'une IA « bénéfique à l'humanité », le cas sera réglé, et l'on pourrait plier bagage en pensant avoir mis le doigt sur *la* tradition éthique dominante actuellement. En effet, en 2017, le Future of Life Institute (FLI), une organisation sans but lucratif, située à Cambridge, organisait une conférence intitulée « *Beneficial AI* » à Asilomar, en Californie, où les vingt-trois principes d'Asilomar ont été adoptés, constituant ainsi une des premières chartes de principes entourant le développement de l'IA (Assemblée nationale et Sénat de France 2017, 177-178). La lettre ouverte que publie le FLI, deux auparavant, dans le but d'attirer l'attention des chercheurs sur les implications de l'IA, va dans le même sens : « en raison du grand potentiel de l'IA, il est important de rechercher comment en récolter les bénéfices tout en évitant les pièges potentiels » [Traduction libre] (Future of Life Institute 2015, §2).

Dans le milieu industriel, DeepMind assure se préoccuper de ce que l'IA soit bénéfique à la société (DeepMind 2017, §1). Le laboratoire de recherche OpenAI désire que l'intelligence artificielle générale (AGI) soit bénéfique à l'humanité (OpenAI 2018a, §1), tandis que la compagnie ElementAI a été fondée dans la foulée des investissements de Microsoft pour une intelligence artificielle qui soit bénéfique (Assemblée nationale et Sénat de France 2017, 180). Puis, dans leur seconde étude sur les standards techniques à adopter pour l'IA, l'Institut des ingénieurs électriques et électroniques (IEEE) élargit sa mission pour faire en sorte que les systèmes autonomes et employant l'IA soient « […] mis de l'avant au profit de l'humanité » [Traduction libre] (Institute of Electrical and Electronics Engineers, Incorporated (IEEE) 2017, 3). La récurrence du principe de non-malfaisance, dans les directives éthiques touchant à l'IA (Jobin, Ienca et Vayena 2019, 15), peut elle-même être tributaire d'un utilitarisme négatif, ou encore des influences du principlisme (Beauchamp et Childress 2001, x). Pourtant, une lecture approfondie des documents montre que l'état de la question est plus corsé qu'il n'y paraît.

Ces énoncés d'intention ont un certain caractère de généralité. Autrement dit, même si ce type d'objectif global est mentionné, cela ne fait pas du document en question un exemple de démarche « utilitariste » à proprement parler. Certes, la valeur mise de l'avant, dans ces énoncés d'intention, est prisée par les utilitaristes. Cependant, il faut reconnaître qu'il est du propre d'une réflexion éthique de s'orienter vers un bien. Dans le cas de développements technologiques, il

paraît intuitif de souhaiter que tous, dans une société donnée, en bénéficient de manière égale. Même si cette idée peut faire penser à la maximisation de l'utilité, dans une théorie utilitariste au sein de laquelle chacun compte également, il faut regarder un peu plus loin pour voir le statut de cette idée au sein de la directive. Souvent, elle constitue un principe parmi d'autres, parfois dans une charte de type pluraliste, qui contient aussi des énoncés pouvant se rattacher soit au déontologisme, soit au conséquentialisme, mais sans hiérarchisation entre eux.

Un autre élément récurrent dans les directives éthiques recensées est le principe de proportionnalité. De prime abord, il est difficile de déterminer, hors de tout doute, de quelle école éthique relève ce principe. D'une certaine façon, il pourrait être rattaché à toutes. On pourrait entendre la proportionnalité comme une forme de « juste milieu », et il relèverait alors de l'éthique de la vertu. On pourrait aussi le concevoir comme un « accommodement raisonnable », conclus dans un contexte où l'on sait pertinemment que les pertes sont inévitables, mais au sein duquel on a cherché à minimiser les dégâts : dans un tel contexte, il s'agirait d'une idée pluraliste. Le principe de proportionnalité pourrait aussi être entendu de manière déontologique, à savoir un impératif à suivre dans toutes les circonstances. Finalement, on pourrait le considérer comme étant utilitariste, relevant d'une sorte de fusion entre ses versions positive (maximiser l'utilité) et négative (minimiser la souffrance). Tout dépend, en définitive, du contexte plus large dans lequel le principe est invoqué. Si l'on y a recours comme à une manière d'atteindre une proportion de type quantitative ou mathématique, on se retrouve manifestement du côté utilitariste et on s'éloigne de la raison prudentielle caractérisant l'éthique de la vertu de même que plusieurs penseurs pluralistes.

Dans les exemples suivants, j'aurais tendance à associer le principe de la proportionnalité soit à l'éthique déontologique, soit à l'utilitarisme, mais certes pas à l'éthique de la vertu ni au pluralisme des valeurs. En 2018, Google met de l'avant ses sept principes éthiques entourant le développement de l'intelligence artificielle, dont l'un d'entre eux, nous l'avons vu, est que l'IA soit « socialement bénéfique » (Google 2018, s.p.)[27]. La compagnie désire s'assurer que les produits soient 1) socialement bénéfiques, ce qui signifie que les produits seront développés quand « […] les avantages globaux probables dépassent largement les risques et les inconvénients

[27] Il faut noter que la compagnie a annoncé que d'autres services en éthique de l'IA seraient développés dans le futur, touchant par exemple aux biais et discriminations raciales (Simonite 2020, §2).

prévisibles » [Traduction libre] (Google 2018, §5). Ce principe pourrait être compris comme évoquant le souhait général que les produits de la compagnie génèrent des retombées positives, ce qui n'est pas nécessairement utilitariste. Il pourrait aussi être un signe d'utilitarisme dans la directive, s'il était érigé comme *le principe* auquel les autres sont soumis. Autrement dit, il faudrait qu'il y ait une hiérarchie au moins implicite entre les principes avant de conclure au monisme. La suite de la déclaration pourrait clarifier la situation.

Les six principes suivants sont 2) « éviter de créer ou de renforcer des préjugés injustes », 3) « être construits et testés pour assurer la sécurité »; 4) « être responsables devant les gens »; 5) « intégrer les principes de conception du respect de la vie privée »; 6) « maintenir des normes élevées d'excellence scientifique » et 7) « être mis à disposition pour des utilisations conformes à ces principes », dans le but de « […] limiter les utilisations préjudiciables ou abusives » [Traduction libre] (Google 2018, §6-11). Un premier constat est qu'il n'y a pas de mention de conflit potentiel entre ces principes et valeurs. Par contre, l'on pourrait objecter que ces collisions sont implicites. Néanmoins, il n'est pas possible d'appuyer cette affirmation par quelque exemple que ce soit. Ce que l'on observe, c'est que le septième principe pourrait évoquer le monisme, au sens où une façon de le comprendre serait comme une « exigence d'harmonie » entre les principes dans la pratique. Parmi ces derniers, le premier peut avoir une résonnance utilitariste selon le contexte dans lequel il est placé, à savoir que les bénéfices doivent excéder les risques prévisibles. Une manière de s'en assurer est par une forme d'évaluation, voire de calcul, bien que cela ne soit pas précisé explicitement dans la déclaration.

De plus, tous les trois mois, Google met à jour des pratiques recommandées pour les SIA, de manière à ce que ces derniers soient bénéfiques. Si l'on devait s'attendre à ce que les principes publiés par le géant du Web entrent en conflit dans la dimension pratique (c'est-à-dire, dans leur cas, dans la fabrication de produits), Google n'aurait peut-être pas ajouté ce septième principe, qui enjoint au respect des sept principes, et non à ceux qu'il est possible d'honorer. Dans ce sens, on pourrait suggérer que ce septième élément est passible d'être vu en tant qu'élément « unificateur » de la déclaration. Et, bien sûr, la notion d'unité renvoie au monisme. Le septième principe peut donc impliquer une certaine subordination des autres principes, dans le but de limiter les retombées négatives. Cette clé d'interprétation des autres principes peut être vue comme la valeur centrale à

l'utilitarisme négatif, soit d'éviter le plus possible les préjudices. Cela étant dit, il m'apparaît que cette interprétation, quoique défendable, n'est pas exigée, et que le document pourrait être lu d'une autre façon. C'est la raison pour laquelle j'accole avec hésitation l'étiquette d'utilitariste à cette directive.

Un an plus tard, dans son étude sur le développement responsable de l'IA, le géant du Web publie une étude sur le développement responsable de l'IA, dans lequel il affirme qu'

> outre ces principes, nous nous engageons à ne pas donner suite à certaines applications. […] Plus généralement, pour toute application d'IA présentant un risque matériel de dommage, nous ne poursuivrons que *si nous estimons que les avantages l'emportent largement sur les risques* et nous intégrerons des contraintes de sécurité appropriées. [Traduction libre, je souligne] (Google 2019b, 4)

Cette prise de position fait suite au soulèvement d'employés de Google contre un contrat de drones militaires du Pentagone, que la compagnie a finalement décidé de ne pas renouveler en 2018 (Shane et Wakabayashi 2018). Dans une telle analyse, il me semble que le principe de proportionnalité a un caractère mesurable, voire quantitatif, qui peut rappeler une doctrine éthique utilitariste.

Une seconde illustration du principe de proportionnalité à saveur utilitariste provenant d'un document ne figurant pas à l'échantillon retenu pour l'analyse, peut contribuer à éclairer la catégorisation métaéthique du positionnement de Google. Le Groupe de Développement des Nations Unies (GDNU), dans la série de principes devant réglementer la collecte et l'usage des données massives (big data), mentionne explicitement le critère de proportionnalité. Le troisième principe qu'il met de l'avant, portant sur l'atténuation des risques, doit être interprété selon le critère de proportionnalité, peut-on lire dans le document :

> l'usage des données devrait être basé sur le *principe de proportionnalité*. [Plus] particulièrement, *n'importe quel risque ou dommage potentiel ne devrait pas excéder les impacts positifs (les bénéfices) de l'utilisation des données*. En outre, il est recommandé, autant que possible, d'analyser les effets des données sur les droits individuels conjointement les uns aux autres, *plutôt que d'opposer les droits entre eux*. [Traduction libre, je souligne] (United Nations Development Group [UNDG] 2017, 5)

Les rédacteurs du rapport préconisent de ce fait une approche éthique unifiée, voulant éviter les chocs de valeurs en soupesant des droits les uns contre les autres. On se retrouve manifestement

ici dans une démarche moniste, à l'opposé du pluralisme sur le continuum métaéthique. Puis, sous le huitième principe concernant « les données ouvertes [*open data*], la transparence et la responsabilité », il est stipulé que de « faire preuve de transparence en ce qui a trait à l'utilisation des données […] est généralement encouragé lorsque les *bénéfices de la transparence sont plus importants* que les risques et les dommages possibles » [Traduction libre, je souligne] (United Nations Development Group [UNDG] 2017, 7). L'approche que met de l'avant le GDNU par rapport au principe de proportionnalité est de toute évidence moniste et semble pouvoir être rattachée à une démarche utilitariste, qui pourrait mesurer ou « apprécier » les bénéfices de manière quantifiable par rapport aux dommages. Il s'agit certes d'une interprétation puisque ni le GDNU, ni Google, n'affirment explicitement se cantonner du côté de l'utilitarisme en éthique. Néanmoins, leur conception et usage du principe de proportionnalité peut donner à penser qu'ils y souscrivent implicitement. Une dernière approche moniste doit être explorée avant de passer aux directives éthiques pluralistes et à celles présentant des tensions : il s'agit de l'éthique déontologique.

c) L'éthique déontologique

1. Une précision sur le « dilemme du tramway »

Une mise en situation est constamment mise de l'avant lorsqu'il est question d'enjeux éthiques entourant l'usage de l'intelligence artificielle. Il s'agit du dilemme du tramway (« the trolley problem » en anglais). Présenté sous une forme ou une autre, ce dilemme a originalement été posé par l'éthicienne de la vertu Philippa Foot (1978). Il faut spécifier d'emblée que le « dilemme du tramway » n'est pas seulement invoqué dans les démarches déontologiques, bien au contraire. Il est utilisé pour des raisonnements conséquentialistes également, le plus souvent dans le cas concret des voitures autonomes (par exemple De Luca-Baratta 2019; Della Foresta 2020). Une raison pour laquelle je l'invoque ici est qu'un raisonnement déontologique peut aussi être proposé dans le but de résoudre ce dilemme. Une autre est que ce genre de dilemme ne concerne pas seulement les concepteurs de systèmes autonomes, mais encore les décideurs politiques et la société civile. En effet, à l'issue d'un sondage effectué auprès de concepteurs auxquels on présentait ce dilemme, 77 % d'entre eux étaient d'avis qu'il ne leur appartenait pas de décider des modalités de sa résolution. Au contraire, ils croyaient que c'était aux décideurs politiques, tout

comme aux usagers de la technologie, de se pencher sur la question et de trancher (Millar 2016, 792).

Foot a initialement illustré le dilemme du tramway au moyen de deux exemples. Dans le premier cas, une décision doit être prise par un juge entre tuer un seul homme innocent ou laisser mourir cinq personnes innocentes. Dans le deuxième cas, un chauffeur de tramway doit choisir s'il doit ou non dévier sa route, la conséquence étant soit de tuer un homme innocent sur les rails s'il ne dévie pas, soit de tuer cinq personnes innocentes s'il bifurque (Woollard et Howard-Snyder 2016, s.p.). La philosophe américaine Judith Jarvis Thomson soutient, à la différence de Foot, que dans un type de dilemme comme celui du tramway, il ne suffit pas de simplement stipuler que tuer est mal, tandis que de laisser mourir serait admissible (2003, 140-141, 145). Il faudrait plutôt réfléchir à la situation, et à ses différentes caractéristiques hypothétiques (que Thomson met tour à tour de l'avant, en ajoutant même la figure d'un passant qui aurait la possibilité de détourner le tramway). Dans son analyse, elle fait appel à la seconde formulation de l'impératif catégorique kantien, soit de traiter l'humanité comme une fin et non seulement un moyen vers autre chose (Thomson 2003, 145). Cela revient à faire appel à la notion de « droit », qui supplante celle d'utilité, pour aider à analyser moralement ce type de dilemme. Dans la perspective déontologique qui est celle de Thomson, l'utilité, même maximisée, n'est pas une raison suffisante pour enfreindre les droits de ceux qui les portent (2003, 147-148).

Il est toutefois frappant de noter que, dans son argumentaire, il n'y a pas de considérations sur la tragédie réelle de ces situations. En effet, c'est le propre de ces dilemmes de toujours impliquer une fin désastreuse, peu importe l'alternative choisie. Il n'est pas non plus question de honte, ce sentiment qui vient avec le fait d'avoir « les mains sales ». Plutôt, Thomson conclut en proposant simplement le principe suivant :

> […] si une personne est confrontée à un choix entre faire une chose, ici et maintenant, à cinq personnes, par laquelle elle les tuera, et faire une autre chose, ici et maintenant, à une seule personne, par laquelle elle n'en tuera qu'une seule, alors (toutes choses étant égales par ailleurs) elle devrait choisir la seconde alternative plutôt que la première. [Traduction libre] (Thomson 2003, 159)

Quelques années plus tard, Thomson changera d'avis et soutiendra qu'en fait, elle n'avait pas tenu compte de la possibilité que le passant puisse faire dévier le tramway sur lui-même, et que cette dernière alternative serait préférable (Thomson 2008 dans Woollard et Howard-Snyder 2016, s.p.).

Ce n'est pas sous influence de l'utilitarisme — que réfutent les kantiens — que Thomson arrive à sa dernière recommandation. L'idée n'est pas de tuer le moins de personnes possible en raison d'un seul calcul de maximisation quantifiable du bien-être ou, dans ce cas, de réduction quantifiable de souffrance. Il s'agit d'un raisonnement par *principe moral*. Selon ce principe, le passant, qui dévierait le tramway sur une personne plutôt que cinq, s'apparenterait à une personne qui dépouillerait une autre de ses biens pour les offrir en charité (Woollard et Howard-Snyder 2016, s.p.). Plus encore, on pourrait y voir une certaine forme d'instrumentalisation de la personne seule, sur les rails, qui serait sacrifiée pour éviter de la souffrance aux cinq autres, alors que l'impératif catégorique kantien stipule que toute personne doit être traitée comme une finalité, jamais comme seulement un moyen vers une fin, aussi bonne soit-elle.

La nouvelle technologie que représente la voiture autonome est le cas de figure par excellence pour le dilemme du tramway et les exemples pullulent dans la littérature. Un scénario fréquemment invoqué est celui de la voiture autonome qui, devant l'impossibilité d'éviter de heurter un ou des piétons, doit choisir entre cette éventualité, ou celle de la mort certaine de l'un ou plusieurs de ses passagers. Il s'agit finalement d'une forme de dilemme moral, au sujet duquel toutes les traditions éthiques ont quelque chose à dire. Le déontologisme n'étant pas exclu du lot, il existe en effet des approches inspirées de cette tradition pour des variantes du dilemme du tramway, comme celle de Thomson. Il s'agit évidemment de dilemmes moraux pour des machines : l'éthique concernerait dans ce cas directement des algorithmes, plutôt que des personnes. Néanmoins — on l'a vu —, il est difficile de séparer, dans l'absolu, l'éthique des machines de l'éthique humaine : en effet, il faut reconnaître que ce sont les humains qui inculqueront certains principes à l'algorithme. Plus encore, c'est souvent à partir des raisonnements et réflexes humains que l'on peut désirer programmer les machines. En réalité,

> indépendamment de la capacité d'action morale, certaines de ces machines seront confrontées à des situations dans lesquelles elles devront prendre une décision morale, ou du moins, ce que nous, les humains, comprendrions comme tel si c'était nous qui

> devions prendre la décision ou qui l'interprétions de notre point de vue […]. » [Traduction libre] (Nørskov et Rodogno s.d., 2)

Le « dilemme de la voiture autonome » est encore, à ma connaissance, irrésolu, au sens où il n'existe pas de consensus final dans la littérature. Cela n'est pas surprenant puisque, d'un point de vue déontologique kantien, il est difficile de trouver une « loi morale » dans un tel type de situation. En effet, celui ou celle qui trouverait un principe qui pourrait être universalisé, dans toutes les instances de ce type de dilemme, tiendrait probablement du génie — ou d'autre chose — puisqu'il est du ressort de ce genre d'impasse d'être pratiquement insoluble. C'est aussi l'avis de la Commission d'éthique nommée par le gouvernement allemand pour enquêter sur l'encadrement éthique de la voiture autonome. La commission dépose son rapport en 2017 — un rapport influencé par de nombreuses approches éthiques diverses —, mais présentant un argument kantien pour traiter de la prise de décision dans les situations de dilemme. Les membres de la commission relèvent

> […] la question élémentaire de savoir quel degré de liberté de choix nous sommes prêts ou autorisés à transférer aux programmeurs ou même aux systèmes d'autoapprentissage […] si, dans l'éthique kantienne, la liberté de l'individu de jouir du droit à l'autodétermination morale constitue la base d'une existence déterminée par la raison. [Traduction libre] (German Government, Federal Ministry of Transport and Digital Infrastructure 2017, 16)

Le problème que soulignent les membres de la commission est que, même si la machine ou l'algorithme prend une décision en conformité avec ce qu'on attendrait d'elle majoritairement, il demeure que la décision est imposée à l'agent humain de l'extérieur. En fin de compte, la décision

> […] ne saisit pas intuitivement une situation spécifique […], mais doit évaluer en termes abstraits/généraux. […] Conduit à sa conclusion logique, l'être humain, dans des situations existentielles de vie ou de mort, ne serait plus autonome, mais hétéronome. Cette conclusion est problématique à bien des égards. […] cela serait contraire au système de valeurs de l'humanisme, dans lequel l'individu est au centre de toutes les considérations. [Traduction libre] (German Government, Federal Ministry of Transport and Digital Infrastructure 2017, 16)

Il est intéressant de noter que l'éthique de Kant est à la fois adoptée et rejetée, d'une certaine façon, dans cette analyse. D'une part, le Kant de l'humanisme et de l'autonomie est mis de l'avant. D'autre part, celui de l'abstraction et de la généralité des principes est placé de côté. Ce raisonnement, par les décideurs politiques allemands, montre toute la complexité qu'une éthique déontologique pour

les machines peut revêtir. Elle permet de souligner que les décideurs politiques ont à voir avec l'éthique des machines, en plus de l'éthique qui touche plus directement aux humains eux-mêmes.

Force est de constater que les frontières sont poreuses entre les différents destinataires de l'éthique, quand il s'agit de guider les décisions de systèmes autonomes. Étant donné l'applicabilité de l'éthique déontologique à au moins trois groupes de destinataires de l'éthique, le propos qui suit sera divisé en autant de sous-parties. Dans mon esquisse d'une éthique déontologique face à l'IA, j'examinerai l'éthique pour les systèmes en IA, puis, pour les développeurs en IA et finalement, pour les décideurs politiques.

2. Esquisse d'une éthique déontologique face à l'IA

a. Une éthique déontologique pour les SIA

Il pourrait sembler farfelu de se tourner vers les œuvres de science-fiction pour penser une éthique machine en intelligence artificielle. Et pourtant, non seulement certains y voient une avenue pédagogique pour la sensibilisation à l'éthique (Burton, Goldsmith et Mattei 2015), mais également, d'aucuns ont emprunté la voie tracée par le romancier Isaac Asimov et ses trois lois de la robotique, tantôt pour les critiquer, tantôt pour les poser en modèles. Les lois d'Asimov sont un exemple parfait d'éthique déontologique, contenant des impératifs dont on ne peut déroger et qui sont par ailleurs imbriqués dans un système hiérarchique avec un ordre sériel, comme nous avons vu dans la théorie de la justice rawlsienne. Plus encore, ce « système juridique » robotique se veut un tout évitant les conflits entre ses éléments constituants.

On l'a vu, l'œuvre d'Asimov a inspiré à Oren Etzioni, le directeur général du Allen Institute for Artificial Intelligence, aux États-Unis, de développer ses propres lois pour encadrer le développement et l'usage de l'intelligence artificielle. Jugeant les lois d'Asimov « élégantes, mais ambiguës » (Etzioni 2017, s.p.), Etzioni suggère plutôt que l'IA soit « […] soumise à toute la gamme de lois qui s'appliquent à son opérateur humain » [Traduction libre] (Etzioni 2017, s.p.). La première loi que propose Etzioni est donc que le droit soit modifié, de façon à ce qu'on ne puisse simplement s'excuser du comportement d'une IA que l'on n'avait pas prévu ou que l'on n'aurait

pas suffisamment compris. La deuxième serait que tout système d'IA se dévoile tel quel, au lieu de leurrer l'humain quant à son « identité ». La troisième serait qu'aucune information confidentielle ne soit obtenue par un système d'IA sans le « consentement explicite de la source de cette information » [Traduction libre] (Etzioni 2017, s.p.). Il conclut en soutenant : « mes trois règles d'IA sont, je crois, solides, mais loin d'être complètes » [Traduction libre] (2017, s.p.), et qu'il désire lancer la discussion par leur entremise. Cette proposition de lois systématiquement imbriquées participe aussi d'une logique déontologique à l'éthique.

Le même chercheur a par ailleurs lancé une autre idée pour l'éthique de l'IA, idée qu'a reprise le Forum économique mondial dans son exploration d'avenues en la matière. Il s'agirait d'un système IA qui agirait comme gardien de l'éthique d'un autre système IA. Cette intelligence artificielle qui agirait comme « gardien éthique » contrôlerait la machine « […] en conformité avec un ensemble prédéterminé de lois et de règles éthiques qui seraient développées au niveau méta » [Traduction libre] (World Economic Forum [WEF] 2019a, 8; Etzioni et Etzioni 2016b, 29-31). Cette idée d'un système de règles éthiques qui en régit un autre est foncièrement déontologique, en ce qu'elle implique des impératifs à suivre, érigés en système unifié, avec une certaine procédure. Comme il s'agit d'une machine qui en réglemente une autre, il n'est pas question de déroger des lois qui sont programmées.

Un autre exemple de cette approche est évoqué dans la stratégie nationale américaine pour la recherche et le développement en intelligence artificielle. On y propose un système à deux niveaux, « […] qui séparerait l'IA opérationnelle d'un "agent système" responsable de l'évaluation éthique ou légale de toute action opérationnelle » [Traduction libre] (National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee 2016, 27). Encore une fois, une IA serait perçue comme une gardienne de la légalité ou de l'éthique des décisions prises.[28] L'éthicien Martin Gibert va jusqu'à suggérer qu'« il est même permis d'imaginer qu'un jour une IA générale nous aidera à identifier ce qui est bon, juste ou vertueux — bref, à faire de l'éthique » (Gibert 2019, s.p.). Cette idée est, à certains égards, étonnamment proche de la notion de « raison pratique pure » que met Kant de l'avant. En effet,

[28] Il ne faut pas entendre par là, cependant, qu'éthique et légalité sont des synonymes; ou encore que la loi, entendue comme le droit régissant une société, participe nécessairement d'une logique déontologique.

l'éthique semblerait relever, dans la situation hypothétique qu'avance Gibert, de l'exercice d'une forme de rationalité désengagée, en quête d'une certaine « neutralité ». Il m'apparaît que ce type de solution, pour une éthique machine, s'apparente à l'éthique déontologique kantienne.

Il est assez intuitif, quand on évoque la métaéthique déontologique, comme je l'ai fait plus haut, que cette tradition est idéale pour des machines ou, en l'occurrence, pour des systèmes d'intelligence artificielle. Il s'agit de procéder par ensemble de règles et par procédures de déduction, pour éviter des contradictions entre les impératifs. L'abstraction et la désincarnation de tels systèmes se marient parfaitement avec la raison pure pratique, et les systèmes d'arbitrage entre les règles. C'est quand ce type d'éthique est transposé au plan humain que certaines difficultés semblent se poser.

b. Une éthique déontologique pour les développeurs en IA

Si l'on voulait dresser un portrait de ce dont aurait l'air une éthique déontologique pour les humains qui conçoivent l'intelligence artificielle, il faudrait se tourner vers les codes d'éthique qui régissent les professions entourant la conception et le développement de l'IA ou d'autres technologies informatiques. Ce type de code d'éthique a été l'objet d'une étude de chercheurs de la Caroline du Nord qui voulaient en vérifier la portée normative concrète. C'est auprès de la plus importante société d'informatique du monde, l'Association for Computing Machinery (ACM), et à son code d'éthique que l'enquête a été menée (Shipman 2018, s.p.). Adopté en 1972, et mis à jour pour la première fois depuis 1992 en 2018, le code d'éthique de l'ACM affirme pouvoir « […] "servir de base à une prise de décision éthique" […] » [Traduction libre] (McNamara, Smith et Murphy-Hill 2018, 1).

Le code d'éthique de l'ACM a fait l'objet d'une étude quant à sa portée normative. Après une enquête-analyse auprès de soixante-trois étudiants de troisième année en génie logiciel et cent cinq professionnels dans le domaine, « aucune différence statistiquement significative dans les réponses n'a été constatée entre les personnes qui ont vu et celles qui n'ont pas vu le code de déontologie, que ce soit pour les étudiants ou pour les professionnels » [Traduction libre] (McNamara, Smith et Murphy-Hill 2018, 4). Les auteurs de cette étude mentionnent toutefois que,

quand les participants à l'étude avaient pris conscience d'une situation réelle, donnée en exemple, dans laquelle les conséquences d'une décision éthique étaient clairement mises de l'avant, leur décision éthique semblait en être affectée (McNamara, Smith et Murphy-Hill 2018, 4). Cette conclusion pourrait discréditer l'approche déontologique en éthique, tout en fournissant des munitions à la tradition de l'éthique de la vertu, pour qui la croissance en vertu passe par l'exposition et une certaine imitation de modèles. La chose est inconcevable pour une éthique déontologique qui se systématise le plus loin possible de toute contingence (Kant 1848a, 38, 40).

En conséquence, penser une éthique déontologique pour les développeurs en IA n'est pas une chose simple. L'enjeu n'est pas nouveau pour autant. De la même façon que plusieurs ordres professionnels ont leurs codes de déontologie, leurs conseils d'éthique (justement parfois appelés « conseils déontologiques ») et leurs éthiciens professionnels, il peut sembler logique de vouloir faire de même avec les développeurs de SIA. Une manière d'assurer l'éthique, du moins dans le domaine biomédical, est en adoptant l'approche du « principlisme », tel que systématisé par Tom L. Beauchamp et James F. Childress. Les quatre principes qu'ils mettent de l'avant sont 1) le respect de l'autonomie, 2) la non-malfaisance, 3) la bienfaisance et 4) la justice (2001, x). Le principlisme a été développé en réaction à la pratique des décisions éthiques dans le domaine médical, lesquelles « […] ont révélé que la prise de décision éthique sur le terrain implique souvent un compromis ou une "pondération" des intérêts pour décider de la ligne de conduite éthiquement appropriée » [Traduction libre] (Mittelstadt 2019, 2). En développant quatre principes, l'éthique médicale est ainsi formalisée. Ces principes peuvent par ailleurs être mis en balance selon le cas (Mittelstadt 2019, 2).

De prime abord, on pourrait penser que le principlisme emprunte au pluralisme des valeurs, puisqu'il est question de soupeser des principes pour la pratique. En réalité, il s'agit d'une approche éthique moniste qui me semble très près du déontologisme. Leur méthode de mise en balance des principes, plutôt que de faire appel au pluralisme des valeurs, renvoie à un équilibre réflectif ou réfléchi, tel que systématisé par John Rawls, ou encore Norman Daniels en éthique biomédicale (Beauchamp et Rauprich 2016, 2287). L'équilibre réflexif est ici distinct de l'exercice de mise en balance des valeurs, relevant du pluralisme et exposé au chapitre précédent. Il renvoie plutôt à un mouvement de va-et-vient entre la théorie et la pratique jusqu'à l'atteinte d'une approche unifiée.

Manifestement, cette approche n'implique pas le sentiment de culpabilité et les « mains sales » associées au pluralisme des valeurs. Ainsi, à l'aide des mécanismes de spécification et d'équilibrage spécifiquement monistes, Beauchamp et Childress ont développé un modèle « algorithmique » de résolution des conflits potentiels entre les principes (Beauchamp et Childress 2011, 16, 18 cités dans Thornton 2006, § 11, 16, 9). Plus encore, les tenants du principlisme affirment que

> contrairement à Rawls, qui ne se réfère pas au concept de la moralité commune, le principlisme considère ses jugements de base non seulement comme provisoirement acceptables comme points de départ, mais aussi comme une base morale solide pour la formation d'un équilibre réfléchi. [Traduction libre] (Beauchamp et Rauprich 2016, 2288)

C'est dans ce sens que ces principes se veulent universels et acceptables pour toutes les traditions morales (Beauchamp et Rauprich 2016, 2289-2290). Le monisme du principlisme est ici patent, tout comme ses parallèles avec l'éthique déontologique, ce qui en fait un candidat potentiel pour une éthique de type déontologique pour les « praticiens » et ingénieurs de l'IA.

Cela étant dit, des difficultés demeurent. Mittelstadt suggère que

> […] les initiatives en matière d'éthique ont à ce jour produit des principes et des déclarations de valeur vagues et de haut niveau, qui promettent d'orienter l'action. Cependant, dans la pratique, elles ne fournissent que peu de recommandations spécifiques et n'abordent pas les tensions normatives et politiques fondamentales, ancrées dans les concepts clés (par exemple, l'équité, la protection de la vie privée). [Traduction libre] (2019, 1)

Autrement dit, les directives éthiques en IA sont vagues et peinent à guider l'action — par exemple, celle des développeurs de SIA. Son commentaire rejoint la critique de Vallor entourant le respect de règles abstraites, qui sont, le plus souvent, éloignées du contexte spécifique à chaque enjeu. Malgré cela, le principlisme se revendique d'être une approche prônant la sensibilité au contexte (Beauchamp et Rauprich 2016, 2290). L'observation de Mittelstadt présente une certaine affinité avec le pluralisme, selon lequel même les valeurs morales entrent en tension les unes avec les autres, aux plans personnel aussi bien que politique. Les principes d'une approche comme le principlisme ne peuvent fonctionner de manière optimale pour les développeurs en IA, soutient Mittelstadt, puisque ces derniers ne partagent pas « des finalités communes ni des obligations

fiduciaires » et le champ de l'éthique de l'IA exhibe une carence « d'histoire professionnelle et de normes » [Traduction libre] (Mittelstadt 2019, 3-5). Si une éthique déontologique est plus facilement concevable pour des machines, elle est difficile à systématiser pour le champ professionnel du développement des SIA.

c. Une éthique déontologique pour les décideurs politiques

Une idée est déjà énoncée dans quelques études sur l'éthique de l'intelligence artificielle et des systèmes autonomes (par exemple European Commission 2018c, 5) : soit celle d'inscrire les obligations éthiques entourant ces technologies dans des normes juridiques déjà existantes, qu'elles soient nationales ou internationales. Une éthique déontologique des décideurs politiques pour le développement et l'encadrement de l'intelligence artificielle pourrait s'inscrire dans le cadre des chartes de droits de la personne qui ont une certaine force de loi. Par exemple, la Déclaration de Toronto sur l'intelligence artificielle se fonde sur le socle des droits de la personne. De l'avis de ses rédacteurs,

> ces lois et normes *universelles*, *contraignante*s et applicables, fournissent des moyens tangibles de protéger les individus contre la discrimination, de promouvoir l'inclusion, la diversité et l'équité, et de sauvegarder l'égalité. Les droits de l'homme sont « *universels*, *indivisibles*, *interdépendants* et *intimement liés*. » [Traduction libre, je souligne] (Bacciarelli et al. 2018, 1)

Le fait de proclamer les droits de la personne comme une doctrine universelle et contraignante, et plus encore de concevoir les droits comme systématiquement liés et interdépendants, relève d'une approche moniste de l'éthique (Blattberg 2016, 9), en l'occurrence l'éthique déontologique kantienne. Une atteinte aux droits de la personne peut mener à des procédures judiciaires devant des cours, qui sont en dernière instance des procédures d'arbitrage entre les valeurs. Le tout peut s'apparenter au fait de suivre la procédure énoncée dans un livre de règlement, qui fait sens lorsqu'il est pris en lui-même, dans son système fermé (Blattberg 2004, 22). Cela serait une façon pour les décideurs politiques de développer un cadre éthique déontologique face à l'IA.

3. L'éthique déontologique dans les démarches recensées

Les acteurs des entreprises privées recensées dans mon échantillon n'ont pas adopté de démarche entièrement déontologique dans leurs énoncés de principes pour l'éthique de l'IA. Cela dit, on le verra dans un chapitre ultérieur, l'éthique déontologique est fréquemment invoquée, même inconsciemment, aux côtés d'éléments pluralistes, faisant de ces documents des démarches que je désigne comme « affichant une forme de tension métaéthique ». On trouve néanmoins un exemple d'une approche majoritairement déontologique dans les documents émanant de la société civile et de multiples partenaires. On a vu que la Déclaration de Toronto s'ancre sur le socle des droits de la personne, compris dans une perspective plutôt kantienne que pluraliste. Il faut par ailleurs noter que certaines valeurs sont énoncées séparément, dans la Déclaration. Elles sont considérées comme étant essentielles au bon usage de l'intelligence artificielle et visent à « atténuer les dommages » que pourrait causer l'IA. Il s'agit de la transparence et la responsabilité (Bacciarelli 2018, 1). L'accent placé sur l'atténuation des dommages et la mention de deux valeurs, prises isolément, pourraient mettre en doute ma catégorisation de la Déclaration de Toronto du côté du déontologisme. Or, à défaut d'avoir de signes plus explicites d'autres appartenances éthiques — ces mentions n'étant pas entièrement concluantes — je considère cette Déclaration comme un document majoritairement inspiré d'une éthique déontologique politique.

Si l'on se tourne vers les directives publiées par des organisations de gouvernance internationale, on peut trouver un exemple d'éthique déontologique dans une étude de l'UNESCO sur l'éthique de la robotique. Il est intéressant de noter que c'est un rapport qui émane d'une instance à caractère politique, mais s'adresse aux concepteurs et utilisateurs des technologies de la robotique. Les rédacteurs suggèrent qu'

> un cadre de valeurs et de principes éthiques peut être utile pour établir des réglementations à tous les niveaux — conception, fabrication et utilisation — et de manière cohérente, des codes de conduite des ingénieurs aux lois nationales et conventions internationales. *Le principe de la responsabilité humaine est le fil conducteur qui unit les différentes valeurs énoncées dans le rapport*. Ces principes et valeurs éthiques pertinents comprennent : (i) dignité humaine; (ii) valeur de l'autonomie; (iii) valeur de la vie privée; (iv) principe de « ne pas nuire » [do not harm]; (v) principe de responsabilité; (vi) valeur de bienfaisance et (v) valeur de justice. [Traduction libre] (Commission mondiale d'éthique des connaissances scientifiques et des technologies 2017, 8[)]

Le fait que le principe de responsabilité ait un rôle unificateur parmi les autres principes donne à penser qu'il aurait un statut en quelque sorte supérieur aux autres valeurs, ce qui laisse deviner une hiérarchie implicite entre les principes. Peut-être pourrait-on argumenter que les principes de ne pas nuire et de bienfaisance ont une connotation utilitariste. Cependant, à l'intérieur de ce système où un principe unifie tous les autres, et que ce principe a une résonnance plus kantienne (« responsabilité » plutôt qu'« utilité »), même s'il est difficile de classer hors de tout doute une telle approche, elle m'apparaît principalement déontologique sous ce regard. Par ailleurs, les principes de l'UNESCO ne sont pas sans rappeler ceux du principlisme. Une étude de la Fondation pour la robotique responsable le cite d'ailleurs comme un exemple d'éthique dont l'usage des drones devrait s'inspirer (van Wynsberghe et al. 2018, 15).

Bref, un peu comme Rawls l'a fait en établissant des principes devant être respectés, ainsi qu'un ordre lexical entre eux, pour assurer la justice sociale, une éthique déontologique pour les décideurs politiques pourrait procéder par voie de hiérarchisation de principes, en incluant des procédures d'arbitrage entre eux. Dans des situations de dilemmes éthiques, comme celui du tramway, l'idée sera de procéder en identifiant des principes à suivre (par exemple, la protection des droits des personnes). En effet, selon Blattberg, un déontologue soutiendrait que « […] nous pouvons garder nos mains propres en respectant certains principes formels plutôt qu'en promouvant une fin quelconque » [Traduction libre] (Blattberg 2018, 156). Il est toutefois important que les lois qui sont mises de l'avant dans les codes d'éthique aient une applicabilité réelle, tangible. Sinon, les principes n'auront aucune portée concrète.

Conclusion

Ce premier portrait de la littérature grise en éthique de l'IA présente des traits fort intéressants. D'une part, il est difficile de trouver des documents entièrement inspirés par une approche éthique moniste comme l'éthique de la vertu, l'utilitarisme ou encore une éthique déontologique. Cela ne signifie pas que ces traditions ne figurent pas dans les démarches à l'étude. Elles y sont dans une certaine mesure, sous certains aspects, et ce travail de détection est évidemment interprétatif. Plus encore, l'inclusion de certaines écoles éthiques recèle des

conséquences politiques. On a vu, par exemple, qu'il est difficile d'adopter l'entièreté de l'éthique de la vertu sans brimer la liberté individuelle. D'autres approches éthiques peuvent se révéler intéressantes pour des robots, comme le « procéduralisme », mais mal adaptées pour les humains. C'est le cas de l'éthique déontologique. Le chapitre suivant, qui complète la deuxième section de la thèse, vient mettre en lumière un phénomène intéressant et inattendu dans mon éthique de l'IA, soit ce que j'appellerai la « tension métaéthique ». Avant de s'y plonger, il faut faire un tour d'horizon des directives éthiques informées par le pluralisme des valeurs.

Chapitre 5 — Démarches pluralistes et en tension

> « *[…] les principes éthiques de l'IA peuvent placer les acteurs dans une situation morale difficile en établissant des responsabilités éthiques envers les différentes parties prenantes sans offrir de conseils sur la manière de trouver des compromis lorsque les besoins ou les attentes de ces parties prenantes sont en conflit*. »
> — Michael A. Madaio et al. [Traduction libre] (2020, 2)

Introduction

Il est à présent clair que les décideurs politiques contemporains sont confrontés à une panoplie d'enjeux posés par l'essor de l'intelligence artificielle. Les chartes et directives éthiques, on le sait, se sont multipliées lors des dernières années. S'y retrouver est une tâche herculéenne qu'un élu politique n'a probablement ni le temps, ni les moyens d'affronter. C'est pour cette raison que les chapitres quatre et cinq, formant la deuxième section de cette thèse, proposent un portrait des ancrages métaéthiques d'un échantillon de ces directives. On verra, dans ce chapitre, le rôle crucial — mais implicite — du pluralisme des valeurs dans ces démarches. Il faut réitérer que les auteurs des directives ne prennent pas explicitement position sur le continuum s'étirant entre le monisme et le pluralisme. En revanche, ils laissent des indices quant à leur positionnement sur ce dernier, et c'est à partir de ces éléments que je tâcherai de « catégoriser » ces initiatives.

La manière de procéder est semblable à celle du chapitre précédent. Dans un premier temps, je traiterai des directives pluralistes telles que trouvées dans l'échantillon à l'étude, non sans avoir, auparavant, donné une idée de ce dont aurait l'air une réponse pluraliste à l'éthique de l'IA. Dans un deuxième temps, je me pencherai sur les démarches présentant une tension métaéthique, en précisant d'emblée ce que j'entends par cette notion de « tension », et en amenant quelques précisions la concernant au moyen des travaux de deux penseurs, W.D. Ross et Amartya Sen. Puis, j'esquisserai ce dont pourrait avoir l'air une démarche laissant percer des tensions métaéthiques, pour ensuite en suggérer des illustrations dans les directives à l'étude. Je conclurai ce chapitre avec

quelques réflexions sur les implications concrètes de la « tension métaéthique » pour les décideurs politiques.

1. Les démarches pluralistes

a) Esquisse d'une éthique pluraliste face à l'intelligence artificielle

Pour les penseurs pluralistes, une approche à l'éthique de l'intelligence artificielle ne pourrait prendre la forme d'une théorie ou d'un ensemble systématique. Plutôt, on aura recours à un état des lieux de la réalité pratique dans laquelle des valeurs, des principes et des intérêts s'affrontent, se révélant souvent incompatibles. La « théorie éthique unifiée » est un non-sens pour les pluralistes. Une telle vision

> […] peut suggérer que les pratiques morales sont assez simples*, régies par des règles claires dont les applications ne prêtent pas à controverse*. Comme n'importe quel agent moral mature le saurait, cependant, il serait tout à fait irréaliste de caractériser la vie morale en ces termes. Comme nous le rappelle l'expérience courante, il est souvent compliqué de déterminer ce que nous devrions faire, et lorsque nous le découvrons, il se peut fort bien que nous ne soyons pas motivés à faire ce qui s'impose. [Traduction libre, je souligne] (Nørskov et Rodogno s.d., 4-5)

Cela ne signifie pas que les tenants d'une éthique procédurale tiendraient pour acquis que toute procédure est empreinte de simplicité. Ces derniers soutiennent également que de déterminer ce qu'il faut faire est parfois très ardu. Cependant, des procédures, des théories ou des systèmes éthiques, ancrés dans la raison théorique, existent, et peuvent servir de guide sans entacher l'agent moral. C'est précisément ce dernier aspect que réfutent les penseurs pluralistes. Ce n'est pas la complexité de la réalité éthique qui est comprise différemment chez les monistes et les pluralistes, mais la manière de concevoir le conflit. Les monistes penseront qu'une divergence donnée est simplement un désaccord, tandis que les pluralistes se déclareront devant un véritable « conflit », c'est-à-dire un différend potentiellement insoluble, impliquant un compromis ou une perte.

Dans le même ordre d'idées, une approche éthique pluraliste aux enjeux de l'intelligence artificielle informerait les parties prenantes que, si tous les intérêts peuvent certes être accommodés dans une certaine mesure, ils ne peuvent toutefois pas être réconciliés sans perte. Des compromis

devront être effectués entre des valeurs — par exemple, entre la protection de la vie privée et la garantie de la sécurité des utilisateurs d'un système d'IA en particulier. Ou encore, la restriction de la liberté d'expression en ligne, pour le profit d'une navigation plus sûre en ce qui concerne l'origine des nouvelles. Les nouveaux enjeux éthiques sont plutôt ardus, en ce qu'ils « […] n'ont pas de réponses objectives[29] […] » [Traduction libre] (Millar 2016, 789).

Une manière de parvenir à ces compromis de valeurs serait par l'entremise de la négociation. Par ce mode de dialogue, les interlocuteurs parviendront à mettre en balance ces valeurs (ou principes, intérêts), en visant une minimisation des pertes. On l'a vu : la négociation est bien présente dans la pensée de Hampshire, comme l'est la notion de compromis chez Berlin. L'acte de négocier est en réalité une « nécessité » pour un pluraliste (Blattberg 2004, 27). La négociation

> […] consiste à soupeser les différentes valeurs en présence et à les confronter en recherchant un équilibre, comme dans un jeu à somme nulle. Par conséquent, si la métaphore de la théorie neutraliste est l'arbitre qui suit les indications du livre de règles, on pourrait dire que celle du pluralisme est le groupe de joueurs qui, sans qu'aucun arbitre n'intervienne, échangent sur la façon dont ils pourraient créer un équilibre entre leurs valeurs respectives, en plaçant celles-ci sur les plateaux d'une balance aux rouages extrêmement complexes. (Blattberg 2004, 29)

L'appel aux parties prenantes, à la négociation avec elles et à la pesée des valeurs préalablement identifiées, seraient des traits distinctifs d'une approche éthique pluraliste. Cela étant dit, la seule mention de « parties prenantes » n'est pas suffisante pour attribuer à toute une directive éthique l'étiquette de « pluraliste ». Il faudrait d'autres indicateurs pour en arriver à cette conclusion. Le fait demeure, bien entendu, que les intentions des rédacteurs de ces documents restent quelque peu opaques.

La nécessité de dialoguer pour arriver à un compromis entre les tenants de différentes valeurs s'impose, pour les pluralistes, dans la mesure où chaque innovation technologique générera d'elle-même des gains et des pertes. On pourrait illustrer cette affirmation en ayant recours à l'exemple des voitures autonomes. Pilotées par une intelligence artificielle guidée par des gigaquantités de données agrégées dans un système infonuagique, ces automobiles pourraient

[29] La manière dont je comprends l'objectivité ici est au sens de la théorie, et non au sens de la pratique.

générer des gains de temps et d'efficacité et de sécurité. D'une part, le gain de temps viendrait du fait que l'on pourrait s'occuper autrement, une fois en voiture, que de la conduite elle-même. D'autre part, l'apport en efficacité pourrait s'expliquer par le fait que les décisions stratégiques de la voiture, pour se rendre d'un point A à un point B, ne seraient pas ralenties par des faiblesses typiquement humaines comme la fatigue ou diverses émotions potentiellement négatives.

Cependant, si la technologie « donne », elle « reprend » aussi, affirme le critique social Neil Postman (1998, 1). Cela est vrai au plan sociétal comme pour celui des utilisateurs individuels. Ainsi, on pourrait identifier de possibles pertes de valeurs dans un rythme de vie accéléré par l'impératif croissant d'efficacité, de même qu'un appauvrissement des capacités motrices, observatrices et de jugement pratique de l'humain en locomotion. D'autres conséquences négatives pourraient être observées dans une déculpabilisation potentielle face à la tragédie (par exemple, un accident de la route), entraînant la détérioration des relations humaines interpersonnelles (Nørskov et Rodogno s.d.) et, en fin de compte, une perte de confiance envers la technologie. Des penseurs éthiques pluralistes auront donc le souci, que ce soit dans la conception et le développement des produits (Millar 2016, 789-790), dans le conseil à leur usage ou dans l'élaboration de politiques publiques les concernant, d'identifier ces conflits de valeurs d'emblée et de tenter de les équilibrer, dans le but de minimiser les pertes aussi tôt que possible.

Finalement, une éthique pluraliste de l'intelligence artificielle pourrait prendre appui sur des chartes de droits humains, dans le but d'élaborer des principes concernant l'usage des SIA. Au chapitre précédent, il a été question de la compréhension moniste (déontologique) des droits de la personne, par exemple. Dans une éthique informée par le pluralisme des valeurs, la compréhension de ces chartes et listes de principes ne se fera pas selon un système unifié ou hiérarchisé. Plutôt, les droits seront perçus comme autant d'entités pouvant entrer en collision les unes avec les autres. Par ailleurs, le développement d'une éthique de l'IA sur la base de chartes de principes connaît actuellement une certaine popularité. On le verra dans l'analyse de l'échantillon de directives éthiques sur l'IA que je propose dans ce chapitre, visant à mettre en lumière les influences du pluralisme dans les réflexions qui les ont mises au jour. Comme pour les écoles monistes, je présenterai en premier lieu les directives émanant des compagnies du secteur industriel, puis de la

société civile et des groupes à multiples partenaires, pour ensuite me diriger vers celles des instances gouvernementales ou intergouvernementales internationales.

b) Le pluralisme des valeurs dans les démarches recensées

1. Les directives des entreprises privées

Les directives éthiques en provenance du secteur industriel sont nombreuses à présenter des affinités avec le pluralisme des valeurs. De toutes les traditions éthiques, il s'agit manifestement de la plus populaire dans les entreprises privées. Ces dernières ne présentent en effet qu'une seule directive plutôt moniste, soit les principes de Google entourant l'IA, que l'on a étudiés sous la loupe de l'utilitarisme. On pourrait aussi les considérer comme influencés par le pluralisme des valeurs, comme on l'a vu au chapitre précédent, puisque la catégorisation est quelque peu malaisée, et relève d'une interprétation que j'ai offerte, mais que je suis loin de considérer comme tranchée de façon finale.

Le même géant du Web publie, en 2019, deux études sur l'avenir de l'intelligence artificielle : « Perspectives on Issues in AI Governance » (Google 2019a) et « Responsible Development of AI » (Google 2019b). La deuxième étude, tout en s'adressant aux gouvernements, ne traite toutefois pas d'éthique, ce qui explique que je ne l'ai pas retenue dans mon échantillon, contrairement à la première. Cette première directive consiste en un livre blanc à l'intention des gouvernements et de la société civile, dans laquelle Google fait la promotion de l'autorégulation de même que de la corégulation, par les compagnies, en ce qui a trait à l'éthique de l'IA. Cela ne l'empêche pas de souhaiter que des normes internationales soient développées pour encadrer le développement et l'usage de l'intelligence artificielle (Google 2019a, 4). La compagnie dit s'attendre à ce que, si la communauté globale s'implique dans un débat sociétal et gouvernemental — ce qui lui apparaît souhaitable — davantage de nuances seront amenées, ainsi qu' « […] une compréhension des *compromis* et des possibles *conséquences involontaires* que des *choix difficiles* tendent à impliquer » [Traduction libre, je souligne] (Google 2019a, 2). Ces allusions aux compromis, à des conséquences involontaires que l'on devine négatives, ainsi qu'aux « choix difficiles » peuvent renvoyer au pluralisme des valeurs. Le fait que la compagnie procède à une

analyse qualifiée d' « […] approche collaborative multipartite [*multi-stakeholder*] […] » (Google 2019a, 4) contribue aussi à considérer cette réflexion comme éminemment pluraliste, en raison du fractionnement avéré des intérêts en jeu.

Google mentionne également la possibilité de pertes dues à des différences de points de vue. Malgré cela, l'entreprise affirme qu' « […] il devrait être faisable de s'entendre sur une liste de facteurs à prendre en considération » [Traduction libre] (Google 2019a, 7). Plus concrètement, elle soutient que dans le *design* de produits,

> il est également important de tenir compte de *tout compromis potentiel* par rapport à l'exactitude du système. Dans le cas des applications pour lesquelles un rendement « acceptable » suffit, on pourrait *privilégier l'explicabilité plutôt que l'exactitude*; dans d'autres cas où la sécurité est primordiale, on pourrait *accorder la priorité à l'exactitude* tant et aussi longtemps que d'autres mécanismes sont en place pour assurer la responsabilité. [Traduction libre, je souligne] (Google 2019a, 11)

On peut comprendre de cet extrait que même la définition de concepts, ou plus précisément des principes auxquels les rédacteurs de chartes éthiques font appel, peuvent générer des conflits.

Plus encore, des glissements de sens pourraient survenir, dans l'usage des notions et valeurs mentionnées. Le principe d'équité en est un bon exemple. Google l'a retenu comme concept clé pour sa réflexion sur la gouvernance de l'IA, admettant qu'il existe une panoplie de définitions « conflictuelles » de l'équité (Google 2019a, 13). Par conséquent,

> différentes approches techniques aboutiront à des modèles qui sont équitables de différentes façons. Décider lequel utiliser nécessite un raisonnement éthique, et cela est très *spécifique au contexte. Étant donné la diversité des points de vue et des approches en matière* de définition de l'équité, *certaines définitions peuvent entrer directement en conflit les unes avec les autres*, tandis que d'autres peuvent favoriser l'équité uniquement *au détriment* de l'exactitude ou de l'efficience. […] Il serait utile d'avoir une clarté morale sur la façon dont le secteur public fait des *compromis* d'équité dans le contexte de décisions spécifiques. Bien que *les réponses varieront probablement selon les cultures et les régions géographiques*, il serait utile pour les entreprises qui doivent faire de tels compromis d'avoir une compréhension commune de l'incidence de ces décisions et de certains repères d'orientation. [Traduction libre, je souligne] (Google 2019a, 13)[30]

[30] Comme je l'ai évoqué au chapitre trois, les pluralistes des valeurs reconnaissent un « noyau conceptuel » universel aux valeurs, hors des contextes spécifiques dans lesquels elles sont promues (Apfel 2011,18-19). Quant à la clarté morale désirée par Google, elle peut être d'ordre théorique (par exemple, une définition) ou pratique (comme une manière de procéder à sa mise en œuvre), toujours selon une perspective pluraliste.

Un des aspects frappants de cet extrait est de constater qu'un acteur influent du secteur privé s'intéresse à la manière dont le secteur public fait face aux compromis de valeurs, dans le but de s'en inspirer. Autrement dit, on tient pour acquis qu'entre autres choses, la politique est une affaire de négociations, un jeu à somme nulle. De plus, l'importance de la sensibilité au contexte (notamment les contextes culturels variant selon les situations géographiques), garant de la diversité des points de vue, est aussi mise de l'avant. Les échos des fondements de l'approche éthique pluraliste précédemment analysés dans le chapitre trois sont patents.

L'enjeu de la sécurité représente un compromis particulièrement épineux. Google reconnaît que la sécurité est « […] un enjeu avec lequel [ils] ont longtemps été aux prises », puisque l'ouverture et la mise en disponibilité des ressources comme Android ou TensorFlow (la bibliothèque Google pour l'apprentissage machine) entraîne le risque de leur mauvais usage. Ainsi, « au fur et à mesure que l'écosystème évolue [Google] continue d'*évaluer les compromis* entre les bénéfices de l'ouverture et le risque d'abus […] » [Traduction libre, je souligne] (Google 2019a, 13). L'allusion aux bénéfices et aux abus pourrait faire penser à l'utilitarisme. Cependant, la centralité de la notion de compromis vient camper cette réflexion du côté pluraliste, dans la mesure où ces compromis impliquent des pertes. En effet, une théorie éthique comme l'utilitarisme reconnaît la possibilité de compromission de valeurs. Toutefois, ces accommodements ne salissent pas les mains des agents qui les effectuent. La valeur de la sécurité est souvent sujette à des arbitrages, en ce qui a trait aux technologies émergentes. Il suffit de penser aux technologies de surveillance qui fonctionnent grâce à l'IA, comme des caméras urbaines permettant la reconnaissance faciale. D'un côté, il y a perte quant à la protection de la vie privée. D'un autre côté, force est d'admettre que de tels outils, bien employés, pourraient aider à combattre le crime ou aideraient à retrouver des personnes disparues (Google 2019a, 29).

La compagnie multinationale d'informatique Microsoft a aussi publié une directive éthique d'inspiration pluraliste. Contrairement à celle de Google, il ne s'agit pas d'un livre blanc, mais plutôt d'une charte de principes éthiques pour des produits employant l'intelligence artificielle. On y trouve une liste de « valeurs intemporelles » : l'équité, l'inclusion, la fiabilité, la sécurité, la transparence, le respect de la vie privée ainsi que la responsabilité (Microsoft 2018a, s.p.). Ces dernières ne sont pas liées entre elles selon un ordre sériel. Microsoft ne semble pas fournir de clé

d'interprétation et d'harmonisation de ces valeurs, comme une théorie ou une procédure. Elles ne sont pas non plus réductibles à une valeur primaire qui les supplanterait potentiellement, comme on le verrait dans une tradition utilitariste. Face à ce constat, on pourrait considérer cette démarche éthique comme relevant du pluralisme des valeurs.

De manière semblable à Microsoft, le laboratoire de recherche OpenAI, fondé par Elon Musk et Sam Altman, a lancé sa propre charte éthique, concernant plus spécifiquement l'intelligence artificielle générale (*AGI*). Dans le but d'agir « dans l'intérêt supérieur de l'humanité », OpenAI met de l'avant quatre principes, soit « des avantages largement répartis, la sécurité à long terme, le leadership technique [et] l'orientation coopérative » [Traduction libre] (OpenAI 2018a, § 1-5). Encore une fois, il s'agit d'une liste de valeurs et de principes qui n'est pas unifiée en un tout. Ces derniers sont tout simplement énoncés. On peut déduire que dans la pratique, ils pourront entrer en conflit, et aucune procédure d'arbitrage n'est fournie par OpenAI pour les accommoder. Également, la préoccupation pour l'intérêt de l'humanité peut s'apparenter à des énoncés d'intentions semblables à ceux que l'on retrouve dans d'autres directives éthiques. Il a été question, au chapitre précédent, de la mention fréquente de vouloir être « bénéfique à l'humanité ». Le fait que la charte d'OpenAI porte plus particulièrement sur l'intelligence artificielle générale pourrait contribuer à expliquer la mention à l'humanité. Selon cette analyse, je dirais qu'il s'agit ici aussi d'une démarche éthique pluraliste.

De son côté, Google DeepMind a, dès 2017, lancé son programme *Ethics and Society*, affirmant souscrire à des « […] valeurs scientifiques telles que la transparence, la liberté de pensée et l'égalité d'accès […] » [Traduction libre] (DeepMind 2017, s.p.). L'entreprise fait la promotion de thèmes de recherche comme « la vie privée, la transparence et l'équité, la moralité et les valeurs des IA, [ainsi que] la gouvernance et la responsabilité » [Traduction libre] (DeepMind, s.p.). DeepMind procède aussi en posant des questions qui sont propres à chaque champ de recherche. Il n'y a pas beaucoup de détails entourant l'approche éthique du laboratoire. Si les valeurs scientifiques sont pensées comme pouvant, dans la pratique, entrer en collision de manière à nécessiter une forme d'accommodement, alors peut-être conclurait-on que la tradition éthique à laquelle DeepMind fait appel est le pluralisme des valeurs. L'absence de mention d'une règle, d'un

ordre ou d'une procédure ne permet pas vraiment de camper l'approche du côté moniste du continuum métaéthique, du moins en ce qui concerne l'utilitarisme et le déontologisme.

Il est intéressant de mentionner que le fondateur de DeepMind, Mustafa Suleyman, a cofondé, avec Eric Horvitz, de Microsoft, la coalition à but non lucratif Partnership on AI (PAI). Cette coalition œuvre dans le but de développer l'IA « pour le bien » (Horvitz et Suleyman 2016, §1-4). En janvier 2020, le PAI a lancé une initiative cherchant à combler l'écart entre les bonnes intentions et la pratique en ce qui a trait aux principes éthiques pour l'IA, intitulée *Closing Gaps in Responsible AI*. Le projet comprend « […] plusieurs phases impliquant diverses parties prenantes, visant à faire émerger la sagesse collective de la communauté pour identifier les changements saillants et évaluer les solutions potentielles » [Traduction libre] (Cavello 2020, §2). D'emblée, la mention de multiples parties prenantes peut traduire une fragmentation des intérêts et des participants. Cependant, elle révèle peut-être seulement la multiplicité des points de vue différents sur la question. Un premier volet du projet est le *Closing Gaps Ideation Game*, qui consiste en un processus interactif au cours duquel le joueur est encouragé, tout seul ou avec l'aide de réponses fournies par des participants ayant joué précédemment, à identifier des écarts, défis et des solutions potentielles entre les intentions et la pratique éthiques. Encore à ses débuts, l'initiative présentera sans doute d'autres aspects éthiques qui mériteront une analyse comme celle-ci. Comme pour DeepMind, on pourrait penser que l'initiative aurait des affinités pluralistes, mais il est difficile de l'affirmer hors de tout doute.

Il est toutefois clair que le secteur privé fournit des exemples de directives éthiques inspirées du pluralisme des valeurs, même si les exemples proposés sont plutôt gris que blancs ou noirs. Des entreprises qui en mènent large comme Google et Microsoft, de même que des laboratoires de recherche en intelligence artificielle, présentent des traits et réflexions qui rappellent les fondements du courant pluraliste des valeurs. Malgré cela, le pluralisme des valeurs, en tant qu'école éthique, n'est pas explicitement nommé dans les directives que j'ai consultées. Cela n'est pas surprenant quand on considère que ce courant tend à être passé sous silence quand l'occasion se présente de nommer les traditions éthiques principales qui informent nos considérations sur le bien et le mal.

2. Les directives de la société civile et des organisations à multiples partenaires

Si on se tourne du côté du secteur public, les directives entourant l'éthique de l'IA ont des origines diverses. Une série d'acteurs de la société civile se sont rencontrés pour produire plusieurs déclarations de principes qui ont eu une résonnance importante entre 2016 et 2020. Ainsi, les principes d'Asilomar ont été endossés par DeepMind, GoogleBrain, Facebook, Apple et OpenAI, de même que par plus de mille cinq cents chercheurs et experts en intelligence artificielle (Future of Life Institute [FLI] 2017, §5). Le FLI, qui organisait la rencontre à l'origine de cette charte, est avare de détails sur les participants. Néanmoins, on sait qu'elle constituait « […] un rassemblement de personnes infiniment accomplies et intéressantes » [Traduction libre] (The FLI Team 2017, §1). Les principes et valeurs ont été adoptés moyennant la formule du consensus. Cela dit, ce dernier « […] peut parfois être obtenu aux frais de l'abstraction et du choix de mots, qui peuvent masquer le désaccord » [Traduction libre] (Boddington 2017, 105). C'est du moins ce que suggère la philosophe Paula Boddington dans son analyse critique des principes d'Asilomar. Ce que cette dernière relève, c'est que la discussion pour arriver à une série de valeurs sera nécessairement parsemée de conflits, que la volonté d'arriver à un « consensus » peut masquer. Boddington ne reconnaît pas le pluralisme des valeurs comme une approche éthique normative à part entière (2017, 8), mais sa critique pourrait y être rattachée. Précisément, c'est au caractère sérieux compromis et des pertes que des philosophes comme Berlin et Williams souhaitaient accorder une importance particulière dans leurs travaux.

Séparée en trois catégories, la liste de principes contient des enjeux de recherche, les défis de l'éthique et des valeurs et finalement, les problématiques à long terme. Elle se veut non exhaustive ni exempte de possibles ambiguïtés dans leur interprétation. Force est de constater que l'énumération des principes rappelle une démarche pluraliste. Sont compilés des enjeux comme « la sécurité », « le manque de transparence », « la transparence judiciaire », « la responsabilité », « la vie privée », « la liberté et la protection de la vie privée », « la prospérité partagée » et « le contrôle humain », entre autres [Traduction libre] (Future of Life Institute (FLI) 2017, §3). Encore une fois, la directive ne présente pas de caractéristique moniste, c'est-à-dire d'aspect unifiant la démarche en un tout. La mise en relation des valeurs et principes est vraisemblablement vouée à engendrer des conflits, voire à mettre au jour des incompatibilités dans des problèmes de la vie

pratique. En somme, simplement énoncer des valeurs qui ont généré une forme de consensus dans la discussion, en les posant comme des principes éthiques, est une façon de procéder qui implique, sans le nommer, le pluralisme. Si ces derniers étaient considérés comme des maximes, des impératifs ou encore des lois éthiques, une catégorisation moniste serait plus aisée.

3. Les directives des organisations gouvernementales, intergouvernementales ou internationales

En jetant un regard vers les organisations gouvernementales, intergouvernementales ou internationales, on peut constater que deux instances ont adopté une démarche à sensibilité pluraliste pour leur positionnement éthique. Il s'agit des membres de la Commission européenne en 2018, ainsi que du groupe des pays du G20 en 2019. Premièrement, la Commission européenne a énoncé les valeurs éthiques qu'elle considère comme centrales dans les débats sur l'IA. Dans sa déclaration de 2018 sur la robotique, les systèmes autonomes et l'intelligence artificielle, les auteurs appellent à la création d'une charte d'éthique de l'IA qui soit internationalement reconnue et adoptée, dans le but d'avoir une approche commune aux enjeux de l'IA (European Commission 2018c, 5, 20). C'est ainsi que la Commission

> préconise le lancement d'un processus qui ouvrirait la voie à un cadre éthique et juridique commun et internationalement reconnu pour la conception, la production, l'utilisation et la gouvernance de l'intelligence artificielle de la robotique et des systèmes « autonomes ». La déclaration propose également *un ensemble de principes éthiques fondamentaux*, fondés sur les *valeurs* énoncées dans les traités de l'UE et la Charte des droits fondamentaux de l'UE, qui peuvent guider son développement. [Traduction libre, je souligne] (European Commission 2018c, 5)

L'idée de cadre commun est ici ambiguë, comme l'est la mention d'« un ensemble de principes éthiques fondamentaux ». Elle pourrait signaler une conception moniste de l'éthique. Néanmoins, la suite de la déclaration vient clarifier les choses. En effet, le développement rapide de l'intelligence artificielle soulève des questions morales de plus en plus pressantes, soutient la Commission. Plusieurs enjeux se posent, ayant trait à

> […] l'explicabilité et à la transparence de l'IA et des systèmes « autonomes ». Quelles sont les *valeurs* que ces systèmes servent efficacement et de manière démontrable? *Quelles valeurs* sous-tendent la manière dont nous concevons nos politiques et nos

> machines? Autour de *quelles valeurs* voulons-nous organiser nos sociétés? Et quelles sont les *valeurs que nous laissons saper* — ouvertement ou silencieusement — dans les *compromis* entre progrès technologie et services publics? [Traduction libre, je souligne] (European Commission 2018c, 8-9)

Il est clair, à la lecture de ce passage de la déclaration, que l'approche préconisée par la Commission rappelle le pluralisme. Il est question de centrer la réflexion autour de valeurs identifiées, dont des pertes sont évoquées, à l'issue de compromis. Il s'agit d'une terminologie très pluraliste.

Par ailleurs, Commission préconise aussi des réflexions éthiques élargies, qui dépassent les exercices de pensée comme le dilemme du tramway. Autrement, des questions peuvent en venir à être négligées,

> […] comme « quelles décisions de conception [des systèmes IA] ont été prises dans le passé, et qui ont conduit à cette situation morale », *« quelles valeurs* devraient guider la conception », *« comment les valeurs de conception* doivent-elles être *pesées* en cas de *conflit*, et par qui » […]. [Traduction libre, je souligne] (European Commission 2018c, 11)

Le conflit entre les différentes valeurs et la nécessité de soupeser ces dernières entre elles sont tenus pour acquis par les rédacteurs de la déclaration. Plus encore, l'approche par valeurs est mentionnée à de nombreuses reprises dans le document. De fait, les auteurs présentent une liste de « principes éthiques et de prérequis démocratiques » qui sont les suivants : 1) la dignité humaine, 2) l'autonomie, 3) la responsabilité, 4) la justice, l'équité et la solidarité, 5) la démocratie, f) l'État de droit et l'imputabilité, 8) la sécurité, l'intégrité mentale et physique, 9) la protection des données et de la vie privée, puis 10) la durabilité (*sustainability*) (European Commission 2018c, 16-19). Le fait de mentionner la démocratie comme « prérequis démocratique » relève peut-être d'un argument circulaire.

Enfin, la Commission européenne, dans son engagement à développer des lignes directrices éthiques pour la fin de l'année 2018, s'engage à « […] rassembler toutes les *parties prenantes* concernées afin de contribuer à l'élaboration de ce projet de lignes directrices » [Traduction libre, je souligne] (European Commission 2018a, 15). On l'a vu, cette terminologie n'est pas nécessairement emblématique de pluralisme des valeurs. Toutefois, en combinaison avec le reste

du propos présenté dans la déclaration, il semble clair qu'il s'agit d'un document pouvant être rattaché au positionnement métaéthique du pluralisme des valeurs.

Deuxièmement, les pays du G20 ont adopté cinq « Principes pour une gestion responsable d'une IA digne de confiance ». Ces derniers sont 1) la croissance inclusive, le développement durable et le bien-être pour toutes les parties prenantes, 2) l'équité, les valeurs centrées sur l'humain, 3) la transparence et l'explicabilité, 4) la robustesse et la sécurité ainsi que 5) la responsabilité (G20 Countries 2019, 1-2). Chaque principe est suivi d'une petite spécification, ou encore de quelques sous-points concernant les « acteurs de l'IA ». Par exemple, la précision suivant le premier principe stipule que

> les parties prenantes devraient s'engager de manière proactive dans la gestion responsable d'une IA digne de confiance afin d'obtenir des résultats bénéfiques pour les personnes et la planète, tels que l'augmentation des capacités humaines et le renforcement de la créativité, l'avancement de l'inclusion des populations sous-représentées, la réduction des inégalités économiques, sociales, de genre et autres, et la protection des environnements naturels, stimulant ainsi la croissance inclusive, le développement durable et le bien-être. [Traduction libre] (G20 Countries 2019, 1)

Même s'il est question de bénéfices généraux et de bien-être, la diversité de valeurs mentionnées et leur non-ordination dans une hiérarchie poussent à penser que la tradition éthique informant ce raisonnement est le pluralisme des valeurs. De plus, il est également question de parties prenantes, c'est-à-dire d'intérêts que l'on peut supposer fragmentés d'emblée. Certes, les pays rédacteurs soutiennent qu'il importe de respecter les chartes de droits en place, mais cela est mentionné aux côtés de l'impératif de respecter également des valeurs démocratiques, comme « […] la liberté, la dignité et l'autonomie, la protection de la vie privée et des données, la non-discrimination et l'égalité, la diversité, l'équité, la justice sociale et les droits du travail reconnus au plan international » [Traduction libre] (G20 Countries 2019, 1). Il semble que les chartes de droits ainsi que les valeurs démocratiques sont conçues comme des collections d'éléments qui eux, ont le potentiel d'entrer en conflit les uns avec les autres. Enfin, le groupe de pays signataires fournit aussi cinq recommandations pour la coopération internationale devant se faire, pour assurer une intelligence artificielle « digne de confiance » (G20 Countries 2019, 3-4).

En définitive, il apparaît distinctement que le pluralisme des valeurs informe un certain nombre de directives éthiques, émanant des secteurs privé et public. Néanmoins, le pluralisme des valeurs, comme positionnement métaéthique, n'est pratiquement jamais mentionné lorsqu'il s'agit de lister les approches éthiques existantes. Les auteurs ne tendent qu'à recenser les approches monistes qui ont dominé dans les derniers siècles, à savoir le déontologisme, l'utilitarisme et l'éthique de la vertu. La démonstration ci-dessus apporte, je le crois, un élément important au portrait des traditions éthiques existant dans les enjeux ayant trait à l'intelligence artificielle. Le pluralisme des valeurs est peut-être négligé, mais il est bien présent. De plus, il se présente également sous d'autres formes auxquelles on ne s'attend pas, à savoir, en combinaison avec des éléments monistes. C'est ce que j'ai identifié comme la « tension métaéthique ».

2. Les démarches en tension métaéthique

a) Précisions sur la tension métaéthique

De la vingtaine de documents recensés pour esquisser le portrait des traditions éthiques à l'œuvre dans les études des secteurs privé et public sur l'IA, le plus grand nombre est à trouver du côté des démarches à sensibilité pluraliste (huit directives) et affichant une tension métaéthique (neuf directives). Étant donné la combinaison de traditions éthiques ayant des postulats contradictoires sur la résolution de conflits, la tâche du décideur politique, appelé à légiférer quant à l'éthique de l'IA, est rendue plus complexe. Par exemple, la posture à adopter face aux dilemmes éthiques peut poser problème.

J'entends par « tension éthique » la réalité de la cohabitation, dans une même réflexion, d'éléments provenant d'une tradition éthique moniste, ainsi que d'éléments pluralistes. Une combinaison de traits éthiques distincts, originaires de traditions monistes différentes, ne serait pas n'exhiberait pas cette difficile cohabitation. Elle serait plutôt « hybride ». Donc, pour parler de tension, il faut que le monisme et le pluralisme se côtoient dans la même directive. Comme cela n'est pas explicité dans les documents à l'étude, ou ne semble pas être pris en compte, la tension, telle que je la décris, pose ou posera problème.

La catégorie de directives exprimant une tension métaéthique est directement issue de mon analyse des documents. Je ne l'avais pas prévue, puisque j'ignorais son existence avant le début de ma lecture plus approfondie des directives. La découverte de cette réalité, mais aussi de sa popularité, est un des apports que j'espère intéressants pour cette thèse. D'une part, il s'agit d'une contribution descriptive, au sens où elle clarifie le portrait de ce qui se fait actuellement en éthique pour l'intelligence artificielle. D'autre part, cette spécification sera utile pour mon argumentation normative dans la troisième et dernière section de la thèse.

Le constat et la critique de la tension métaéthique ne se veulent pas une invalidation de ce qui a été accompli en éthique de l'IA jusqu'à présent. Souvent, les réflexions éthiques sont diverses parce qu'elles présentent la richesse épistémique d'avoir inclus toute une portion d'une population donnée au processus d'élaboration de principes. Néanmoins, quand on analyse en profondeur les démarches éthiques des secteurs privé et public, la réalité est que plusieurs choisissent des éléments provenant de postulats métaéthiques incompatibles entre eux, mais sans en être conscients ni en approfondir les implications. Cette inconscience est ce qui pose réellement problème, puisqu'elle ouvre la porte à une incohérence interne des démarches éthiques proposées. D'une certaine façon, des éléments « dépareillés » ont été agencés, présumément sans réflexion préalable sur leurs ancrages métaéthiques respectifs. Parfois, au sein d'une même directive, on évolue dans un cadre moniste, pour ensuite se déplacer, sans crier gare, dans les repères du pluralisme.

Au chapitre deux, on a vu que certaines écoles éthiques monistes peuvent exhiber une rigidité mal adaptée aux contextes de la pratique, dans la procédure à suivre pour résoudre des dilemmes éthiques. Je développerai davantage cet argumentaire au chapitre six. Par contraste, le pluralisme préconise une sensibilité au contexte avec une certaine prudence. Sous cet angle, monisme et pluralisme sont en tension. Plus saillante encore est la question des « mains sales ». Une approche moniste, dans un document, peut n'impliquer théoriquement aucune compromission de valeurs. À l'opposé, dans la même démarche, non seulement le pluralisme croit ces compromissions incontournables, mais parfois même au point parfois d'en exagérer les pertes, comme je l'ai exposé au chapitre trois. Il sera donc difficile, pour un décideur, d'intégrer tous les éléments d'une directive pour la mettre en pratique. Avant de présenter au lecteur des exemples concrets de cette tension à partir de l'échantillon utilisé pour cette thèse, j'aimerais apporter une

dernière précision concernant la cohabitation du monisme et du pluralisme, cette fois dans la pensée de certains philosophes.

1. Sur la tension entre monisme et pluralisme chez certains philosophes

J'aimerais préciser que les catégories décrites ci-dessus peuvent être liées à des « idéaux types » métaéthiques. Dans la littérature académique en éthique, de nombreux penseurs sont difficiles à catégoriser, car ils empruntent à plus d'une tradition et il arrive que ces traditions ne soient pas compatibles. Certains d'entre eux peuvent même donner l'impression de fusionner sans problème le monisme et le pluralisme. Cela dit, si l'on creuse davantage, il est possible de faire entrer leur pensée dans l'une des catégories que j'ai décrites, ou dans un de leurs sous-ensembles.

Par exemple, un penseur tel que W.D. Ross, qui rejette une conception de l'éthique comme étant dérivée d'une théorie morale similaire aux sciences naturelles, pourrait sembler étranger à une approche procédurale à première vue (alors que je l'ai traité, au chapitre deux, comme un déontologiste). En effet, Ross attache une grande importance à la sensibilité au contexte, à l'image d'Aristote et des pluralistes des valeurs. Cependant, il est bien un éthicien des devoirs, proposant une catégorisation de six types de « devoirs *prima facie* » (2003, 59). Ceux-ci reposent sur des principes tels que la bienfaisance, la justice et la non-malfaisance. C'est de lui que Beauchamp et Childress s'inspirent pour développer le principlisme, qui contient ces trois principes ainsi que celui du respect de l'autonomie (Beauchamp et Childress 2001, x; Thornton 2006, §6).

Pour Ross, les devoirs *prima facie* doivent être honorés, mais, si les circonstances exigent même qu'une promesse soit rompue

> […] pour soulager la détresse de quelqu'un, nous ne cessons pas un instant de prendre conscience d'un devoir *prima facie* de tenir notre promesse, et cela nous amène à ressentir, *non pas vraiment de la honte ou du repentir*, mais certainement de la componction, pour avoir agi comme nous le faisons […]. [Traduction libre, je souligne] (Ross 2003, 64-65).

En d'autres termes, un déontologiste peut prendre conscience des conflits de devoir sans admettre une perte morale qui conduit à une véritable culpabilité. Il ressort clairement de ce passage que

pour Ross, les conflits de devoir sont malheureux, mais ils ne salissent pas les mains de l'agent. À cet égard, dans sa réflexion sur l'éthique aristotélicienne, Karen M. Nielsen abonde dans le même sens. Selon elle, « […] il n'est pas nécessaire d'accepter le paradoxe des mains sales pour accepter que "notre vie [puisse] être moralement difficile" » [Traduction libre] (2007, 294). Dans ce cas, la réalité éthique peut être unifiée en principe, sans culpabilité morale. L'argumentaire de Ross, comme celui de Nielsen, est finalement moniste. Ainsi, la reconnaissance d'une *multiplicité* de devoirs, de valeurs ou de biens n'implique pas automatiquement un *pluralisme* des valeurs au sens métaéthique. Ce qui caractérise ce dernier est l'incompatibilité des valeurs et, par conséquent, le fractionnement irrémédiable de la réalité éthique, ce qui implique une culpabilité morale ou des mains sales.

Un autre exemple de pluralisme apparent, mais dans un contexte utilitariste cette fois est celui d'Amartya Sen, qui admet « […] une bonne part de pluralisme dans le cadre de l'utilité elle-même […] » [Traduction libre, je souligne] (Sen 1980, 208). La diversité des biens admise dans sa vision « vectorielle » de l'utilité reconnaît cependant

> […] la nécessité de *justifier toutes les valeurs morales par rapport à un aspect de l'utilité* (par exemple la satisfaction de certains désirs) [et elle] continue d'agir comme une contrainte contraignante et puissante. [Traduction libre, je souligne] (Sen 1980, 210)

Manifestement, le principe de l'utilité est la clé à partir de laquelle interpréter et ordonner tous les autres, ce qui campe la pensée de Sen dans le champ moniste. Selon lui, une théorie de la justice qui ne tombe pas dans le piège de l'abstraction universelle (comme une théorie kantienne) peut accueillir une multiplicité de « considérations non concordantes » sans rendre ces dernières « […] incohérentes, ou ingérables, ou inutiles. Des conclusions définitives peuvent émerger malgré la pluralité » (Sen 2009, 397).

Peut-être Sen est-il influencé par le pluralisme des valeurs, mais cela ne fait pas de lui un penseur pluraliste. Il m'apparaît plus moniste que pluraliste en raison de son attachement à la valeur de l'utilité et à son positionnement hiérarchique. Il en va de même pour Ross. Ainsi, un philosophe qui accorderait une place égale ou semblable au monisme et au pluralisme dans ses œuvres serait hautement paradoxal — un « pluramoniste », selon Blattberg (2018, 159-161). Ce positionnement

ne pourrait être qu'un oxymore, car on ne peut pas affirmer l'unité et la fragmentation ultime de la réalité simultanément. Il n'est pas possible d'avoir les mains propres et sales en même temps.

Ross et Sen ne semblent pas être, au premier abord, des penseurs du paradoxe. Néanmoins, on peut être un philosophe moniste « sensible » aux idées du pluralisme des valeurs, sans déroger de manière ultime à son monisme — c'est-à-dire sans verser dans l'incohérence interne. C'est de cette manière que les philosophes Martha Nussbaum et Charles Taylor seraient des « monistes non orthodoxes ». Je reviendrai en détail sur cette idée au chapitre suivant. Pour le moment, il convient de se pencher sur la tension métaéthique que j'ai évoquée, de même que sur le type d'éthique de l'IA qui en pourrait en découler.

b) Esquisse d'une démarche affichant une tension métaéthique face à l'intelligence artificielle

Il n'est pas aisé de déterminer à quoi ressemblerait « typiquement » une directive éthique qui présenterait une tension métaéthique. Les possibilités de combinaisons sont plutôt vastes. Ce qui caractérise tout spécialement une démarche éthique en tension, c'est que l'on a du mal à la classer soit dans le monisme, soit dans le pluralisme. Lui attribuer une catégorie entre les deux génère le sentiment d'une inadéquation que l'on se doit de mentionner. Ma recherche m'a fait voir qu'un document éthique peut receler une démarche combinant de l'éthique de la vertu, l'éthique déontologique ou encore de l'utilitarisme, avec certains aspects du pluralisme de valeurs. Il est, dans de tels cas, difficile d'accoler une étiquette à une directive éthique semblable. Cette tension est présente du côté des compagnies privées, de celui de la société civile et des groupes de multipartenaires, ainsi que du côté des organisations de gouvernance internationale, soit de tous les types de documents à l'étude.

Un exemple hypothétique pourrait proposer une éthique utilitariste qui accepte la possibilité de conflits de valeurs insolubles et la tragédie qui peut suivre. Conséquemment, cette éthique proposerait une chose et son contraire. D'une part, la doctrine utilitaire selon laquelle, en suivant la procédure de la maximisation de l'utilité, l'agent ne se compromettra pas moralement. D'autre

part, la conviction pluraliste que les accommodements impliquant des pertes, voire des compromissions morales, sont inévitables dans l'absolu. Il y aurait là tension métaéthique. Par contraste, chez un utilitariste de la règle, la situation serait tout autre. Au chapitre deux, il a été question que l'utilitarisme de la règle puisse donner l'impression de faire ménage avec le déontologisme. Si c'était le cas, l'on assisterait une hybridité éthique tout au plus, non à ce que j'appelle une forme de tension. De fait, l'utilitarisme de la règle fait bon accueil aux principes, les règles ainsi que les droits, pris en système. Cependant, cette acceptation n'est faite que pour garantir l'utilité à long terme. L'objectif ultime est toujours de maximiser l'utilité. Le système de droits qu'un utilitariste de la règle pourrait évoquer est donc un moyen vers cette fin. Les droits seraient, en dernière instance, commensurables au principe suprême de l'utilité. À présent, au lieu de penser la tension métaéthique de façon hypothétique, c'est dans ses expressions réelles, dans les documents à l'étude, qu'on l'examinera.

c) La tension métaéthique dans les démarches recensées

1. Les directives des entreprises privées

Dans son livre, *The Future Computed : Artificial Intelligence and Its Role in Society*, Microsoft appelle à une forme d'entente autour de six principes devant encadrer le développement et l'usage de l'IA (Microsoft 2018b, 57, 9). L'ouvrage est adressé non seulement à ceux qui travaillent en IA, mais aussi aux acteurs gouvernementaux, aux universitaires et aux travailleurs du monde des affaires. Il est intéressant de noter que, si l'approche d'identification de principes non érigés en système unifié peut renvoyer au pluralisme des valeurs, l'objectif final de développer des règles à suivre peut vouloir se diriger vers une démarche plus déontologique, procédurale. Cette approche de réglementation, Microsoft l'associe à la sphère publique. Effectivement, ce qui constitue l'objectif final est d'élaborer

> […] un cadre commun de principes pour guider les chercheurs et les développeurs dans la mise au point d'une nouvelle génération de systèmes et de capacités basés sur l'IA, et les gouvernements dans leur réflexion sur une *nouvelle génération de règles et de réglementations* visant à protéger la sécurité et la vie privée des citoyens et à garantir que les avantages de l'IA soient largement accessibles. [Traduction libre, je souligne] (Microsoft 2018b, 49)

On peut voir le mélange métaéthique de cette façon. D'une part, le passage donne à comprendre que les principes éthiques pourraient être mis en pratique « directement » par les développeurs. On ne sait pas exactement comment : cela dit, les principes sont perçus comme pouvant guider les concepteurs de nouveaux systèmes IA. Les principes ne sont pas organisés en hiérarchie dérivée de la théorie, mais ils forment partie d'un « cadre ». L'idée de cadre est unificatrice, mais sa seule mention ne fait pas de tout un document une démarche moniste.

D'autre part, on perçoit que ces mêmes principes peuvent inspirer de nouvelles « règles et réglementations », mais cette fois du côté gouvernemental. Autrement dit, les principes énoncés, sans être mis en application dans la réalité pratique de manière immédiate, seraient plutôt à la base de nouveaux systèmes de règles ou de lois. Cette façon de voir renvoie plutôt au monisme. Enfin, Microsoft suggère que les

> […] gouvernements travaillent avec les entreprises et les autres *parties prenantes* pour trouver *l'équilibre nécessaire* afin de *maximiser le potentiel de l'IA pour améliorer la vie des gens* et relever les nouveaux défis à mesure qu'ils se présentent. [Traduction libre, je souligne] (Microsoft 2018b, 75)

La mention des parties prenantes et d'un processus d'équilibrage des valeurs peut faire penser à une compréhension pluraliste du principe de proportionnalité. Au fond, les valeurs sont soupesées entre elles pour arriver à un équilibre. La référence à la maximisation de ce que l'IA a à offrir de positif pour favoriser des retombées positives dans la vie des gens pourrait renvoyer à l'utilitarisme, sans qu'on puisse le garantir. Néanmoins, avec la combinaison de tous les éléments mentionnés jusqu'à présent, il appert que la directive de Microsoft présente une forme de tension métaéthique. La compagnie propose par ailleurs d'avoir recours à une norme de négligence (*negligence standard*), un outil provenant du droit. Cet outil, qui serait de nature plutôt déontologique, est mis de l'avant par la compagnie pour que les décideurs politiques puissent réaliser l'équilibrage (pluraliste) entre la sécurité et l'innovation (Microsoft 2018b, 82). Devant cette combinaison complexe de monisme et de pluralisme, la pertinence d'une catégorie de démarches « en tension » apparaît peut-être plus clairement au lecteur.

Toujours dans le milieu industriel, la multinationale International Business Machines (IBM) a publié en 2014 un rapport concernant l'éthique de l'intelligence artificielle, mis à jour en 2019,

qui s'adresse à ses employés, mais aussi aux développeurs en IA (International Business Machines Corporation [IBM] 2019). Un peu semblable aux deux exercices de l'IEEE (que les auteurs du rapport citent d'ailleurs comme un exemple à reproduire [International Business Machines Corporation [IBM] 2019, 11], la réflexion éthique porte surtout sur les systèmes employant l'intelligence artificielle, plutôt que les personnes qui en font et feront usage. En ce sens, le document touche de plus loin les décideurs politiques. Je l'ai retenu, toutefois, en raison du sujet plus large de l'éthique de l'IA, et parce que, comme mentionné précédemment, il est difficile de séparer « dans l'absolu » les catégories d'éthique de l'IA. Dans le document, il est question des « vertus spécifiques » que les systèmes d'intelligence artificielle devront exhiber [International Business Machines Corporation [IBM] 2019, 10]. De même, à plusieurs endroits dans le document, revient ce concept de vertu. L'influence de l'éthique de la vertu n'est cependant pas la seule dans cette réflexion, comme dans la plupart des autres documents de ce genre. En effet, s'il est mention de vertus, il est aussi mention de valeurs, un terme qui peut renvoyer à l'éthique pluraliste ou encore tout simplement un langage courant sur la morale.

En ce qui a trait à ces valeurs, IBM favorise ce qu'on appelle « l'alignement des valeurs ». Cette approche, fréquemment mentionnée quand il est question d'éthique de l'IA (notamment, encore une fois, dans les démarches du IEEE), consiste à veiller à une certaine concordance entre les valeurs humaines et celles que l'on désire voir respectées par les machines. Cet alignement de valeurs, même s'il est techniquement l'affaire des concepteurs et des développeurs, nécessite l'apport de la société dans son ensemble, et plus précisément des décideurs politiques, du moins pour l'identification des valeurs désirées. En effet,

> la prise de décision éthique n'est pas seulement une autre forme de résolution de problèmes techniques. […] Une approche centrée sur la technologie, qui ne vise qu'à améliorer les capacités d'un système intelligent, ne tient pas suffisamment compte des besoins humains. Une IA éthique, centrée sur l'humain, doit être conçue et développée de manière à s'aligner sur les valeurs et les principes éthiques d'une société ou de la communauté qu'elle touche. [Traduction libre] [International Business Machines Corporation [IBM] 2019, 10]

Ainsi, étant donné qu'il s'agit d'éthique et que la portée des technologies employant l'IA est appréciable, l'enjeu des valeurs ne peut demeurer une question uniquement « technique ».

Ce qui est intéressant avec la question de l'alignement des valeurs, c'est qu'elle n'est pas ici traitée comme relevant d'un système unifié. Au contraire, il s'agit plutôt d'un principe parmi d'autres. À ce titre, les rédacteurs du rapport structurent leur propos autour de cinq champs pour l'étude de l'IA, soit la responsabilité, l'alignement des valeurs (*value-alignment*), l'explicabilité, l'équité, et les droits des utilisateurs sur les données (*user data rights*) (2019, 12). IBM prône aussi le design d'un encadrement éthique par un *Ethics Canva*, en suivant la méthode du *human-value design* (2019, 24). Cela renvoie à une approche pluraliste où une multiplicité de valeurs sont mises en jeu les unes avec les autres, sans ordre de préférence ni procédure ordonnée de sélection.

Dans ce même document, toutefois, on retrouve des influences monistes. De fait, il est question de notions comme des droits et des obligations qui, eux, peuvent être associés à l'éthique déontologique ou encore l'utilitarisme de la règle. Pour les rédacteurs du document,

> l'éthique est basée sur des normes bien fondées de bien et de mal qui prescrivent ce que les humains doivent faire, généralement en matière de *droits, d'obligations, d'avantages pour la société, d'équité ou de vertus spécifiques*. [Traduction libre, je souligne] [International Business Machines Corporation [IBM] 2019, 10].

De toute évidence, cette façon de concevoir l'éthique penche du côté du monisme, que ce soit l'éthique de la vertu ou le déontologisme. Conséquemment, IBM exhibe, dans son positionnement éthique, une sorte de combinaison entre le monisme et le pluralisme, et il n'est pas aisé de déterminer si une approche métaéthique devrait avoir préséance sur une autre dans la pratique et si oui, la forme qu'elle prendrait.

2. Les directives de la société civile et des organisations à multiples partenaires

L'Institut des ingénieurs électriques et électroniques (IEEE) se trouve à la jonction des secteurs privé et public. Le projet *Principled Artificial Intelligence* de Harvard la classifie comme une démarche de multiples parties prenantes, que je classe à mon tour dans celle des directives émanant de la société civile et de multiples partenaires. En effet, les contributeurs aux deux directives éthiques qui ont été publiées, en 2016 et 2017, comportent des professeurs d'université, des juristes, des consultants du monde des affaires, des spécialistes de l'industrie technique, des hauts placés d'organisations non gouvernementales, et ce, du monde entier [Institute of Electronics

and Electrical Engineers, Incorporated [IEEE] 2019, 2-11]. Pour chaque principe ou notion mise de l'avant dans le document, une équipe multidisciplinaire d'experts des quatre coins de la planète avait préalablement été consultée (IEEE 2019).

Les directives du IEEE font partie des documents les plus cités quand il est question d'éthique pour l'intelligence artificielle, en particulier en ce qui a trait à l'élaboration des SIA. Par exemple, le livre blanc de la compagnie Google, *Perspectives on Issues in AI Governance*, voit dans un partenaire comme le IEEE une bonne piste pour le développement de standards globaux à long terme (Google 2019a, 7). Les auteurs du livre blanc du Forum économique mondial recommandent le IEEE comme approche pratique pour la mise en pratique de l'éthique dans les systèmes d'intelligence artificielle [World Economic Forum [WEF] 2019a, 10]. Puis, lors du Sommet de 2017 tenu par les Nations Unies sur le potentiel bénéfique de l'intelligence artificielle, on note que l'initiative du IEEE en éthique a été mentionnée lors de plusieurs sessions sur l'éthique et la vie privée (ITU & the XPRIZE Foundation 2017, 43, 67). L'Institut en question a même profité du Sommet pour rencontrer ceux qui le désiraient dans le cadre de ses consultations publiques (ITU & the XPRIZE Foundation 2017, 43). Le Comité sur la science et la technologie de la Chambre des communes britannique, dans son rapport sur la robotique et l'IA (House of Commons of the United Kingdom, Science and Technology Committee 2016, 23), tente de donner une impulsion au gouvernement pour l'adoption de sa stratégie digitale ainsi que de sa stratégie pour le monde du travail. Déjà, des cadres de gouvernance avaient été adoptés, par l'Union européenne, mais également par le IEEE. Le rapport conjoint de l'Assemblée nationale et du Sénat de France mentionne aussi le travail effectué par le IEEE (2017, 185). Bref, il s'agit d'une initiative qui a eu, et continue d'avoir, une importante résonnance en éthique de l'intelligence artificielle.

Dans la première version de l'étude, le IEEE lance une série P7000 de standards industriels pour les systèmes autonomes (SA) et l'IA. Ces standards se retrouvent à la jonction entre les standards éthiques et les standards techniques (National Institute of Standards and Technology, U.S. Department of Commerce 2019, 28-29). De tous les documents que j'ai consultés dans mon analyse, ceux du IEEE sont les seuls à mentionner explicitement l'*eudaimonia* aristotélicienne. Les auteurs la définissent comme

> une pratique qui définit le bien-être humain comme la plus haute vertu d'une société. […] En alignant la création de l'IA et des SA sur les valeurs de ses utilisateurs et de la société, nous pouvons donner la priorité à l'augmentation du bien-être humain comme mesure du progrès à l'ère de l'algorithme. [Traduction libre] [Institute of Electrical and Electronics Engineers, Incorporated [IEEE] 2016, 2]

On retrouve, comme dans la démarche d'IBM décrite plus haut, la volonté de procéder par alignement des valeurs, et d'opérer de façon « prioritaire » dans le cadre plus large de la « bonne vie », concept inspiré d'Aristote, mais non repris dans son entièreté.

Malgré la mention de l'*eudaimonia*, il serait inexact de dire que les études du IEEE sont des documents complètement inspirés de l'éthique de la vertu. On y trouve plutôt des influences déontologiques ainsi que quelques traces d'utilitarisme positif et négatif [Institute of Electronics and Electrical Engineers, Incorporated [IEEE] 2016, 16, 5]. En effet, les principes généraux du document forment un appel au respect « des plus hauts idéaux des droits humains » en tenant compte des chartes de droits humains déjà élaborées et en vigueur [Institute of Electronics and Electrical Engineers, Incorporated [IEEE] 2016, 16], à « maximiser les bénéfices à l'humanité et à l'environnement naturel », à « atténuer les risques et les retombées négatives de l'intelligence artificielle et des systèmes autonomes ». Une série de questions est associée à chacun de ces enjeux [Traduction libre] [Institute of Electronics and Electrical Engineers, Incorporated [IEEE] 2016, 5].

Les principes qui sont mis dans l'avant dans le rapport sont le bénéfice humain, la responsabilité, la transparence, l'éducation et la sensibilisation [Institute of Electronics and Electrical Engineers, Incorporated [IEEE] 2016, 16-21]. L'énumération de principes peut faire penser au pluralisme des valeurs. Cependant, certains principes, pris isolément, semblent renvoyer à l'utilitarisme, sans être pour autant inscrits dans un cadre purement conséquentialiste. Dans le cas de l'intelligence artificielle générale (ou généralisée) et de la superintelligence, les auteurs du rapport soulignent l'importance des principes de sécurité et de bienfaisance (IEEE 2016, 7).

L'idée maîtresse de la démarche menée par le IEEE est de mettre en œuvre (*implement*) des valeurs dans les systèmes intelligents autonomes (*AIS*), en suivant trois étapes principales :

> 1) « identifier les normes et valeurs d'une communauté spécifique touchée par les systèmes intelligents autonomes »,

> 2) « mettre en œuvre les normes et les valeurs de cette communauté à l'intérieur des systèmes intelligents autonomes » et
> 3) « évaluer l'alignement et la compatibilité de ces normes et valeurs entre les humains et les systèmes intelligents autonomes à l'intérieur de cette communauté ». [Traduction libre] (IEEE 2016, 5-6)

Il est spécifié que ces normes et ces valeurs « […] ne sont pas universelles, mais plutôt largement spécifiques à la communauté d'utilisateurs et aux tâches » [IEEE 2016, 24). L'importance de la sensibilité au contexte peut faire penser à l'éthique de la vertu ou encore au pluralisme des valeurs, en ce qu'elle ne se prétend pas universelle. Néanmoins, l'analyse ultérieure révèle que le document est difficile à étiqueter. En effet, quand il est question de l'alignement des valeurs entre les humains est les SIA, les auteurs affirment que leur recommandation est de

> […] donner la *priorité aux valeurs* qui reflètent l'ensemble des valeurs communes des *parties prenantes* les plus importantes. […] Par exemple, le « Principe du Bien Commun »[31] pourrait être utilisé comme ligne directrice pour *résoudre les différences dans l'ordre de priorité* des différents groupes d'intervenants. Nous recommandons aussi que *l'ordre de priorité des valeurs prises en compte à l'étape de la conception des systèmes autonomes ait une politique claire et explicite de justification*. […] [Avoir une telle politique] […] surtout lorsque *ces valeurs sont en conflit les unes avec les autres*, non seulement encourage les concepteurs à réfléchir sur les valeurs mises en œuvre dans le système, mais fournit également une fondation et un point de référence pour un tiers [de façon à ce qu'il puisse] comprendre le processus de pensée du ou des concepteurs. […] Nous reconnaissons en outre que, selon le système autonome en question, *la priorité de l'ordre des valeurs peut changer dynamiquement d'un contexte d'utilisation à l'autre, ou même au sein du même système au fil du temps*. [Traduction libre, je souligne] (IEEE 2016, 25)

Plusieurs éléments entrent en ligne de compte dans ce seul passage. Il est question de valeurs, de conflits entre elles et d'échelles de valeurs qui peuvent différer d'une personne, voire d'un contexte à l'autre. La hiérarchisation des valeurs, qui dérive ici non de la raison théorique, mais de la sensibilité au contexte, est entièrement compatible avec le pluralisme des valeurs (non décisionniste). Autrement dit, le pluralisme des valeurs peut accueillir une hiérarchie de valeurs, si elle ne relève pas d'un système unifié. Cette ordination changera selon le contexte et les personnes en jeu, sans être figée ou rigide. La mention des « parties prenantes » pourrait achever de dresser un portrait pluraliste de la démarche.

[31] Les auteurs renvoient, par lien hypertexte, au texte de Velasquez, Andre et Shanks (2014).

Toutefois, la question de la priorisation des valeurs, de leur interprétation à partir d'un principe (le « Principe du Bien Commun »), l'impératif d'avoir une « politique claire et explicite de justification », peuvent évoquer des conceptions monistes de l'éthique. Certes, il serait possible d'assigner à cet extrait une étiquette ou une autre. À mon sens, quand on le regarde attentivement, il révèle des postulats métaéthiques incompatibles. De plus, la tension métaéthique n'est pas surprenante s'il est question d'alignement des valeurs entre les humains et les machines. Le seul exercice de relever les valeurs qui sont communes aux êtres humains se fait souvent par l'approche pluraliste, car elle permet de rendre compte de la multiplicité des points de vue. C'est du moins l'approche qu'a préconisée le FLI avec les principes d'Asilomar, les pays du G20, la Commission européenne ainsi que plusieurs acteurs du secteur privé, dont Google et Microsoft. En revanche, la question de la « mise en œuvre » ou de l'intégration de ces valeurs dans des systèmes intelligents, étant une question plus technique, ferait meilleur ménage avec une approche moniste comme l'utilitarisme ou l'éthique déontologique. C'est d'ailleurs une des conclusions de la seconde étude du même Institut, en 2017.

Parmi d'autres recommandations qui sont mises de l'avant dans le document, on retrouve celle d'inclure davantage d'éthique dans les programmes des futurs développeurs et concepteurs en IA; d'être sensibles à des valeurs non occidentales, dans le but de les incorporer aux différents systèmes; le besoin d'inclure les différentes parties prenantes dans la démarche éthique; de même que la nécessité d'une étude sur l'intelligence artificielle générale ou superintelligence artificielle (IEEE 2016, 37, 29, 44, 49-55). Les données personnelles, les armes autonomes, les enjeux économiques et humanitaires et le développement des nations forment également une grande partie de l'analyse (IEEE 2016, 56-95).

Toujours dans l'étude de 2016, de nouveaux comités sont formés pour se pencher sur la pertinence des différentes traditions éthiques pour les technologies employant l'intelligence artificielle. Ces études doivent paraître dans le second rapport d'IEEE. On y reconnaît de prime abord qu'

> une approche éthique de la vertu a le mérite d'accomplir cela même sans avoir à postuler un « caractère » dans une technologie autonome, puisqu'elle met l'accent sur l'action habituelle et itérative axée sur l'atteinte de l'excellence dans un domaine choisi ou en accord avec un but directeur [Traduction libre] (IEEE 2016, 95)

Néanmoins, l'éthique classique a ses limites, soutiennent les auteurs. En effet,

> l'éthique classique peut grandement contribuer à explorer les préoccupations, mais elle n'offre pas toujours de solutions concrètes aux chercheurs, aux innovateurs et aux entreprises. L'éthique classique est-elle accessible et applicable en ce qui concerne les projets technologiques? Peut-être que les théories éthiques classiques traditionnelles ne sont pas adéquates pour la tâche à accomplir, qui est d'éclairer la conception de la valeur des machines. Une méta-analyse de « l'éthique classique » est nécessaire pour répondre aux objectifs énoncés. [Traduction libre] (IEEE 2016, 95)

Différents comités ont la tâche d'explorer ce que l'éthique classique a à dire au sujet de la responsabilité. Une autre avenue ouverte dans le document est de dépasser les traditions éthiques classiques occidentales identifiées comme le déontologisme, l'utilitarisme et l'éthique de la vertu. Enfin, les questions du « jargon » éthique et de la compréhension populaire sont aussi soulevées (IEEE 2016, 96).

En définitive, ce qui ressort de ce coup d'œil au premier rapport de l'IEEE est qu'il présente un assortiment de déontologisme, d'utilitarisme, de pluralisme des valeurs, le tout couronné par quelques allusions à l'éthique de la vertu. Les auteurs se défendent d'avoir en tête d'élaborer un code d'éthique, leur objectif étant plutôt la production d'un outil de référence, sans constituer pour autant une prise de position, une politique, ou encore un rapport formel. Au final, à défaut de se camper dans une tradition éthique ou une autre, les auteurs les invoquent toutes.

L'année suivante, dans le second rapport de l'IEEE, l'éthique classique fait son retour, cette fois bonifiée par l'éthique de l'information, l'éthique des technologies, l'éthique machine (IEEE 2017, 193), ainsi que par la volonté d'inclure des principes éthiques de traditions non occidentales. Les auteurs parlent d'un « monopole » des traditions éthiques occidentales, qui ne tiennent pas compte des approches Ubuntu, du confucianisme, du bouddhisme ou du shintoïsme (IEEE 2017, 203). L'objectif établi dans cette étude est de

> […] [créer] des systèmes autonomes et intelligents qui honorent explicitement les *droits humains inaliénables* et les *valeurs bénéfiques de leurs utilisateurs*, [pouvant] *donner la priorité à l'augmentation du bien-être humain* comme mesure du progrès à l'ère de l'algorithmique. Mesurer et honorer le potentiel d'une prospérité économique holistique devrait devenir plus important que de poursuivre des objectifs unidimensionnels comme l'augmentation de la productivité ou la croissance du PIB. [Traduction libre, je souligne] (IEEE 2017, 2)

Dans ce seul énoncé d'intention, on peut retrouver des influences déontologiques dans les droits humains inaliénables, possiblement pluralistes dans la mention des valeurs, et à saveur potentiellement utilitariste ou de l'éthique de la vertu dans l'insistance sur le bénéfice et la croissance du bien-être. Le bien-être (*well-being*) est décrit dans la démarche comme renvoyant à

> […] la satisfaction de l'humain par rapport à la vie et aux conditions de vie, l'épanouissement (*eudaimonia*) et les affects positifs et négatifs, conformément aux lignes directrices de l'Organisation de coopération et de développement économiques (OCDE) sur la mesure du bien-être subjectif. La définition holistique du bien-être englobe les circonstances individuelles, sociales, économiques et gouvernementales ainsi que les droits humains, les capacités, la protection de l'environnement et le travail équitable, car ces circonstances et bien d'autres encore constituent la base du bien-être humain. [Traduction libre] (IEEE 2017, 242)

Charles Ess voit dans cette conception du bien-être un héritage de l'éthique de la vertu (2019, 83-84). Originellement, j'y avais lu une influence utilitariste, en raison de la quantification du bien-être par des « métriques ». Cela étant dit, il n'est pas exclu que l'éthique de la vertu comprenne des mesures quantitatives de ses concepts et donc, dans cette optique, l'interprétation de Ess pourrait être plus exacte. Ce dont on peut être certain, somme toute, c'est qu'il s'agit ici d'une influence incontestablement moniste.

Puis, comme dans la démarche précédente, une grande importance est accordée à l'implication des différentes parties prenantes (IEEE 2017, 3), de même qu'à l'apport de plusieurs traditions éthiques. Les principes ou standards éthiques qui sont mis de l'avant dans cette analyse ont quelque peu changé. On retrouve encore les droits humains, ainsi que le bien-être, mais ce dernier, on l'a vu, est désormais mesuré en « métriques ». Ces unités de mesure contiennent des facteurs sociaux, psychologiques et environnementaux. S'y trouvent aussi la responsabilité, la transparence et une « sensibilisation au mauvais usage » [Traduction libre] (IEEE 2017, 6, 8). Diverses méthodologies sont employées, dont une forme de « conception axée sur la valeur » (*value-based design* en anglais), une approche pluraliste.

De son côté, l'éthique de la vertu a toujours une certaine influence dans l'approche derrière cette démarche pour intégrer des normes éthiques dans les systèmes autonomes. Ess relève ses différentes mentions au fil du document, par exemple avec l'importance de la *phronesis* (IEEE 2017, 207 dans Ess 2019, 83). L'éthique de la vertu est aussi évoquée dans la

programmation de codes et lexiques pour l'ordinateur, basés sur les approches éthiques occidentales (IEEE 2017, 199). Elle est également rebaptisée comme « comportement axé sur des objectifs » (*goal-directed behaviour*) (IEEE 2017, 214). Dans cette optique, l'une des recommandations de l'IEEE est de

> programmer des systèmes autonomes pour qu'ils soient capables de reconnaître le comportement de l'utilisateur comme étant celui de types spécifiques de comportements, et de conserver des attentes en tant qu'opérateur et co-collaborateur, de sorte que l'utilisateur et le système reconnaissent mutuellement les décisions du système autonome comme étant fondées sur l'éthique de la vertu. [Traduction libre] (IEEE 2017, 214)

Dans cette optique, la notion de « comportement axé sur des objectifs » pourrait tout aussi bien — sinon mieux — renvoyer au conséquentialisme qu'à l'éthique de la vertu.

Néanmoins, l'équipe derrière le rapport du IEEE se doit d'admettre que, en ce qui a trait à la programmation d'une forme d'éthique, les approches basées sur des règles (*rule-based ethics*) sont plus efficaces (IEEE 2017, 215). Dans ce sens, force est de constater que le pluralisme des valeurs serait reconnu comme non efficace pour la production de SIA éthiques, du moins dans la perspective du IEEE (même si une éthique pluraliste pourrait reconnaître l'efficacité comme l'une de ses valeurs). En effet, une éthique fondée sur des règles, que le IEEE désire privilégier, serait nécessairement moniste. Malgré cela, comme je l'ai exposé, les rédacteurs des deux études font montre non seulement d'hybridité éthique, mais de tension métaéthique, et aucune école n'est présentée comme primant sur les autres. En conséquence, prises ensemble, les deux études du IEEE présentent une sorte de pot-pourri de traditions éthiques.

Il est par ailleurs intéressant de noter que Charles Ess a lui aussi analysé les documents du IEEE en vue d'y déceler les approches éthiques à l'œuvre. Il constate que dans la seconde étude, celle de 2017, « […] le document du IEEE reprend explicitement l'éthique de la vertu, ainsi que l'éthique déontologique, l'utilitarisme et de multiples autres traditions mondiales comme fondements d'une telle conception éthique » [Traduction libre] (Ess 2019, 83). La seconde étude du IEEE est toutefois plus favorable à des approches monistes que pluralistes. Elle exhiberait moins de tension métaéthique, si les approches éthiques non occidentales qui sont évoquées peuvent être catégorisées comme monistes (ce que je n'ai pas vérifié, puisque la thèse porte sur les traditions

occidentales). Quant à lui, Ess relève que dans cette étude, le pluralisme éthique est mentionné de manière explicite (IEEE 2017, 207 dans Ess 2019, 83). En revanche, le pluralisme est ici entendu, selon ma compréhension, non au sens du pluralisme des valeurs, mais comme le désir de donner une voix à une multiplicité ou une diversité de traditions éthiques. C'est d'ailleurs ce que les rédacteurs affirment, en suggérant que de « faire intentionnellement de la place pour le pluralisme éthique est un antidote potentiel à la domination de la conversation par la pensée libérale, avec son héritage du colonialisme occidental » [Traduction libre] (IEEE 2017, 207).

Un autre exemple de tension métaéthique, tiré des initiatives de la société civile, peut être localisé dans la « Déclaration de Montréal pour un développement responsable de l'intelligence artificielle ». Elle est adressée à « […] toute personne, toute organisation de la société civile et toute compagnie désireuses de participer au développement de l'intelligence artificielle de manière responsable […] » (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018a, 5). La Déclaration de Montréal prend la forme d'une charte de principes, incluant

> le bien-être, le respect de l'autonomie, la protection de l'intimité et de la vie privée, la solidarité, la participation démocratique, l'équité, l'inclusion et la diversité, la prudence, la responsabilité et le développement soutenable. (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018a, 5)

Chaque principe est accompagné de quatre à dix questions visant à faciliter sa mise en œuvre. Les principes ne sont pas hiérarchisés entre eux. De plus, ils sont flexibles. Dans l'éventualité où l'un d'entre eux pourrait peser davantage qu'un autre selon le contexte (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018a, 5) — ce qui implique un jugement pratique — on pourrait supposer que la Déclaration de Montréal est une initiative éthique informée par le pluralisme des valeurs.

En revanche, la Déclaration présente aussi des aspects qui penchent vers le monisme. Le Comité d'élaboration de la Déclaration écrit que ses membres désirent, par les principes qu'ils mettent de l'avant, « […] orienter le développement de l'intelligence artificielle vers des finalités moralement et socialement désirables » (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018a, 7). À cet effet, les principes adoptés sont basés

> […] sur l'idée commune que les êtres humains cherchent à s'épanouir comme des *êtres sociaux* doués de sensations, d'émotions et de pensées, et qu'ils s'efforcent de *réaliser leurs potentialités* en exerçant librement leurs *capacités affectives, morales et intellectuelles*. Il incombe aux différents acteurs et décideurs publics et privés, aux niveaux local, national et international, de s'assurer que le développement et le déploiement de l'intelligence artificielle sont compatibles avec la protection et l'*épanouissement des capacités humaines fondamentales*. [je souligne] (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018a, 7)

La définition de l'être humain comme étant fondamentalement social et orienté vers le développement de ses potentialités, et visant l'épanouissement de ses capacités, n'est pas sans rappeler l'éthique de la vertu. On pourrait même suggérer un lien avec la vision de la démocratie sociale aristotélicienne que Nussbaum (1990) appelle de ses vœux. L'influence de cette idée d'épanouissement de l'être humain au moyen de ses potentialités trouve aussi un écho dans les réflexions de Norbert Wiener sur la cybernétique.[32] En effet, Wiener reprend le triple idéal français de liberté, d'égalité et de fraternité pour invoquer, entre autres choses, « […] la liberté de chaque être humain de développer dans sa liberté la pleine mesure des possibilités humaines qui s'incarnent en lui […] » [Traduction libre] (Wiener 1969, 143-144).

Même si, de toute évidence, la démarche et son résultat (la Déclaration) ne peuvent être identifiés à l'éthique de la vertu, une forme de monisme vient tout de même se profiler ici. L'aspect le plus moniste de la directive se trouve dans la volonté de cohérence entre les principes. Cette idée de cohérence dans l'interprétation s'éloigne tout à fait du pluralisme, qui tend plutôt à placer l'accent sur les compromis et les pertes. Les auteurs de la Déclaration soutiennent que les principes, « bien qu'ils soient divers […] doivent faire l'objet d'une interprétation cohérente afin d'éviter tout conflit qui empêche leur application. […] Il est impératif que l'interprétation soit cohérente » (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018a, 5). Il est clair que cette clé de lecture de la Déclaration se rattache au monisme.

Ainsi, on peut suggérer que dans la Déclaration de Montréal, on retrouve une forme de tension métaéthique. L'énonciation des principes et la manière d'y arriver sont informées par le pluralisme des valeurs. Son but premier est d'« […] identifier les principes et valeurs éthiques qui

[32] Je suis redevable ici, comme à d'autres endroits dans la thèse, au professeur Charles Ess pour avoir mis au jour cette connexion. On consultera par ailleurs avec profit (Ess 2019, 82-83), citant le même passage de Wiener.

promeuvent les intérêts fondamentaux des personnes et des groupes » et servir de « […] base pour un dialogue interculturel et international » (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018a, 5). La forme que prend la Déclaration est la plus éloquente quant au pluralisme, puisqu'il s'agit d'une liste de valeurs ou de principes qui ne sont pas hiérarchisés, ou organisés en système, et que les rédacteurs ne fournissent pas de procédure pour les honorer tous dans la pratique. Toutefois, la Déclaration se veut aussi une « boussole » et un « cadre éthique » (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018a, 7). Ces termes renvoient à une vision moniste de l'éthique, comprise comme un cadre, un système, avec une orientation claire et déterminée d'avance. De plus, sa méthode d'interprétation doit être moniste, c'est-à-dire exhiber une cohérence qui empêche les contradictions et, ultimement, les pertes.

3. Les directives des organisations gouvernementales, intergouvernementales ou internationales

Des directives éthiques émanant d'organisations internationales, trois semblent présenter un caractère métaéthique exprimant des tensions. Tout d'abord, la Commission européenne a mis sur pied, en juin 2018, un groupe d'experts de haut niveau en intelligence artificielle (GEHN IA). Pour ces derniers, une intelligence artificielle « digne de confiance » doit être légale, éthique et robuste (European Commission 2019b, § 2). Les lignes directrices que la Commission publie concernent les concepteurs et les développeurs de SIA, mais visent aussi à conseiller les décideurs politiques en matière d'IA. Ainsi, sept exigences essentielles sont mises de l'avant : 1) la robustesse technique, la supervision et l'agentivité humaine, 2) le respect de la vie privée et la gouvernance des données, 3) la transparence, 4) la diversité, 5) la non-discrimination et l'équité, 6) le bien-être sociétal et environnemental et enfin 7) la responsabilité (European Commission 2019b, § 3).

Le groupe d'experts de haut niveau admet que des tensions peuvent survenir entre ces principes éthiques et qu'il n'y a pas de solution unique à ce type de conflit (Groupe d'experts de haut niveau en IA [GEHN IA] 2019, 16). L'idée du rejet de la « solution unique » est combinée au rejet d'une éthique entièrement déontologique. De l'avis de ce groupe,

> un code de déontologie spécifique à un domaine donné — quel que soit le niveau de cohérence, d'élaboration et de détail de ses futures versions — *ne pourra jamais se*

> *substituer à un raisonnement éthique en tant que tel*, qui doit en toutes circonstances *rester sensible aux éléments de contexte*, qui *ne peuvent jamais être rendus dans des lignes directrices générales*. [Je souligne] (Groupe d'experts de haut niveau en IA [GEHN IA] 2019, 11)

Il est clair que cette prise de position peut présenter des affinités avec l'éthique de la vertu ou encore le pluralisme des valeurs, vu l'importance accordée au contexte et au rôle de la raison pratique. Sans mentionner cette dernière spécifiquement, les auteurs parlent tout de même du « raisonnement éthique », qui ne peut être remplacé par un code ou des lignes directrices.

En étant dirigées vers les parties concernées, les lignes directrices touchent à la fois les programmeurs et les machines elles-mêmes, sans vouloir remplacer la législation politique, mais en conseillant les décideurs politiques en matière d'IA (Groupe d'experts de haut niveau en IA [GEHN IA] 2019, 2, 4). Cette directive européenne se veut aussi fondée sur les droits humains tels qu'énoncés dans la Charte de l'Union européenne (Groupe d'experts de haut niveau en IA [GEHN IA] 2019, 7). Pour les experts,

> le respect des droits fondamentaux, dans le cadre de la démocratie et de l'État de droit, est le fondement le plus prometteur pour recenser les principes et les valeurs éthiques abstraits pouvant être concrétisés dans le contexte de l'IA. (Groupe d'experts de haut niveau en IA [GEHN IA] 2019, 12).

La préoccupation essentielle est d'assurer le bien-être humain, « […] que ce soit du point de vue de la qualité de vie ou de l'autonomie humaine et de la liberté nécessaire pour une société démocratique » (Groupe d'experts de haut niveau en IA [GEHN IA] 2019, 11).

La diversité des approches éthiques, dans cette même directive, est patente. On y retrouve, notamment, le pluralisme des valeurs dans l'énonciation et la mise en balance de valeurs et de principes, ainsi que dans la reconnaissance des conflits potentiels. On peut aussi dénoter une préoccupation pour le « bien-être humain », notion que l'on pourrait associer soit à l'éthique de la vertu (dans le sens d'un épanouissement ou de la « bonne vie »), ou encore à l'utilitarisme, dans le sens de bonheur, de plaisir ou d'utilité. Dans ce passage, le bien-être peut être appréhendé par la loupe de l'autonomie, de la liberté ou de la qualité de vie, des valeurs qui sont mentionnées de manière éparse.

Le groupe d'experts adopte aussi quatre principes qui sont mis de l'avant dans les lignes directrices, en plus des sept exigences essentielles susmentionnées, soit : 1) le respect de l'autonomie humaine, 2) la prévention de toute atteinte, 3) l'équité et 4) l'explicabilité (GEHN IA 2019, 9). Il faut noter que bien que ces principes soient énoncés sans hiérarchie, comme s'il s'agissait d'une liste de valeurs devant être sensibles au contexte, l'on trouve des traces de la tradition déontologique dans le raisonnement éthique des auteurs. En effet, les rédacteurs affirment que ces quatre principes

> […] sont présentés comme **des impératifs éthiques**, si bien que les professionnels de l'IA devraient en toutes circonstances s'efforcer d'y adhérer. Sans imposer de hiérarchie, [ils présentent] les principes […] de manière à refléter l'ordre d'apparition, dans la charte de l'Union, des droits fondamentaux sur lesquels ils se fondent […]. [Accent dans le texte] (GEHN IA 2019, 14)

Les auteurs admettent eux-mêmes ne pas ordonner leurs principes en hiérarchie. Néanmoins, l'ordre d'énonciation est calqué sur un autre ordre, dans lequel sont organisés en système ces droits, soit des chartes de droits fondamentaux. Les auteurs énoncent aussi aussi que

> si ces principes fournissent clairement des orientations destinées à trouver des solutions, ils n'en demeurent pas moins des *prescriptions éthiques abstraites*. On ne peut attendre des professionnels de l'IA qu'ils trouvent la solution adaptée sur la base des principes ci-dessus. *Il leur faut toutefois aborder les dilemmes et arbitrages éthiques selon une réflexion raisonnée et fondée sur des éléments probants, plutôt que sur la base de l'intuition ou d'un jugement aléatoire*. Il pourrait cependant exister des situations dans lesquelles *aucun arbitrage acceptable du point de vue éthique ne peut être déterminé*. Certains droits fondamentaux et principes connexes *sont absolus et ne peuvent dépendre d'un exercice de mise en balance* (par exemple, la dignité humaine). [Je souligne] (GEHN IA 2019, 16)

La combinaison de pluralisme et de monisme est tout à fait frappante dans cet extrait. La reconnaissance du fait que des principes abstraits orienteront l'action grâce à l'exercice de la raison pratique, dans des situations de dilemmes éthiques, rappelle le pluralisme des valeurs. De même, le fait d'admettre que dans certains cas, il n'y aura pas de « bonne solution », relève de la même approche métaéthique. Néanmoins, l'affirmation selon laquelle certains principes sont absolus et « ne peuvent être mis en balance », dans le même passage, exprime un point de vue éthique moniste. Ce monisme serait très près de l'éthique déontologique kantienne, mais d'un déontologisme qui rejetterait l'intuitionnisme moral. Enfin, comme on ne peut affirmer que la

démarche du groupe d'experts soit entièrement moniste, ou encore seulement pluraliste, force est d'admettre qu'il s'agit d'un exemple très clair de tension métaéthique.

Une autre illustration de cette difficile cohabitation peut être fournie avec le livre blanc du Forum économique mondial (FEM). Le FEM propose aux décideurs politiques une approche « holiste » à la gouvernance de l'IA pour que les valeurs et les principes éthiques soient programmés dans des SIA de manière à présenter des bénéfices pour l'être humain (World Economic Forum [WEF] 2019a, 4). Le document affiche des affinités avec le pluralisme des valeurs, mais en exprimant également des tendances déontologiques et des traits aristotéliciens. Un premier élément sur lequel se pencher est la « définition neutre » de l'éthique que propose l'organisme :

> on désigne couramment l'éthique comme l'étude de la moralité. La « moralité », aux fins de ce [livre blanc] est comprise comme un *système de règles et de valeurs guidant la conduite humaine, ainsi que des principes pour évaluer ces règles*. Conséquemment, un comportement éthique ne veut pas nécessairement dire un « bon » comportement. *Plutôt, un comportement éthique renvoie à la conformité avec des valeurs spécifiques*. De telles valeurs sont communément acceptées comme faisant partie de la nature humaine (par exemple, la protection de la vie humaine, de la liberté et de la dignité humaine). Ou encore, elles peuvent renvoyer à des attentes morales qui caractérisent *les croyances et les convictions de groupes de personnes spécifiques* (par exemple, des règles religieuses). Les attentes morales peuvent aussi être de nature individuelle (par exemple, l'attente d'un entrepreneur que ses employés acceptent le code de conduite spécifique de la compagnie). [Traduction libre, je souligne] (World Economic Forum [WEF] 2019a, 5)

Il est frappant de voir la cohabitation du déontologisme (l'idée de « système de règles »), le pluralisme (les valeurs et croyances portées des groupes distincts) et l'éthique de la vertu (la référence à la nature humaine) dans cet extrait du livre blanc. L'aspiration à la neutralité a peut-être entraîné le Forum économique mondial (FEM) vers un buffet de traditions qui sont entre elles contradictoires, à défaut d'en choisir une seule au détriment des autres. Toutefois, ce à quoi le rapport semble vouloir faire référence avec cette tentative de neutralité est l'exploration de différentes écoles éthiques selon lesquelles les systèmes d'intelligence artificielle pourraient être programmés. Par exemple « […] l'utilitarisme, l'éthique kantienne, les trois lois d'Asimov sur la robotique[33] (qui forment un système moniste) ou encore la règle d'or […] » [Traduction libre]

[33] Ces lois ont été proposées dans le roman d'Isaac Asimov, 2013 [1977] et s'énoncent comme suit : 1) « Un robot ne doit pas blesser un être humain ou, par son inaction, permettre à un être humain d'être blessé », 2) « Un robot doit

(World Economic Forum [WEF] 2019a, 8). Dans ce contexte, cependant, les rédacteurs du rapport penchent pour « […] une programmation technique de principes éthiques dans des robots de manière "casuistique" plutôt qu'une structure décisionnelle spécifiquement programmée pour la prise de décisions » [Traduction libre] (World Economic Forum [WEF] 2019a, 8). Dans cette optique, on se rapprocherait de la sagacité aristotélicienne (la *phronesis*) qui est promue à la fois par les pluralistes des valeurs et par les éthiciens de la vertu. Véritablement, le document du Forum économique mondial présente une volonté de sensibilité au contexte de chaque décision éthique, puisque les auteurs mentionnent les « […] applications spécifiques aux situations […] » [Traduction libre] (WEF 2019a, 8).

Outre ce qui a été évoqué, l'étude du Forum économique mondial propose aussi quelques pistes éthiques qui relèvent du monisme. Par exemple, les auteurs suggèrent de développer

> […] un système de surveillance piloté par l'IA qui contrôle la conformité d'une machine à *un ensemble prédéterminé de lois et de règles éthiques au niveau méta* (« une IA gardienne ») […]. Cette IA gardienne peut techniquement interférer dans le système de l'IA de base et corriger directement les décisions illégales ou contraires à l'éthique. […] De même, une IA tutélaire correspondante pourrait être programmée pour signaler la décision illégale ou contraire à l'éthique de l'IA de base à une autorité ou un organisme d'exécution approprié. [Traduction libre, je souligne] (WEF 2019a, 11)

Cette idée d'un système de règles éthiques qui en régit un autre est foncièrement déontologique, en ce qu'elle implique des impératifs à respecter, érigés en système unifié, avec une certaine marche à suivre. Parallèlement, une éthique déontologique dans la sphère politique, pour réguler les technologies d'intelligence artificielle, peut simplement consister à construire des hiérarchies de principes, qui seraient régulés par des corps gouvernementaux ou judiciaires différents, selon le cas.

Dans ce but, le FEM développé une pyramide de valeurs, de même qu'un « carré magique » (WEF 2019a, 13). À la base de la pyramide sont placées les valeurs éthiques fondamentales et inaliénables : la dignité humaine, l'humanité, la vie humaine, la santé et l'autonomie. Ces valeurs fondamentales sont réglementées par les conventions internationales, comme les Nations Unies

---

obéir aux ordres donnés par des êtres humains, sauf si ces ordres entrent en conflit avec la Première Loi » et 3) « Un robot doit protéger sa propre existence tant et aussi longtemps que cette protection n'entre pas en conflit avec la première ou la deuxième loi » [Traduction libre] (Anderson M. 2017, s.p.).

ainsi que les législations nationales (WEF 2019a, 13). Le niveau juste au-dessus comprend les valeurs éthiques constitutionnelles comme l'État de droit, la justice sociale, la démocratie, l'équité, le respect de la vie privée, le développement durable) et sont réglementées par les mêmes instances que le niveau inférieur des valeurs éthiques fondamentales et inaliénables (WEF 2019a, 13). Un peu plus haut encore se retrouvent les valeurs éthiques spécifiques aux groupes, selon les cultures et les croyances, qui sont prises en charge par les systèmes de certification et les statuts d'associations. Au sommet de la pyramide, sans surprise, se situent les valeurs éthiques procédant des croyances personnelles, donc relevant de l'individu. Elles sont gérées par des systèmes de certification contractuels (WEF 2019a, 13). La hiérarchisation et la réglementation caractérisant cette pyramide en font un exemple clair de monisme métaéthique.

Quant au « carré magique » du FEM, il renvoie à la régulation de la technologie qui « […] implique une décision concernant des valeurs qui doit être prise à la lumière de plusieurs principes fondamentaux, qui entrent parfois en conflit » [Traduction libre] (WEF 2019a, 13). Parmi ces principes potentiellement contradictoires, on note « […] le principe de compétition […] ainsi que d'autres principes normatifs de base tels qu'on les retrouve dans les droits fondamentaux, les principes constitutionnels et l'éthique » [Traduction libre] (WEF 2019a, 13). Dans le cas où des valeurs entreraient en conflit, il faudra procéder à un processus de « mise en balance » des valeurs (WEF 2019a, 12). Ce peut être le pluralisme des valeurs qui se glisse ici, ou encore une forme de conséquentialisme : « la question difficile pour les régulateurs est de savoir *comment équilibrer* les retombées positives potentielles, en tenant compte des obligations additionnelles qui peuvent surgir pour les compagnies d'intelligence artificielle » [Traduction libre, je souligne] (WEF 2019a, 12).

L'exemple donné pour illustrer ce conflit de valeurs est celui des jouets pour enfants qui emploient l'intelligence artificielle et par l'entremise desquels les compagnies productrices recueillent des données sur ces derniers. Les compagnies ont-elles une obligation de signaler des situations potentiellement dangereuses dont la connaissance leur est parvenue par cette « collecte de données », comme un enfant qui présenterait des symptômes de pensées suicidaires? (WEF 2019a, 12) Il s'agit du genre d'impasse devant laquelle des distributeurs de produits propulsés par des algorithmes d'intelligence artificielle pourraient être placés, et que la tradition éthique du pluralisme des valeurs présente comme une dynamique à somme nulle. En dernière

analyse, il est inévitable qu'un agent subisse une perte de quelque nature que ce soit. En somme, à la lumière des éléments de la directive du FEM, il est indéniable qu'une forme de tension métaéthique est, ici aussi, à l'œuvre.

La tension telle qu'elle s'exprime dans les directives du Groupe d'experts de haut niveau de la Commission européenne et du Forum économique mondial est assez aisée à percevoir. Toutes les démarches qui exhibent une tension métaéthique ne sont pas aussi simples à appréhender. C'est le cas des principes de l'Organisation de coopération et développement économiques (OCDE), qui sont présentés sous la forme d'une charte. Il s'agit des premiers principes de ce type à avoir été officiellement ratifiés par des gouvernements nationaux : ceux des trente-six pays de l'OCDE, mais aussi par l'Argentine, le Costa Rica, le Pérou, le Brésil, la Colombie et la Roumanie. Des représentants de plusieurs d'entre eux, de même que de différentes sphères de la société, constituaient un groupe de spécialistes qui ont fait valoir « cinq principes complémentaires fondés sur des valeurs pour une gestion responsable d'une IA digne de confiance » [Traduction libre] (Organisation for Economic Co-operation and Development [OECD] 2019, § 2).

La forme de la charte rappelle le pluralisme des valeurs, même si le contenu des principes contient d'autres influences éthiques. Le premier principe stipule que « l'IA devrait profiter aux personnes et à la planète, au développement durable et au bien-être », ce qui peut rappeler une forme douce d'utilitarisme, ou bien une référence s'apparentant davantage à l'éthique de la vertu. Il est difficile de le dire avec ce seul énoncé. Le second soutient que « les systèmes d'IA doivent être conçus de manière à respecter l'État de droit, les droits de l'homme, les valeurs démocratiques et la diversité, et ils doivent comporter des garanties appropriées […] pour assurer une société juste et équitable », ce qui peut rappeler la pyramide déontologique de principes du FEM (OECD 2019, § 2). D'autres principes comme la sécurité, la transparence et la robustesse sont aussi mis de l'avant, en plus de cinq recommandations émises à l'endroit des gouvernements, cohérentes avec les cinq principes précédemment énoncés.

Alors que cette charte éthique penche clairement plus du côté du pluralisme des valeurs, la notion de « principes complémentaires » vient brouiller les pistes. Sans le dire explicitement, l'idée que les principes soient entre eux « complémentaires » suggère une forme d'unité et d'harmonie,

conditions sous-jacentes au monisme sur le plan métaéthique. Si les principes avaient été présentés comme « potentiellement contradictoires », en tension les uns avec les autres, il m'apparaît que l'on aurait pu simplement camper cette directive éthique du côté du pluralisme des valeurs, du moins pour sa forme. Je suggère donc que la directive de l'OCDE présente aussi un caractère métaéthique en tension, bien que son expression soit moins limpide que dans les autres documents consultés. Une chose est sûre, en définitive : la tension métaéthique existe bel et bien dans les directives éthiques concernant l'IA et, si elle se présente sous diverses formes et de manières différentes, elle n'a pas le même degré de clarté dans tous les documents.

Conclusion

Les chapitres quatre et cinq forment la deuxième section de la thèse, qui proposait un portrait pouvant se rattacher à la sociologie de la connaissance. L'esquisse des ancrages métaéthiques de la vingtaine de directives portant sur l'éthique de l'IA met au jour de véritables pots-pourris de traditions éthiques différentes. Dans le secteur privé (neuf documents), le pluralisme des valeurs est l'approche qui domine (cinq directives), suivie de près par les approches tentant de conjuguer des éléments incompatibles (trois directives). Un seul document des compagnies privées pourrait avoir été élaboré selon une approche plutôt moniste, si l'on admet l'hypothèse selon laquelle les principes de Google seraient informés par l'utilitarisme. Le secteur public (onze documents) est divisé entre les groupes émanant de la société civile et des associations multipartenaires d'une part, et les organisations internationales d'autre part. Dans le premier sous-groupe, on retrouve un document à saveur moniste (la Déclaration de Toronto, qui est plutôt déontologique), un document pluraliste (les principes d'Asilomar) et trois documents affichant une forme ou une autre de tension métaéthique (les deux études du IEEE ainsi que la Déclaration de Montréal). Du côté des organisations de gouvernance (ou d'intergouvernance) internationale, ce sont ces approches qui dominent (trois documents), suivies par deux documents pluralistes et une seule démarche moniste, celle de l'UNESCO, qui emprunte principalement à l'éthique déontologique. Ainsi, dans le secteur public, on trouve deux démarches monistes, trois pluralistes et six présentant des tensions métaéthiques. Ce qui en ressort, c'est qu'en tout et partout, sur vingt documents analysés, trois seraient monistes, huit pluralistes, et neuf en tension.

Ce constat est intéressant à plusieurs niveaux. Quand vient le temps de nommer les traditions éthiques existantes, dans le but de réfléchir aux enjeux posés par l'IA, le pluralisme ne fait presque jamais partie du lot. Le pluralisme (des valeurs) n'est en somme presque jamais nommé[34]. Toutefois, force est de constater son omniprésence dans les réflexions éthiques. En effet, il informe chaque document de la catégorie la plus nombreuse, soit la catégorie de la tension métaéthique. On peut évidemment en dire autant du monisme. Le pluralisme des valeurs constitue la seconde catégorie présentant le plus de directives éthiques. On peut alors se demander pourquoi ce dernier n'est pas explicitement reconnu quand il est question de tirer le portrait des traditions éthiques occidentales existantes, étant maintes fois employé dans les démarches elles-mêmes.

Plus encore, des traditions incompatibles au plan métaéthique cohabitent fréquemment dans un même document éthique, et ce, sans que cela soit reconnu de manière explicite. Les conséquences de cette cohabitation ne sont ni anticipées ni approfondies dans l'écriture de ces directives. Ce constat n'est pas anodin, dans le sens où la façon d'aborder les dilemmes ou conflits éthiques diffère de manière importante, selon que l'on adopte une posture métaéthique moniste ou pluraliste. Un éthicien qui adopterait une position mitoyenne devrait d'abord en être conscient, puis expliquer en quoi sa combinaison est viable ou cohérente à l'interne. La raison en est que les deux postures métaéthiques sont, ultimement, en contradiction l'une avec l'autre. La question à se poser, en dernière instance, est de savoir l'effet que cela aura pour des décideurs politiques. Une autre interrogation, celle qui, en somme, guidera la prochaine section est de savoir comment conseiller ces derniers dans un tel contexte.

[34] Une exception doit être mentionnée dans la littérature en éthique de l'IA : Morley et al. (2019) ont cherché à développer des outils pour l'application concrète de cinq principes en éthique de l'IA, sur lesquels il existerait un consensus (selon Floridi et al. 2018a). Ils écrivent que « ces outils et méthodes pourraient être utilisés pour aider les concepteurs à traiter de manière proéthique le pluralisme des valeurs (c'est-à-dire la variation de la valeur entre différents groupes de population) » (15-16). Le pluralisme des valeurs est mentionné, mais il n'est pas certain qu'il s'agisse de la tradition éthique associée à Berlin, selon la définition proposée.

SECTION 3 : PROPOSITION

Chapitre 6 — Monisme non orthodoxe

« *L'innocence est en effet une chose glorieuse; mais le monde réel, étant coupable, a besoin de philosophie*. » — Martha Nussbaum [Traduction libre] (2000, 255)

« *L'auditeur de la leçon d'éthique aristotélicienne devait par lui-même être à l'abri d'un danger : celui de vouloir en rester à la théorie pour échapper à l'exigence de la situation. Aristote ne perd jamais de vue ce danger; c'est en cela qu'à mon sens il n'a pas vieilli*. » — Hans Georg Gadamer (1982, 325)

Introduction

Les deux premiers tiers de la thèse ont été employés à deux choses. Premièrement, cerner les fondations des traditions éthiques auxquelles je fais référence. Deuxièmement, j'ai tenté, sur la base de ces caractéristiques, de relever quelles approches étaient à l'œuvre dans un échantillon de directives portant sur l'éthique de l'intelligence artificielle. En tout et partout, le pluralisme des valeurs informe dix-sept documents sur vingt, tandis que le monisme en influence une douzaine. Le portrait que j'ai tiré de l'éthique en IA est certes non exhaustif. Néanmoins, je le crois assez représentatif, étant donné la portée des documents sélectionnés, ainsi que leur influence. Cette nouvelle section de la thèse offre au lecteur la première partie d'une proposition, qui se présente comme une voie alternative aux options glanées dans la littérature grise. Évidemment, ma proposition comporte certaines ressemblances avec les démarches recensées dans la littérature, mais elle possède tout de même ses spécificités propres.

Ainsi, cette troisième section se veut beaucoup plus normative que la première, qui se rapproche d'un « cadre théorique » ou de la seconde, qui contient « l'analyse des données », dans la terminologie de la méthodologie empirique. Dans ce chapitre et le suivant, je mettrai en lumière ce qui m'apparaît optimal, pour les décideurs politiques, de favoriser comme approche éthique pour dialoguer et légiférer sur les enjeux de l'intelligence artificielle. Ce chapitre sera axé sur l'approche métaéthique que je privilégie, et le prochain sur l'approche au dialogue. Il peut être pertinent de rappeler que ces deux chapitres, comme le veut l'orientation de la thèse, sont écrits avec les

décideurs politiques comme destinataires. Conséquemment, la proposition métaéthique que j'avance ici n'est pas pensée pour des ingénieurs, des concepteurs ou des programmeurs en IA. On l'a vu : plusieurs des directives à l'étude étaient destinées à un auditoire fort diversifié. Dès le premier chapitre, j'ai spécifié qu'à mon sens, une éthique politique pour l'IA englobe en quelque sorte tous les champs de l'éthique de l'intelligence artificielle, de manière plus ou moins directe. Toutefois, les destinataires principaux de ce travail de réflexion sont les élus représentant une communauté politique donnée.

Précisions d'entrée de jeu sur la voie alternative

Je ne vais pas réitérer l'exposé des critiques pouvant être adressées à chacune des traditions éthiques dont il a été question jusqu'ici : il figure à la section des fondements. Plutôt, j'aimerais justifier auprès du lecteur pourquoi les trois approches métaéthiques que l'analyse des directives éthiques a permis de mettre au jour me semblent inadaptées pour guider les décideurs politiques en éthique. Je dirais, comme Taylor, que

> la difficulté qui s'attache à la plupart des conceptions que je considère comme inadéquates, et par rapport auxquelles je tiens ici à définir la mienne, est qu'elles restreignent trop le champ de ce qu'elles reconnaissent. (2003, 626-627)

Concrètement, tout en présentant ce que je retiens de l'éthique de la vertu et du pluralisme des valeurs, j'exposerai au lecteur ce qui me semble « rétrécissant » dans le déontologisme et l'utilitarisme. L'éthique de la vertu et le pluralisme n'échapperont pas à l'examen critique, puisque je me propose de mettre au jour les difficultés qu'ils présentent aussi. La voie alternative que je présente exhibe, je l'espère, une cohérence interne, éprouvée par des philosophes comme Martha Nussbaum et Charles Taylor. En ce sens, elle se différencie des directives éthiques faisant montre d'une tension métaéthique, parce qu'il s'agit d'une démarche éthique à part entière, qui se veut cohérente « à l'interne ». Je l'ai dit au début de la thèse : mon apport à cette discussion sociétale sur l'éthique de l'IA est celui d'une citoyenne de plus qui se penche sur la question. Dans cette optique, je suis tout à fait consciente que ma proposition a un poids différent d'une autre qui serait

le fruit de consultations citoyennes[35] ou d'un exercice interdisciplinaire. Ce que je ne possède pas en diversité de points de vue, j'essaie d'y compenser par une cohérence et une profondeur que le temps passé à travailler la question peut m'avoir fournies.

La proposition que je mets de l'avant est composée de quatre éléments. Le premier élément pourrait, à mon sens, englober les deux suivants. La division est ainsi factice, mais je l'opère tout de même pour des fins de clarté et de précision. La première composante est donc la prudence comme vertu intellectuelle,[36] la deuxième une orientation téléologique « douce » vers un bien commun, la troisième une sensibilité profonde au contexte, et la quatrième la reconnaissance de dilemmes potentiellement insolubles. Ces quatre éléments seront aussi, selon moi, les fondements d'un dialogue politique optimal sur l'éthique. Cependant, une nuance s'impose. Ma proposition ne consiste pas en une « théorie éthique » au sens moniste. Le quatrième élément, en effet, vient contrecarrer l'unification « automatique », voire théorique, du réel dans la confrontation des points de vue éthiques dans le dialogue. Si on devait catégoriser ma démarche, elle tomberait probablement dans le monisme « non orthodoxe » en raison de sa proximité avec des penseurs néo-aristotéliciens comme Martha Nussbaum et Charles Taylor. Cela étant dit, s'il y a parenté intellectuelle avec ces penseurs, cela ne signifie pas que sur tous les enjeux métaéthiques, je partage leur position ni que j'aie entièrement articulé la mienne. Cette thèse est mon premier effort dans ce sens, et le travail ira, je l'espère, au-delà de ces quelques lignes.

Je commencerai ce chapitre en traitant du monisme « non orthodoxe ». Par la suite, j'exposerai ma critique des approches éthiques dont j'ai traité jusqu'à présent (en particulier le déontologisme et l'utilitarisme), dans le but d'illustrer en quoi elles ne me semblent pas adéquates, dans leur entièreté, pour guider le dialogue des décideurs politiques. Mon propos portera tout

[35] Taylor, en réfléchissant sur l'approche herméneutique gadamérienne, soutient que le meilleur moyen de parvenir à l'objectivité sur une question donnée est par une démarche d'inclusion dans la conversation (1995, 152). Cela étant dit, il faut aussi tenir compte du fait que la quantité de personnes consultées pour une directive éthique n'est pas une garantie automatique de représentativité de la population donnée. C'est ce que soutient Paula Boddington au sujet des Principes d'Asilomar du FLI (2017, 105).

[36] J'entends ici la prudence comme vertu et non comme principe. La Déclaration de Montréal liste effectivement un principe de prudence, qu'elle décrit comme la nécessité d'« […] anticip[er] autant que possible les conséquences néfastes de l'utilisation des SIA et en prenant des mesures appropriées pour les éviter. » (Comité d'élaboration de la Déclaration de Montréal IA reponsable 2018, s.p.) On verra dans la suite du chapitre que ma compréhension de la prudence est différente de celle-ci.

spécialement sur la raison théorique, l'éthique de la règle, la raison instrumentale, le pessimisme inhérent au pluralisme des valeurs, puis les angles morts des approches présentant des tensions métaéthiques. Ensuite, j'aborderai un par un les quatre éléments qui forment la base de ma proposition, soit une démarche de sagesse pratique alliant des éléments de l'éthique de la vertu et du pluralisme des valeurs. Je fournirai quelques illustrations portant sur des enjeux concrets que posent actuellement des systèmes employant l'intelligence artificielle. Je terminerai ce chapitre avec une réflexion sur l'acceptabilité de ma proposition et des notions mixtes qu'elle combine. Cette dernière section guidera le lecteur vers la suite du propos, soit le dialogue des décideurs politiques au prochain et dernier chapitre.

1. Le monisme orthodoxe et non orthodoxe

Le monisme « pur » (ou pris dans ses expressions « orthodoxes ») présente quelques problèmes dans la sphère politique. On l'a vu, une des choses qui caractérise une démarche moniste est l'unité (qu'elle soit théorique [sous une forme organique ou systématique] ou non). Devant une diversité de points de vue en éthique politique, une des réactions possibles peut être de chercher à générer un consensus, que ce soit par recoupements ou par équilibre réflexif. Or, un risque attaché à cette poursuite est que la recherche de consensus devienne « absolutisée » et que ce dernier « […] puisse parfois être atteint au détriment de l'abstraction et du choix de mots qui peuvent masquer le désaccord » [Traduction libre] (Boddington 2017, 105). Autrement dit, le fait d'aborder les conflits d'interprétation à partir d'une perspective moniste orthodoxe peut occulter la profondeur des différends. Ignorer l'ampleur des pertes générées par un compromis de valeurs est loin d'être idéal pour un décideur politique, qui conduirait ainsi ses décisions sans suffisamment d'égards à ces angles morts. Comme il s'agit de résolutions possédant une incidence réelle sur une communauté politique concrète, ne pas être conscient des pertes encourues, lorsqu'on est dans une situation de responsabilité, pourrait aller jusqu'à s'apparenter à un aveuglement volontaire, dans certains cas.

De fait, on a vu qu'un décideur moniste orthodoxe pourrait faire appel au paradigme des droits humains pour approcher un défi éthique posé par un SIA. Ces droits seraient compris comme faisant partie d'un système ou d'un tout. Une telle interprétation se heurte au fait que « […] les

cultures relèvent des valeurs » [Traduction libre] (Boddington 2017, 108). Les conflits entre ces dernières ne peuvent être simplement oblitérés par une façon unique de les comprendre, ou par une seule interprétation culturelle. Le prétendre reviendrait à masquer une partie de la réalité. De plus, les conflits auront bel et bien lieu, qu'ils soient occultés (consciemment ou inconsciemment) ou non. En revanche, il serait injuste de prétendre que dans tous les cas, le monisme orthodoxe peine à tenir compte de la diversité de points de vue. Plutôt, sa faiblesse réside dans sa prétention à leur unification ou harmonisation sans perte morale.

a) Le « pluralisme raisonnable » comme monisme orthodoxe

Il est vrai que certaines approches monistes sont parfaitement compatibles avec ce que l'on appelle le « pluralisme raisonnable ». On retrouve cette notion, par exemple, dans la pensée de John Rawls, qu'Andrew Lister cite en la définissant comme

> une diversité de doctrines religieuses, philosophiques et morales irréconciliables, mais raisonnables [qui] est « le résultat inévitable à long terme des pouvoirs de la raison humaine à l'œuvre dans le cadre d'institutions libres durables ». [Traduction libre] (Rawls 2005, 36, 135, dans Lister 2015, 700)

Ce pluralisme raisonnable ne met pas en danger l'existence d'une société « approximativement juste », même si on reconnaît que, même entre personnes raisonnables, les désaccords existeront (Lister 2015, 700-701). Cela dit, il ne s'agit pas de « […] la thèse du pluralisme des valeurs associée à Isaiah Berlin » [Traduction libre] (Lister 2015, 701). Le pluralisme raisonnable ne traite pas de la question de l'incompatibilité de certaines valeurs, alors qu'elle est centrale pour les pluralistes des valeurs. Il est en somme inévitable que certains types de concepts, émergeant surtout de la philosophie politique ou des arts, soient interprétés de manières *incompatibles* selon le contexte ou les personnes, étant donné les limites de ces derniers. Il pourrait s'agir de « concepts essentiellement contestés », qui « […] bien que ne pouvant être résolus par un quelconque argumentaire, sont néanmoins soutenus par des arguments et des preuves parfaitement respectables » [Traduction libre] (Gallie 1956, 168-169). Ou encore, plusieurs acceptions d'un même concept peuvent être mises de l'avant, en étant « raisonnables » selon des critères inspirés de Rawls, mais qui sont incompatibles dans la pratique.

Ainsi, la différence entre les deux types de pluralisme est que le pluralisme raisonnable est moniste, au plan métaéthique, ce qui n'est de toute évidence pas le cas du pluralisme des valeurs. En effet, le postulat implicite d'une approche éthique moniste est que la résolution de ces désaccords inévitables n'implique pas de dommage moral, que les valeurs en jeu ne sont pas incompatibles dans l'absolu. Plus encore, la solution à ces désaccords existe, elle est réelle. Si on ne la perçoit pas, c'est par une faute que l'on peut attribuer à l'agent, non à la situation, à la réalité ou encore à la théorie. Cette façon de voir, esquissée dans ses grandes lignes ici, est partagée par les tenants de théories éthiques comme le déontologisme ou l'utilitarisme (ou encore l'éthique de la vertu aristotélicienne), qui sont des façons « orthodoxes » d'exprimer l'unité métaéthique. Par exemple, pour le déontologue, il importe, pour bien agir, de suivre de façon autonome l'impératif catégorique. Pour l'utilitariste, la bonne solution est celle qui maximise le bien-être du plus grand nombre. Une fois ces trames d'action choisies, il n'y a pas de raison de se sentir coupable de possibles pertes encourues, même si on peut les déplorer. Moralement parlant, l'agent tiendra pour acquis qu'il a les mains propres. Au plan métaéthique, la réalité est unifiée : si on ne le voit pas, le problème réside en nous, pour le dire très simplement.

b) Le monisme non orthodoxe et ses spécificités

Dans sa réflexion sur le monisme et le pluralisme, Charles Blattberg fait place à une catégorie de penseurs pour qui les choses ne sont pas si « arrêtées ». Ceux que l'on pourrait appeler des « monistes non orthodoxes » sont d'avis que « […] la réalité dans son ensemble n'est pas (encore) unifiée […] »; mais qu'éventuellement, « […] rien, en principe, ne l'empêche de le devenir » [Traduction libre] (Blattberg 2018, 157). Charles Taylor s'y inscrit clairement quand il affirme qu'« entre ceux qui ont décrété sans discussion une unité *a priori*, et ceux qui pensent que toute unité est imposée arbitrairement, il ne reste qu'une frange étroite de penseurs capables ne serait-ce que de discuter du problème » (Taylor 1997b, 162). Autrement dit, il ouvre la porte à une unité *a posteriori*, qui ne s'impose pas de manière arbitraire. Même s'il s'agit d'une position métaéthique difficile à cerner, parce que subtile, elle postule tout de même l'unité, mais de manière différente des monistes orthodoxes.

Certains monistes non orthodoxes affirment que l'espérance de conciliation (ou d'unification, d'harmonisation) est freinée par les limites du monde de l'esprit — les limites de la théorie, par exemple. Ce serait le cas d'un Platon ou d'un Léo Strauss. Inversement, d'autres monistes non orthodoxes soutiendraient que c'est le monde de la pratique (comme les institutions, ou encore les actions des êtres humains qui sont libres) qui met des entraves dans le chemin de l'unification de ce qui est fragmenté. Des philosophes comme Martha Nussbaum et Charles Taylor seraient emblématiques de ce second groupe. Ces derniers, à la différence de Strauss ou de Platon,

> […] reconnaissent les problèmes du monde de la pratique plutôt que ceux de la théorie et, par conséquent, ils considèrent que notre défi ultime est de transformer le premier pour qu'il se conforme au second. Pour eux, *les mains sales sont parfois inévitables,* car, si la réforme complète du monde de la pratique est possible en principe, dans la pratique, il y aura des moments où nous ne pourrons tout simplement pas la gérer. [Traduction libre] (Blattberg 2018, 157-158)

Les deux « tendances » du monisme non orthodoxe ont toutefois en commun qu'un changement du monde est nécessaire, possible et évidemment souhaitable : seulement, le premier groupe soutient que c'est le monde de l'esprit qui doit changer, et le second, le monde des pratiques. Nussbaum est très claire quant à la nécessité de ce changement : nous ne vivons pas dans un salon de discussion d'Oxford, dit-elle, mais bien « […] dans un monde plein de mauvaises théories grossières, de passions égoïstes et de jugements entachés, où les bonnes passions et les bons jugements ont besoin de toute l'aide possible pour l'emporter et même survivre » [Traduction libre] (Nussbaum 2000, 248). Cette conscience aiguë des difficultés que présente le monde pratique rend, à mon sens, Martha Nussbaum plus attachée que Charles Taylor à l'idée d'une *théorie éthique.* Taylor, hormis ses travaux sur le multiculturalisme ou la laïcité (voir Maclure et Taylor 2010), semble placer l'accent sur la reconnaissance de la pluralité, toujours dans un cadre moniste implicite (par exemple, Taylor 1982).

Dans ses écrits plus récents, Martha Nussbaum ne se range pas aux côtés de ceux qui

> […] exhortent à rejeter la théorisation systématique de l'éthique et à rejeter également le but qui était celui de la philosophie des Lumières, une vie sociale fondée en raison. […] [C]eux qui voudraient [lui] coller l'étiquette d'« anti-théorie » commettent tout simplement une erreur […]. (2016, xxxvi)

La philosophe reconnaît néanmoins le risque constant de la corruption, ce qu'elle appelle « la fragilité du bien ». Pour elle, « à l'intérieur de la conception aristotélicienne ou de la conception tragique, ils ne peuvent être exclus » (Nussbaum 2016, 522). L'approche même de l'éthique de la vertu comporte ses risques et « […] il y a une perte de valeur chaque fois que les risques de la vertu proprement humaine sont éliminés » (Nussbaum 2016, 522). En effet, « […] chaque vertu aristotélicienne importante semble inséparable d'un risque de dommage » (Nussbaum 2016, 522). On comprend donc l'équilibre précaire sur lequel repose le point de vue de Nussbaum, qui désire la théorie éthique, tout en reconnaissant la réalité de l'échec moral et de la tragédie. Quant à Taylor, il reconnaît que le monde de la pratique peut présenter

> […] d'authentiques dilemmes, que poursuivre un bien unique jusqu'à ses conséquences ultimes peut mener à la catastrophe non pas parce qu'il ne s'agit pas d'un bien, mais parce qu'il en existe d'autres qui ne peuvent lui être sacrifiés sans dommage. (2003, 627)

Il m'apparaît que ce type de positionnement, soit celui du monisme non orthodoxe d'une Nussbaum ou d'un Taylor, qui fait place à la réalité de la culpabilité et des mains sales dans les conflits de valeurs dans le monde de la pratique, permet un regard éthique plus riche, parce que plus réaliste, sur un enjeu sociétal aussi complexe que l'intelligence artificielle et sa gestion. Il permet aussi de faire place à une grande quantité de personnes qui, face à des approches éthiques différentes, adhèrent à des principes moraux qui sont incompatibles entre eux (Taylor 1982, 134-135), sans toutefois souscrire à l'entièreté d'une théorie éthique.

Il faut néanmoins reconnaître que la prise en compte des insuffisances de la théorie et du monde de la pratique, dans le monisme non orthodoxe, vient ouvrir la porte sur une tension au sein même de ce positionnement sur le continuum métaéthique. Les penseurs monistes non orthodoxes sont conscients que leur tentative d'harmonisation des biens qui sont en jeu dans un contexte donné pourrait échouer, en raison des limites de la théorie ou de la pratique (Blattberg 2018, 157-158), voire des deux. Ce qui fait la différence entre ce positionnement métaéthique et ceux que je viens d'analyser dans la section précédente est la réflexivité. Les penseurs monistes non orthodoxes admettent que l'unification des valeurs pourrait fort bien ne pas fonctionner. Ils ne présentent toutefois pas une prise de position « en tension métaéthique » au même titre que celles du chapitre dernier, puisqu'ils ne prétendent pas à la fois à l'unité et à la fragmentation. La conscience de la possibilité d'un échec à l'harmonisation fait qu'ils n'affirment pas garder les mains propres et sales

en même temps, en suivant la même démarche éthique. La démarche moniste non orthodoxe vise l'unité, mais elle fait place à la possibilité d'un échec partiel ou total. En ce sens donc, elle n'est pas incohérente à l'interne. Si l'on saisit bien cette petite nuance, il serait plus aisé de suivre ma critique des deux théories monistes orthodoxes que sont l'éthique déontologique kantienne et l'utilitarisme, qui est surtout informée par ma proximité intellectuelle avec un penseur moniste non orthodoxe comme Charles Taylor : on le verra dans la troisième sous-section de ce chapitre.

2. Critique des approches éthiques recensées

a) Critique de la raison théorique, de l'éthique de la règle et de la raison instrumentale

1. La raison théorique et l'éthique de la règle

On l'a vu précédemment, l'utilitarisme se veut une théorie éthique « scientifique » (Smart 1997, 68), une « science de la moralité » (Sandel 2016, 57). Ce n'est pas le cas de l'éthique de la vertu, qui est d'origine aristotélicienne et qui, en accord avec la pensée du Stagirite, relève, comme discipline, de la raison pratique, et non de la raison théorique, comme le feraient les sciences naturelles. J'ai déjà brièvement exposé la division des fonctions de la raison (ou de l'intelligence) chez Aristote au chapitre deux. Cette catégorisation des rôles de l'intelligence est très utile, car elle permet de différencier, d'un côté, l'éthique déontologique kantienne et l'utilitarisme puis, de l'autre, non seulement l'éthique de la vertu, mais aussi le pluralisme des valeurs, sur le plan des types de la raison auxquels ces différentes approches font appel. Conséquemment, bien que l'éthique de la vertu soit une approche éthique moniste orthodoxe, elle n'est pas incluse dans ma critique de la raison théorique que j'étaie ici.

Je chercherai à mettre en lumière que la conception de la raison (théorique) qui sous-tend les théories utilitariste et déontologique est mal adaptée à la réalité de l'action éthique. En effet, comme dit Thomas d'Aquin, « l'intellect spéculatif est celui qui n'ordonne pas ce qu'il connaît à l'action, mais seulement à la contemplation de la vérité. Au contraire, *l'intellect pratique ordonne*

à l'action ce qu'il connaît » [je souligne] (Aquin 1928, 1a, question 79, article 11, 254-255). La raison théorique s'applique mieux à l'étude ce qui est immuable et universel, par exemple les sciences naturelles, que ce qui est particulier et changeant, comme l'éthique et la politique. La raison théorique n'enjoint pas à pratiquer la vertu de prudence qui, on le verra un peu plus loin, nécessite la flexibilité permise par la sensibilité au contexte. Elle a sa place dans l'investigation scientifique, en ce qui a trait tout particulièrement au domaine de la nature. Ainsi, peut-être la raison théorique a-t-elle quelque chose à dire sur ce qui relèverait de la nature humaine, tant qu'elle est pertinente à la réflexion éthique. Néanmoins, la démarche éthique politique en tant que telle, de même que sa conception intellectuelle, devrait faire appel à la raison pratique, car c'est elle qui oriente vers l'action et ouvre à la vertu de prudence.

Assurément, une posture moniste non orthodoxe pourrait, en principe, faire une place à la raison théorique de pair avec la raison pratique dans une démarche éthique. Par exemple, une théorie pourrait servir à inspirer l'action, sans pour autant la dicter. C'est lorsque la raison théorique s'enjoint à devenir la rationalité à laquelle on fait appel pour l'entièreté — ou presque — de la démarche éthique qu'elle sort complètement de son « domaine de compétence ». Par exemple, lorsque l'éthique devient une question de suivre une règle ou une procédure, il est patent que la démarche est plaquée sur le réel et beaucoup trop rigide. La sensibilité profonde au contexte est entravée par une série d'obligations à suivre. Certes, un éthicien de la règle pourrait soutenir que même les règles doivent être, dans une certaine mesure, adaptées au contexte. Je répondrais qu'à la base, l'éthique n'est pas affaire de règles élaborées dans l'abstraction et que l'on veut universelles, mais de sagesse pratique. Cette dernière exprime le caractère d'une personne et se développe dans les particularités de l'action.

On se rappelle la conclusion des rédacteurs du second rapport du IEEE, soit que les approches éthiques basées sur des règles seront plus efficaces pour réaliser l'alignement des valeurs entre les machines et les humains, ou pour faire en sorte que des systèmes autonomes et d'IA soient « éthiques » dans leurs actions (IEEE 2017, 215). Autrement dit, le respect d'une procédure ou encore d'obligations particulières peut être la meilleure manière de s'y prendre *sur le plan technique*. Toutefois, comme il est question de penser l'éthique dans la perspective d'un dialogue entre décideurs politiques (qui peut, certes, inclure des experts techniques, mais qui demeure, en

définitive, un dialogue humain), ce serait une erreur de se confiner à une éthique de la règle comme si l'on avait affaire à des systèmes informatiques. Les êtres humains ne sont pas que des agents computationnels : ils sont également des « animaux qui s'auto-interprètent » [Traduction libre] (Taylor 1985, 189). À ce sujet, Charles Taylor explique que

> nous attribuons en effet certains des mêmes termes aux humains et aux machines. Nous parlons des deux comme étant capables de « calcul », ou de « déduction » et ainsi de suite pour une longue liste de termes de performance mentale. Mais l'attribution n'a pas la même force dans les deux cas, car nous ne pouvons pas vraiment attribuer une action à une machine au sens plein du terme. [Traduction libre] (1985, 192)

En raison de la différence importante entre les humains et les machines, l'on ne peut pas tenir pour acquis que, si une éthique de la règle était optimale pour un SIA, elle le serait aussi pour un être humain.

Le déontologisme et l'utilitarisme (mais surtout le déontologisme) consistent en des approches éthiques attachées à des règles, des impératifs, et conséquemment elles sont plus rigides. On peut quand même admettre qu'une telle théorie peut contenir des éléments intéressants à amener dans le dialogue. Un exemple serait, comme je l'ai mentionné un peu plus haut, la centralité de la rationalité de l'être humain comme trait distinctif de ce dernier (par exemple Kant 1848a, 100-109).[37] Une éthique procédurale n'est pas à rejeter comme entièrement mauvaise, mais comme trop rigide et, par conséquent, inadéquate au contexte dynamique et changeant de l'action éthique ou politique. À l'inverse, ce contexte dont les particularités sont mouvantes peut être perçu, dans une certaine mesure, comme une entrave aux théories éthiques procédurales. Pour un déontologue, l'agent véritablement éthique est justement celui qui ne se laissera pas détourner de son devoir par des considérations sentimentales émanant de la situation particulière, par exemple, qui peuvent en rendre l'exercice plus ardu.

Une autre difficulté avec les théories éthiques procédurales réside dans ma conviction en l'impossibilité d'articuler l'entièreté du réel en propositions abstraites (Taylor 1997c, 168). Pour Taylor, qui s'appuie sur les travaux de penseurs comme Wittgenstein et Bourdieu, « […] une grande partie de notre action intelligente dans le monde, aussi sensible soit-elle à notre situation et

[37] Cependant, je ne crois pas que les êtres humains soient des « êtres purement rationnels » comme le suggère l'interprétation que fait Taylor de Kant (Taylor 2003, 641).

à nos objectifs, se poursuit sans être formulée. Elle découle d'une compréhension qui est largement inarticulée », bien que non dépourvue de sens [Traduction libre] (Taylor 1997c, 170, 168).[38] Le dialogue des décideurs politiques se devra d'être sensible, non seulement au contexte, mais à la difficulté réelle de l'articuler en premier lieu. La prise de décision, qui implique l'usage de la raison pratique, ne peut être traduite en termes algorithmiques pour la simple raison que cette réduction n'est pas possible (Berlin et Williams 1994, 308). C'est pour ces raisons que je considère l'éthique déontologique et l'utilitarisme comme non optimaux pour le dialogue des décideurs politiques en éthique de l'IA.

Cependant, un lecteur qui serait tenant d'une théorie de la démocratie délibérative, par exemple, pourrait à bon droit m'objecter que le dialogue n'est pas l'apanage seul d'une démarche pratique prudentielle, au sens où je l'ai défini jusqu'à présent. Effectivement, une approche éthique de la règle pourrait préconiser un recours à une « théorie de la conversation » (Blattberg 2009, 27). Cette théorie sera néanmoins régie par des règles qui rendront la pratique du dialogue « […] au moins dans une certaine mesure, détachée du contexte pratique […] » [Traduction libre] (Blattberg 2009, 29). Un problème concret de ce détachement est la distorsion des biens que la théorie systématique entraîne dans la pratique (Blattberg 2009, 29-30). La déformation s'opère quand

> […] nous avons tendance à commencer à formuler notre métathéorie d'un domaine donné avec un modèle déjà formé de raisonnement valable, d'autant plus dogmatique que nous sommes inconscients des autres options. […] *Nous coupons et découpons la réalité de la pensée éthique, dans ce cas, pour l'adapter au lit de Procuste de notre modèle de validation*. Puis, comme la métathéorie et la théorie ne peuvent être isolées l'une de l'autre, la conception déformante commence à façonner notre pensée éthique elle-même [Traduction libre, je souligne] (Taylor 1982, 129).

Taylor observe que deux approches éthiques opèrent de cette façon : l'utilitarisme et le formalisme d'inspiration kantienne. Ces modèles sont réducteurs, puisqu'ils suggèrent

> […] qu'il existe un seul domaine cohérent de la « morale », qu'il y a un seul ensemble de considérations, ou mode de calcul, qui détermine ce que nous devons « moralement » faire. […] Le raisonnement moral équivaut simplement à calculer les conséquences pour le bonheur humain, ou à déterminer l'applicabilité universelle des maximes, ou quelque chose du genre (Taylor 1982, 132).

[38] Taylor suggère que, bien que cette « action intelligente dans le monde » soit « largement inarticulée », elle n'en demeure pas moins « articulable ». Mon avis diffère un peu parce que c'est impossible dans l'absolu.

Parallèlement, une autre difficulté que pose une éthique de la règle réside dans l'attention démesurée qui pourrait être accordée (qualitativement ou quantitativement) au respect de la procédure (Taylor 1993, 348) plutôt qu'au contenu de l'enjeu en question, voire au bien commun concret de la communauté politique en question. Dès lors, une théorie de la conversation — comme une théorique éthique — pourrait tomber dans ces travers, se révéler réductrice et, dans un cas extrême, devenir à sa manière un lit de Procuste.

Comme mentionné plus haut, la façon de voir l'action éthique diffère grandement en éthique de la vertu, malgré le fait qu'il s'agisse aussi d'une approche moniste orthodoxe. Dans cette optique, l'éthique n'est pas une théorie dictant l'action, mais

> elle nous guide plutôt *en améliorant le raisonnement pratique avec lequel nous agissons*. […] [Elle] reconnaît que la vie morale n'est pas statique, mais qu'elle est toujours en évolution. Lorsqu'il s'agit d'élaborer la bonne chose à faire, *nous ne pouvons pas passer le travail à une théorie, aussi excellente soit-elle*, car, contrairement aux théories, nous apprenons toujours, et nous aspirons donc toujours à faire mieux. [Traduction libre, je souligne] (Annas 2004, 73-74)

Une éthique de la règle (par exemple, l'éthique déontologique kantienne) serait incomplète ou inadaptée à des situations éthiques particulières sans la prudence ou sagesse pratique. En effet,

> même de nombreux déontologues soulignent aujourd'hui que *leurs règles d'orientation de l'action ne peuvent être appliquées de manière fiable sans sagesse pratique*, car une application correcte exige une appréciation de la situation — la capacité de reconnaître, dans toute situation particulière, les caractéristiques de celle-ci qui sont moralement saillantes. [Traduction libre, je souligne] (Hursthouse et Pettigrove 2018, s.p.)[39]

Ainsi, ce que l'éthique de la vertu permet, c'est de répondre à l'imprévisibilité des particularités de chaque situation. C'est ce qui fait dire à Vallor que

> […] l'éthique de la vertu convient parfaitement à la gestion de paysages moraux complexes, nouveaux et imprévisibles, exactement le genre de paysage que présentent les technologies émergentes d'aujourd'hui. [Traduction libre] (Vallor 2016, 17)

Effectivement, même nos « […] attentes changent en même temps que le paysage technologique » [Traduction libre] (Selinger 2019, s.p.).

[39] Notons néanmoins qu'un aristotélicien verrait dans ces règles un obstacle à l'exercice de la *phronesis*.

Il apparaît à présent clairement que dans les sciences sociales, lorsqu'arrive la question de guider l'action, la rationalité des règles n'est pas suffisante : il faut y ajouter les références au contexte, de même que la prise en compte de l'intuition, du jugement et de l'expérience comme connaissance pratique (Flyvbjerg 2001, 24). Le fait de suivre des règles éthiques pour atteindre une finalité, peu importe le degré de réflexion sur ces fins et ces moyens, ne parvient pas à fournir l'environnement dans lequel la vertu de prudence peut s'épanouir. Elle ne peut être codifiée en règles. De fait,

> […] la réponse à la question pratique de savoir ce qu'il faut faire dans des circonstances particulières *ne peut jamais, pour Aristote, être entièrement codifiée dans le discours ou l'écriture humaine sous la forme d'une série de règles abstraites spécifiées antérieurement* — il y a toujours un reste qui n'est pas saisi dans ou par des *logos* abstraits. […] La philosophie pratique, qu'elle soit éthique ou politique, ne peut jamais être entièrement codifiée dans le langage comme une série de principes pratiques généraux précisés à l'avance. [Traduction libre, je souligne] (Abizadeh 2002, 268-269)

Somme toute, on pourrait résumer le propos en suggérant que « ce qui rend une action bonne », « […] c'est le fait qu'elle prenne en compte cela qui est » (Spaemann 1999, 114). J'espère avoir bien expliqué au lecteur en quoi la raison théorique et une éthique de la règle ne sont pas des avenues éthiques à emprunter pour un dialogue optimal des élus politiques. Il importe à présent de se pencher un peu plus sur la rationalité instrumentale, ainsi que les raisons pour lesquelles elle est également inadéquate pour le raisonnement éthique.

2. La raison instrumentale

Contrairement à la raison théorique, la rationalité instrumentale n'existe pas dans la typologie de l'intelligence que propose Aristote. Elle découle d'une acception éminemment moderne de l'intelligence, qui sépare la fin et les moyens, chose véritablement impensable pour le Stagirite.[40] La raison instrumentale peut être comprise comme relevant de la fonction productive

[40] On peut soutenir cela à la lumière du fait qu'Aristote, par exemple, ne sépare pas la main de l'outil. Plutôt, il les conçoit comme une seule et même chose : « *Car la main est un outil*; or la nature attribue toujours, comme le ferait un homme sage, chaque organe à qui est capable de s'en servir. […] En effet, l'être le plus intelligent est celui qui est capable de bien utiliser le plus grand nombre d'outils : or, la main semble bien être non pas un outil, mais plusieurs. Car elle est pour ainsi dire un outil qui tient lieu des autres. C'est donc à l'être capable d'acquérir le plus grand nombre de techniques que la nature a donné l'outil de loin le plus utile, la main. […] Car la main devient griffe, serre, corne,

de l'intelligence, soit celle à laquelle est associée la vertu de *techne*. Simplement, elle consiste en « […] cette rationalité que nous utilisons lorsque nous évaluons les moyens les plus simples de parvenir à une fin donnée. L'efficacité maximale, la plus grande productivité, mesure sa réussite » (Taylor 1992, 15). Il s'agit d'une forme de la raison qui n'est pas définie de manière substantielle, mais plutôt formelle, « […] en termes de procédures que l'on pense devoir suivre, notamment celles qui consistent en la mise en place de moyens pour atteindre des objectifs […] » [Traduction libre] (Taylor 1993, 341).

Le pluraliste des valeurs Max Weber a distingué deux types de rationalité dans la modernité, soit la rationalité en valeur et la rationalité en finalité (un autre nom pour la rationalité instrumentale). Elles sont en opposition l'une avec l'autre, parce que l'on peut dire, au sujet de la première, que l'action rationnelle est motivée non par ses « conséquences prévisibles », mais par ce qui crée une obligation comme « […] le devoir, la dignité, la beauté, les directives religieuses, la piété ou la grandeur d'une "cause", qu'elle qu'en soit la nature » (Weber 2013, 2). On pourrait de prime abord penser que l'impératif catégorique de Kant se situerait dans la rationalité en valeur, étant donné que ce qui motive l'agent kantien est le devoir fait pour lui-même. Si cela est vrai, il faut toutefois reconnaître que la rationalité en valeur fait aussi place à des raisons d'agir que Kant rejette, parce qu'hétérodoxes (« les directives religieuses », par exemple).

Par contraste, la rationalité en finalité — ou la rationalité instrumentale — caractérise

> […] celui qui oriente son activité d'après les fins, moyens et conséquences subsidiaires [*Nebenfolge*] et qui *confronte* en même temps rationnellement les moyens et la fin, la fin et les conséquences subsidiaires et enfin les diverses fins possibles entre elles. (Weber 2013, 3)

La rationalité instrumentale caractérise des approches éthiques conséquentialistes; on pourrait aussi voir son lien avec les impératifs hypothétiques de Kant (Kolodny et Brunero 2020, s.p.). Comme forme de la raison technique, au sens moderne, elle caractérise l'utilitarisme qui

> transforme les normes morales en normes techniques. Car, d'après lui, on ne peut pas discerner la qualité morale des actions à partir des actions elles-mêmes; on a besoin pour

ou lance, ou épée, ou toute autre arme ou outil. *Elle peut être tout cela, parce qu'elle est capable de tout saisir et de tout tenir.* […] Il est possible de s'en servir comme d'un organe unique, double ou multiple » [je souligne] (Aristote s.d.a, 10, 687b).

> cela d'une fonction universelle d'utilité, et la constitution de cette dernière est affaire d'experts, fussent-ils autoproclamés. (Spaemann 1999, 82)

Or, Flyvbjerg explique que la *phronesis* vient équilibrer les deux types de rationalité proposés par Weber (2001, 3). Bien plus : cette vertu n'équivaut à rien de moins que la disposition de la sagesse politique, pour Aristote.[41] Cela étant dit, il faut spécifier que, pour mon propos, ce n'est pas la rationalité instrumentale en tant que telle qui pose problème, pas plus que la fonction théorique de l'intelligence dont il a été question plus haut. Ce ne sont pas « […] les règles, la logique, les signes et la rationalité en eux-mêmes […] » qui portent à confusion, mais leur domination dans des sphères qui appellent plutôt la fonction pratique de l'intelligence, en raison de leur orientation vers l'action [Traduction libre] (Flyvbgerg 2001, 49). En bref, dans la perspective « phronétique » qui est celle de Flyvbjerg, « […] un développement social et politique basé sur la seule rationalité instrumentale n'est pas viable » [Traduction libre] (Flyvbjerg 2001, 53). Je partage ce point de vue et j'expliquerai pourquoi, un peu plus loin, la raison prudentielle est mieux adaptée à une démarche éthique.

b) Critique du portrait pessimiste du pluralisme des valeurs

Bien que toutes les approches éthiques puissent être en mesure de reconnaître des dilemmes moraux, aucune ne met autant l'accent sur leur inévitabilité et le sérieux de leur caractère que le pluralisme des valeurs.[42] Il me semble donc compréhensible qu'il soit aussi présent dans les documents éthiques étudiés. C'est probablement une raison qui explique que je retienne certains aspects du pluralisme des valeurs dans ma proposition éthique alternative. Les pluralistes des valeurs, avec leur insistance sur l'inévitabilité des compromis issus de la rencontre des valeurs, présentent un apport intéressant à l'éthique. Dans une telle optique, face à l'IA,

> le *défi politique* consiste à savoir *comment gérer ces compromis*, soit en concevant des systèmes technosociaux qui maximisent d'une manière ou d'une autre les valeurs pour tous, soit en adoptant un compromis particulier d'une manière que la société est prête à reconnaître comme valable. [Traduction libre, je souligne] (Calo 2017, 12)

[41] Il spécifie tout de même que « la sagesse politique et la prudence sont une seule et même disposition, bien que leur essence ne soit cependant pas la même. » (Aristote 2014, Livre VI, 8, 1141b24-25)

[42] Le sérieux étant tel que, parfois, il peut être quelque peu exagéré en versant dans la tragédie, comme je l'ai mentionné au chapitre trois.

Il est donc vrai qu'il importe de reconnaître l'insolubilité de certains dilemmes.

Toutefois, comme le dit Taylor à Berlin, « je suppose que là où je ne suis pas d'accord, c'est que je suis réticent à prendre cela comme le dernier mot » [Traduction libre] (1994c, 214). Certes, les décideurs politiques doivent être prêts à ce que leur démarche dialogique prudentielle aboutisse à un « échec » au moins partiel, dans la tentative d'intégrer la multiplicité des valeurs et points de vue sur une question. C'est pour cette raison que je pense que les aspects monistes de l'éthique de la vertu, pour être viables politiquement, doivent être tempérés par la reconnaissance de l'insolubilité de certains dilemmes. Il s'agit là, on l'a vu, d'une articulation d'une position moniste non orthodoxe.

En revanche, si cette possibilité de l'échec est réelle, il serait erroné de la supposer inévitable à tous les coups. De fait, « nous sommes souvent simplement victimes d'une illusion d'optique. Les différences nous impressionnent davantage, parce que les points communs nous semblent évidents » (Spaemann 1999, 15) : d'où l'importance du dialogue et de la mise en commun. Estimer ce processus soldé d'avance équivaudrait à fermer la porte à « […] un mode de vie […] dans lequel ces exigences pourraient être réconciliées » [Traduction libre] (Taylor 1994c, 214).

Malgré ses forces, le pluralisme des valeurs présente certaines difficultés pour penser le dialogue éthique des décideurs politiques face à l'IA. Il a été dit qu'il comporte la faiblesse de présupposer le pire alors qu'une issue favorable à la situation de dilemme est possible. D'emblée, une approche pluraliste à l'éthique suppose que les intérêts sont fragmentés de manière irrévocable. Dans cette perspective, est difficile de penser un bien commun sans possibilité de tragédie. En effet, la démarche éthique divise, de prime abord, les intérêts des différentes « parties prenantes ». Il faut spécifier que tous les conflits de ces parties prenantes n'ont pas forcément la même gravité ni la même portée.

Néanmoins, même s'il présente une vision lucide de la réalité des tensions entre les valeurs ou les principes que les agents ont à cœur, le pluralisme des valeurs comporte la faiblesse

d'exagérer ces différends. Ainsi, « les pluralistes tiennent pour acquis qu'il n'y a pas d'unité, en réalité ou à espérer, dans ce monde ni dans quelque autre à venir » [Traduction libre] (Blattberg 2018, 158). Le pessimisme pluraliste réside dans le fait que non seulement ces compromis sont inévitables, mais qu'ils sont « sales » : autrement dit, les effectuer, même de bonne foi, implique une faute morale qui diffère de degré selon la situation (Blattberg 2018, 159). Malgré la polysémie du terme « pluralisme » (que j'entends ici comme le pluralisme des valeurs à proprement parler), je tends à penser, comme Ess, que le recours au pluralisme est « nécessaire, mais non suffisant » (2020, 553, 566) pour approcher des dilemmes éthiques par l'entremise du dialogue.

On l'a vu précédemment, pour des penseurs comme Martha Nussbaum, la situation de dilemme éthique n'a pas forcément à être si tragique. La principale faiblesse des pluralistes des valeurs est qu'« […] en ne visant qu'à éviter les pentes savonneuses, ils s'assurent que nos vies finiront par être plus sales qu'elles doivent l'être » [Traduction libre] (Blattberg 2018, 164).[43] Charles Taylor abonde dans ce sens en affirmant que « le dilemme de la mutilation est, en un sens, notre plus grand défi spirituel, *pas un destin de fer* » [je souligne] (2003, 649-650).

Par ailleurs, le pluralisme des valeurs, on l'a vu, aurait du mal à accepter de parler de vertu de prudence et de téléologie, même « douce », alors qu'il s'agit, à mon avis, d'un élément indispensable à la réflexion éthique politique. L'orientation de la communauté politique vers un bien commun, compris de manière partagée et intimement lié, dans son contenu, à son contexte d'émergence, n'est pas explicitement mentionnée dans les thèses pluralistes. La vertu de prudence, sans pouvoir être imposée aux décideurs politiques, confère une excellence au raisonnement pratique (Aquin 2008, n^{o} 1124). Il s'agit d'une autre faiblesse de l'approche pluraliste, puisque cette orientation téléologique est difficilement évitable dans un raisonnement éthique prudentiel faisant appel à la raison pratique.

Au fil de mon analyse des directives éthiques, j'ai constaté que le pluralisme des valeurs peut contribuer à des déclarations de généralité quelque peu banales : par exemple, sur des déclarations de valeurs et de principes peu contestés dans une grande partie de la population. En

[43] La traduction française de *slippery slope* amène involontairement un amusant jeu de mots.

effet, il est difficile — quoique possible — d'être en désaccord avec la promotion de plusieurs de ces principes éthiques visant l'IA (Boddington 2017, 106), comme la sécurité, le respect de la vie privée, la transparence ou encore la responsabilité, qui sont parmi les plus récurrents. L'avantage de ce constat est qu'il permet aux agents éthiques de reconnaître que les sources de discorde proviennent plus souvent des moyens pour atteindre des fins, que des finalités elles-mêmes. Conséquemment,

> le vrai problème que la politique doit résoudre n'est pas la fin, mais les moyens spécifiques par lesquels ces questions sensibles peuvent être résolues, avec les ressources disponibles, et en tenant compte des conditions réelles dans lesquelles nous nous trouvons. [Traduction libre] (Luño s.d., 11)

c) Critique des angles morts des directives en tension métaéthique

La tâche de proposer une approche éthique aux décideurs politiques pour faire face aux enjeux que soulèvent des technologies d'intelligence artificielle est de toute évidence très ardue. Cela dit, la prise en compte de la complexité de l'entreprise ne justifie pas un recours irréfléchi à un amalgame d'approches éthiques potentiellement contradictoires entre elles, sans expliquer en quoi certains aspects peuvent (ou non) être combinés. Or, c'est ce que l'on retrouve en majorité dans les documents éthiques étudiés. Je l'ai expliqué au chapitre cinq : c'est l'inconscience avérée de ces mélanges éthiques et de leurs conséquences pour les décideurs politiques qui pose problème dans de telles démarches.

Par exemple, concevoir l'éthique comme devant suivre une procédure d'inspiration kantienne, tout en admettant que les intérêts soient tous fragmentés d'avance, peut laisser les décideurs politiques perplexes quant à la réalité éthique, et aussi face aux conséquences de leurs décisions sur le plan moral. Sont-ils en train d'adopter des compromis qui impliquent des pertes — suggérant qu'une forme de remords accompagnerait normalement de telles décisions — ou bien devraient-ils s'accommoder d'un regret, sans plus, puisque ce qui semble une perte ne leur salit pas les mains?

Si une proposition éthique devait être développée dans le but de guider le dialogue des décideurs politiques par rapport aux enjeux que pose l'intelligence artificielle dans la vie des citoyens de la communauté politique, une certaine transparence des présupposés métaéthiques serait nécessaire. Dans l'éventualité où des penseurs feraient appel à plus d'une tradition éthique, il leur incomberait de le mentionner pour en comprendre les implications. À titre d'illustration, dans leur ouvrage sur l'ingénierie sociale et humaine que présentent certaines innovations en IA actuellement, Brett M. Frischmann et Evan Selinger exposent clairement les fondements métaéthiques de leur argumentaire :

> la perspective normative que nous avons avancée est *conséquentialiste* dans le sens où nous pensons que l'ingénierie technosociale doit être évaluée en fonction de ses conséquences pour les êtres humains et l'humanité. D'autres types de conséquences sont pertinentes pour notre évaluation. Notre base normative est *pluraliste, multidimensionnelle et axée sur les capacités d'épanouissement de l'être humain* [*human flourishing*]. Nous mettons l'accent sur les *capacités humaines* spécifiques et reconnaissons l'importance des autres, et nous souligner *la pertinence du bonheur et des conceptions plus larges du bien-être*. Ainsi, nous rejetons « l'idée qu'il existe *une seule valeur morale fondamentale irréductible* à laquelle toutes les autres valeurs morales peuvent être réduites ». [Traduction libre, je souligne] (Frischmann et Selinger 2018, 272)

Cette prise de position exhibe une tension métaéthique non équivoque (conséquentialisme, éthique de la vertu et pluralisme des valeurs),[44] qui est entièrement assumée. À savoir si elle présente des contradictions internes, cela est une autre question, mais je suis d'avis que oui. Les auteurs n'ont pas explicité comment réconcilier les aspirations au bien-être et à l'épanouissement humain, qui sont des notions tirées d'un cadre moniste, avec le pluralisme dont ils affirment s'inspirer. C'est ce qui, selon mon point de vue, fait défaut dans leur argumentaire.

Une directive éthique destinée à un gouvernement donné, peu importe sa provenance (du milieu industriel et du secteur privé, ou public, ou encore des instances de gouvernance internationale), devrait exposer, de la même façon, son positionnement métaéthique. Si ce dernier

[44] Les auteurs affirment s'inscrire dans l'approche des capabilités développée par Martha Nussbaum et Amartya Sen, entre autres, et mentionnent l'épanouissement humain « en des termes aristotéliciens » (Frischmann et Selinger 2018, 272). Ils affirment pencher davantage vers l'approche de Sen, qui est « […] résolument ouverte et pluraliste, en admettant que les différentes communautés et cultures peuvent valoriser et prioriser des capacités différentes » (*Ibid*., p.273). L'approche de Nussbaum, on le sait, tend plutôt à reconnaître une même liste de « circonstances constitutives de l'être humain » pour l'entièreté du monde, qui sont en réalité des limites et des capabilités, et une liste de dix « capabilités humaines fonctionnelles de base » (Nussbaum 1990, 219-224, 225).

recèle des tensions, il reviendrait aux rédacteurs de cette directive d'expliquer comment ces éléments cohabitent dans une même approche cohérente, ou de proposer une manière de faire face aux dilemmes éthiques de même qu'aux conflits de valeurs. À tout le moins, il serait important de mentionner la possibilité qu'ils surviennent. Autrement, les décideurs politiques ne pourraient savoir, d'entrée de jeu, s'il faut attendre ou non de se salir les mains.

C'est ainsi que j'ai présenté, dans la première partie de ce chapitre, les critiques que j'adresse à des aspects très concrets des approches éthiques que j'ai analysées dans la thèse. La place trop grande accordée à la raison théorique, la compréhension de l'éthique comme la nécessité de suivre une règle ou de maximiser son efficacité grâce à une logique instrumentale de la raison, sont des tendances déformantes du déontologisme et de l'utilitarisme. Parallèlement, l'exagération des différends que l'on retrouve dans le pluralisme des valeurs peut saper la volonté de réconciliation si l'accent, dans le dialogue, est placé sur les différences apparemment ou potentiellement irréconciliables des interlocuteurs. De même, des démarches éthiques mises de l'avant, qui combinent inconsciemment des composantes monistes et pluralistes, peuvent générer de la confusion auprès des décideurs politiques. L'importance que j'ai accordée à la raison pratique, dans ma critique des autres traditions éthiques, ainsi qu'à sa vertu, la prudence, est au cœur de ma « contre-proposition » pour l'éthique de l'IA. La seconde partie de ce chapitre est donc dédiée à mettre en lumière les aspects de la voie alternative que je propose aux élus.

3. Les éléments de ma proposition

a) La prudence comme vertu intellectuelle

Je parle de la « prudence », dans ma proposition, comme une vertu intellectuelle centrale à l'exercice de la raison pratique. Cela est conforme à la tradition aristotélicienne. Néanmoins, je ne l'appelle pas « *phronesis* », puisque je préfère réserver ce mot pour l'acception entièrement aristotélicienne du terme, c'est-à-dire, comme on l'a vu au chapitre deux, orientée par la nature vers la cible du « Souverain Bien » (Aristote 2014, Livre I, 1, 1094a20-25 — 1094b-1095b5) qu'est l'*eudaimonia*. Je ne crois pas qu'en éthique politique, la finalité de tout raisonnement pratique soit

explicitée d'avance de façon exhaustive, pas plus que je ne crois à une harmonie parfaite des biens concurrents vers ce « Souverain Bien ». En effet, il arrive souvent que « le monde de la pratique » ne le permette pas, comme le suggèrent les monistes non orthodoxes que sont Nussbaum et Taylor. C'est la raison pour laquelle je me dissocie partiellement de la compréhension aristotélicienne de cette vertu, qui relèverait plutôt du monisme orthodoxe.

Ma réflexion à ce sujet n'est pas encore entièrement aboutie, mais je peux dire que si la finalité de la prudence — qui est le bien commun, comme on le verra plus bas — n'est pas entièrement déterminée par la nature, elle a tout de même quelque chose à voir avec ce que tous les êtres humains ont en commun *qua* humains. En ce sens, j'accepterais la notion de rationalité (ou de *logos*) comme étant un trait distinctement humain, une idée que l'on retrouve non seulement dans la conception aristotélicienne de l'humain, mais aussi notamment dans l'éthique kantienne. D'ailleurs, Aristote dit que l'homme est le seul, parmi les animaux, à posséder un langage, qui « […] existe en vue de manifester l'avantageux et le nuisible, et par suite aussi le juste et l'injuste » (Aristote 1990, Livre 1, 2, 1253a10). Ce genre de conception de la nature humaine m'apparaît comme un socle du commun acceptable pour le moment, bien que, comme on le verra, les modalités et les « couleurs » que peut prendre le bien commun dans chaque contexte politique sont variables. J'y reviendrai.

Pour comprendre comment fonctionnerait un concept remanié de la *phronesis* dans une réflexion s'inscrivant dans les sciences sociales, on peut se tourner vers les travaux de Bent Flyvbjerg. Ce dernier a repris le concept pour développer une « science sociale "phronétique" » (Flyvbjerg, Landman et Schram 2012, 285). Insatisfait avec l'application aux sciences sociales de méthodologies exportées du domaine des sciences de la nature, Flyvbjerg a voulu réhabiliter la prudence aristotélicienne, en la modifiant quelque peu. Il y ajoute quelques postulats de Michel Foucault et de ses réflexions sur le pouvoir, alors que ni Aristote, ni Hans Georg Gadamer, ni aucun autre philosophe ne l'avaient fait dans leurs compréhensions de la *phronesis*. Selon Flyvbjerg, les phénomènes sociaux sont caractérisés par le conflit et par le pouvoir (Flybjerg 2001, 3, 55). Ce qui m'intéresse de son argumentaire se trouve moins dans cette idée du pouvoir que dans sa critique du positivisme dans les sciences sociales, et sa volonté de ramener la *phronesis*. Au fond, cette

apologie revient à replacer les sciences sociales comme la science politique (ou des enjeux éthiques à portée politique) dans le contexte de l'exercice de la raison pratique.

Ainsi, pour Flyvbjerg, la prudence

> […] va au-delà des connaissances analytiques et scientifiques (*episteme*) et des connaissances ou savoir-faire techniques (*techne*). Elle implique des jugements et des décisions prises à la manière d'un acteur social et politique virtuose. […] la *phronesis* est couramment utilisée dans la pratique sociale et que, par conséquent, les tentatives de réduire la science et la théorie sociales soit à l'*episteme*, soit à la *techne*, ou de les comprendre en ces termes, sont erronées. [Traduction libre] (Flybjerg 2001, 2)

Autrement dit, « "[…] la raison se comporte autrement dans le domaine du technique et autrement dans le domaine du moral" » (Aristote dans Spaemann 1997, 7). L'idée de comparer les sciences sociales aux sciences naturelles est problématique puisque là où sont les forces de l'une, se trouvent aussi les faiblesses de l'autre (Flyvbjerg 2001, 3). Flyvbjerg appuie son argumentaire, entre autres, sur la pensée des frères Hubert et Stuart Dreyfus dans leur critique du cognitivisme et du naturalisme, ainsi que dans leur phénoménologie de l'acquisition de compétences (Dreyfus et Dreyfus 1988),[45] de même que la phénoménologie de Maurice Merleau-Ponty (2016),[46] pour traiter de l'importance de la saisie du contexte (Flyvbjerg 2001, 4-20). La sagesse pratique est orientée vers le concret, le particulier et l'action. En conséquence, dans l'optique où le but de ma réflexion est d'élaborer un document devant guider le dialogue des décideurs politiques concernant les enjeux éthiques que pose l'intelligence artificielle, il est implicite que ce dialogue débouchera forcément sur l'action. De plus, cette action sera particulière, et non générale ou universelle.

Chaque action possède ses modalités propres, car elle est accomplie dans le temps, l'espace, de même qu'un contexte culturel et politique particulier. C'est pour cela que la raison théorique ou spéculative ne convient pas pour penser l'éthique à partir du point de vue des décideurs politiques. Aristote le dit : « […] que la prudence ne soit pas science, c'est là une chose manifeste : elle porte, en effet, sur ce qu'il y a de plus particulier, comme nous l'avons dit, car l'action à accomplir est

[45] Cette phénoménologie est effectivement ancrée dans une perspective pratique de la raison, mais plus technique (*techne*) que prudentielle (*phronesis*).
[46] On peut aussi y associer les travaux de Ludwig Wittgenstein.

elle-même particulière » (Aristote 2014, Livre VI, 9, 1142a20-25). Ce n'est pas une question de science, dit-il donc, mais de perception (Aristote 2014, Livre VI, 9, 1142a25-30).[47]

La différence entre l'emploi de la raison théorique, instrumentale ou prudentielle n'est pas accidentelle en éthique de l'IA. Un exemple peut illustrer cette assertion. Tessa Sproule, la fondatrice de la plateforme « Vubble », servant à la classification et l'organisation de vidéos mises en ligne, a quitté le réseau anglophone de la Société Radio-Canada (CBC), en raison de son insatisfaction devant la gestion de ses produits. La Société les vendait en effet à Facebook, enrichissant de ce fait sa position déjà prééminente dans le monde de la technologie Web et nourrissant indirectement, soutient-elle, des auditoires en quête de sensationnalisme (Daniels 2018, s.p.). Pour répondre à cette problématique, la compagnie Vubble a élaboré un plan d'action à l'intention de dirigeants d'entreprises et de décideurs, « […] pour s'assurer que les initiatives d'IA génèrent un résultat éthique positif net » [Traduction libre] (Daniels 2018, s.p.). Autrement dit, Sproule soutient que la meilleure manière d'employer les technologies de l'IA, c'est en mettant sur pied une « technocratie »,[48] qui facilitera « […] la production de technologies qui ne nourrissent pas seulement notre activité et celle de nos clients, mais qui font le bien et rendent la société meilleure pour nous tous » [Traduction libre] (Daniels 2018, s.p.).

Un décideur politique à qui l'on présenterait l'approche de Sproule pour réglementer la production de technologies employant l'IA pourrait se demander, concrètement, comment assurer « un résultat éthique positif net ». Cet objectif, à saveur conséquentialiste, comprend le sous-objectif de satisfaire non seulement les producteurs et les consommateurs de la technologie, mais la société dans son ensemble. Rappelons que c'est la raison instrumentale caractérise l'éthique conséquentialiste qu'est l'utilitarisme. Le problème potentiel avec cette approche est que des compromis pourraient être effectués en cours de route (par des décisions de politiciens) pour maximiser une cible le plus souvent quantifiable, que l'on désire généraliser (par les retombées positives pour toute la société). Plus encore, ces accommodements seront perçus comme souhaitables puisque, selon la logique instrumentale à l'œuvre, il n'est pas question de délibérer sur la fin et les moyens en faisant appel à la prudence, mais de maximiser le résultat de la fin en

[47] Cela étant dit, le bon fonctionnement de la raison pratique nécessite, à sa place, la raison théorique également, pour Aristote.
[48] Dans le contexte de l'article, mon interprétation du terme est qu'il signifie une démocratie de la technologie.

choisissant les moyens les plus *efficaces* pour cette dernière. Ainsi, dans le but de maximiser les bénéfices généralisés des SIA, un raisonnement de nature abstraite sur le bien-être de la population en général pourrait être mené et des cibles établies à partir de ces données. Un raisonnement quantitatif pourrait s'ensuivre.

Pourtant, une approche prudentielle, à mon sens, procéderait plutôt par questions. Les décideurs politiques devraient se demander « qui gagne, et qui perd » dans la transaction touchant à la mise en marché, par exemple, d'un SIA et « à quel prix ». Le postulat pluraliste des valeurs pourrait venir enrichir la réflexion en amenant une sorte de lucidité politique : même si l'on tente de maximiser les retombées positives d'une technologie en particulier, il y aura forcément des gagnants et des perdants, ainsi que des intérêts qui entrent en ligne de compte. L'insistance du pluralisme sur l'inévitabilité des conflits peut être d'une grande utilité pour la réflexion éthique politique. Le penseur Neil Postman, qui a étudié la portée des changements technologiques au cours de l'histoire (Postman 1993), affirme :

> […] vous seriez surpris de voir combien de personnes croient que les nouvelles technologies sont des bénédictions non mitigées. […] Pensez à l'automobile qui, malgré tous ses avantages évidents, a empoisonné notre air, étouffé nos villes et dégradé la beauté de notre paysage naturel. […] La culture paie toujours un prix pour la technologie. [Traduction libre] (Postman 1998, 1-2)

Il ne s'agit évidemment pas d'être un « luddite », mais bien lucide face aux conflits inévitables que des avancées technologiques peuvent générer. Au-delà de l'exemple de l'automobile, qui est évidemment daté, la question de la sécurité des données d'une maison connectée peut être soulevée. À plus petite échelle, l'emploi de technologies comme des électroménagers robotiques, ou encore des applications de traçage sur les téléphones mobiles, peut aussi soulever des questions au sujet de la protection de la vie privée, en raison des images et données enregistrées dans ces systèmes, puis agrégées et partagées (Nørskov et Andersen 2016; Astor 2017).

Dans ce contexte, Postman propose d'aborder l'enjeu en cherchant à savoir « qui tire avantage spécifiquement du développement d'une nouvelle technologie? Quels groupes, quel type de personne, quel type d'industrie seront favorisés? » [Traduction libre] (Postman 1998, 2). Ce type de questions favorise la prise en compte du contexte et permettra de guider l'action des décideurs politiques en vue du bien commun, tout autrement qu'un objectif abstrait de

maximisation par l'entremise de l'efficacité. La raison est que les questions que pose Postman visent à prendre les personnes qui forment la société dans leur contexte, et non en abstraction de ce dernier. La raison théorique peut penser dans l'abstraction; la raison instrumentale peut faire fi des pertes dans l'obtention du but auquel sont voués les moyens. Quant à la prudence, c'est par l'entremise d'un dialogue vers un bien commun, qui tient compte du contexte, qu'elle jouera son rôle de manière optimale.

Dans la section traitant des fondements de l'éthique pluraliste, il a été question que la différence entre les tenants de l'éthique de la vertu et les pluralistes des valeurs, quant à la sagesse pratique, réside dans la cible vers laquelle l'archer oriente sa flèche. Conséquemment, puisque l'exercice de la raison pratique peut se servir de cette analogie de l'archer que propose Aristote (2014, Livre I, 1, 1094a20-25 — 1094b-1095b5), il est difficile de penser la raison pratique sans une forme ou une autre d'orientation vers une finalité. C'est la raison pour laquelle une forme de téléologie, au sens de « finalité », m'apparaît incontournable.

b) Une orientation téléologique « douce » vers un bien commun

Je l'ai mentionné dans le portrait des démarches éthiques visant à entourer le développement de l'intelligence artificielle : il est clair que toute démarche éthique se dirige vers un bien. Aristote affirme que « […] s'il y a quelque chose qui soit [la] fin de tous nos actes, c'est cette chose-là qui sera le bien réalisable, et s'il y a plusieurs choses, ce seront ces choses-là » (2014, Livre I, 5, 1097a15-20). Comme il est ici question des décideurs politiques et donc d'éthique plus largement *politique*, le bien dont il est question serait, en définitive, le bien commun. C'est ce que soutient Aristote :

> puisque toute cité, nous le voyons, est une certaine communauté, et que *toute communauté a été constituée en vue d'un certain bien* (car c'est en vue de ce qui leur semble un bien que tous les hommes font ce qu'ils font), *il est clair que toutes [les communautés] visent un certain bien*, et que, avant tout, c'est le bien suprême entre tous que [vise] celle qui est la plus éminente de toutes et qui contient toutes les autres. Or c'est celle qu'on appelle la cité, c'est-à-dire la communauté politique. (1990, Livre I, 1, 1252a1)

Il me semble donc qu'une démarche éthique suivant la rationalité pratique sera nécessairement « téléologique » : cela dit, il s'agira d'une téléologie « douce » plutôt que « forte ». Cette façon de voir apparaît dans les travaux de Charles Taylor, pour qui la vie éthique que nous menons vise une certaine unité dans la diversité des biens :

> […] dans la mesure où *nous avons* un certain sens de l'unité de notre vie, de ce que nous nous efforçons de « mener », nous rapporterons les différents biens que nous recherchons, non seulement à leurs degrés d'importance respectifs, mais aussi à la façon dont ils s'accordent entre eux, ou ne parviennent pas à s'accorder, dans le développement de notre vie. [Je souligne] (1997b, 168)

Ainsi, pour Taylor, le sens de l'unité est quelque chose que nous possédons déjà. L'unité est à son sens possible, contrairement à ce que croit un penseur pluraliste comme Berlin. Elle est atteignable. Cet élément de ma pensée, que je partage avec Taylor, se rattache au monisme plutôt qu'au pluralisme, et permet plus facilement d'intégrer la compréhension d'une certaine téléologie.

J'ai expliqué au début de la thèse que l'éthique peut être un champ d'étude politique lorsqu'il est question des actions des individus « réunis en une communauté politique » ou de la communauté politique à proprement parler (Luño s.d., 1-2). Par conséquent, l'évaluation éthique de ces décisions se fait selon le référent de « […] la fin que les individus organisés en société se sont donnée » [Traduction libre] (Luño s.d., 2), soit le bien commun politique. Le contenu de ce bien commun n'est pas complètement déterminé d'avance ou de manière théorique, contrairement à ce que soutient Aristote lorsqu'il parle du « bien suprême » ou du « Souverain Bien ».[49] De fait, dans certaines interprétations de la pensée aristotélicienne, on rappelle qu'« […] il y a identité entre le bien de l'individu et celui de la cité […] » (Aristote 2014, Livre I, 1 1094 b10). C'est que « […] les législateurs rendent bons les citoyens en leur faisant contracter certaines habitudes […] » (Aristote 2014, Livre II, 1, 1103b). D'une certaine façon, la politique est, pour Aristote, l'occasion de faire pratiquer aux citoyens des vertus particulières.

[49] Je suggérerais que la définition du *principe* du bien commun, c'est-à-dire le bien commun compris de manière générale, relèverait de la raison théorique, tandis que la définition du *contenu* du bien commun d'une communauté politique donnée, dans un contexte particulier, relève de la raison pratique. C'est la raison pour laquelle la prudence est nécessaire pour l'élaboration de la cible ainsi que l'orientation de la flèche vers cette dernière.

Comme je l'ai expliqué au chapitre deux, l'identification complète qu'opère Aristote entre le bien de l'individu et le bien de la Cité entre en tension avec la notion (moderne) de la liberté individuelle. À coup sûr, ces deux notions sont liées et interdépendantes, mais elles ne sont pas identiques. Il faut que le bien commun permette l'exercice de la liberté individuelle. C'est pour cette raison que je préfère tracer cette première distinction entre éthique personnelle et éthique politique, qui est différenciée selon le sujet agissant. Puis, je suis d'avis, avec les pluralistes des valeurs, que plusieurs modalités du bien commun doivent souvent faire l'objet de délibérations, de dialogue, par les décideurs politiques eux-mêmes, étant donné les données changeantes des situations et la multiplicité de points de vue valables[50] sur une même question (voir par exemple Gadamer 1982, 323). Ce dialogue consiste en un exercice de la rationalité pratique des personnes impliquées. Le bien commun n'est pas purement le résultat d'un consensus, mais plutôt l'aboutissement d'une démarche mobilisant la sagesse pratique. Dans sa lecture de l'éthique aristotélicienne, Gadamer entend la détermination de la finalité comme un choix de ce qui est faisable et qui dépend de notre être. Plus largement, explique-t-il, « notre activité a pour horizon la Polis et notre choix du faisable s'élargit au point de prendre place dans la totalité de notre être extérieur et social » (Gadamer 1982, 324).

Au fond, une certaine cohabitation entre l'objectif du bien commun et la reconnaissance de la multiplicité de points de vue moraux est possible. Selon Taylor, elle est déjà en œuvre dans le monde occidental dans une certaine mesure, puisque « […] nous sommes beaucoup plus "aristotéliciens" que nous l'admettons […] » [Traduction libre] (Taylor 1994b, 22). Autrement formulé, l'idéal moderne de la liberté individuelle n'a pas complètement étouffé la notion de bien commun, et les deux sont dans une tension presque permanente, même si elle n'est pas explicitée. La cohabitation entre la liberté personnelle et l'éthique de la vertu est possible si on conçoit l'éthique comme émergeant des agents, et non comme quelque chose qui est imposé de l'extérieur, comme une règle ou une procédure à suivre, voire une théorie à laquelle conformer nos actions (Luño s.d., 5). Par conséquent, l'éthique peut se « proposer », mais jamais « s'imposer » (Luño s.d., 6).

[50] C'est-à-dire tout point de vue qui ne fait pas l'apologie de choses comme la violence ou la persécution d'un groupe, par exemple.

Dans la perspective métaéthique à laquelle je souscris, soit une posture moniste non orthodoxe, le bien commun est à la fois indécomposable quantitativement et irréductible à un autre qualitativement. Il s'agit d'un bien partagé, mais non par des « parties prenantes » qui ont une part des intérêts fragmentés d'avance, comme c'est le cas dans une perspective entièrement pluraliste. Dans une telle vision, les valeurs en tension sont perçues comme s'opposant comme le feraient des adversaires. Cette opposition devrait plutôt générer « […] un mode de dialogue véritablement conversationnel, et donc non "adversariale" » [Traduction libre] (Blattberg 2008, 8). Le bien commun est un bien « pour nous[51] », qui se conçoit, se vise et se partage en commun (Taylor 1997a, 138). Malgré cet aspect commun indivisible et irréductible, un même recours à la prudence ne générera pas nécessairement des conclusions identiques de la part de tous les décideurs politiques. C'est peut-être ce qui pousse Taylor à suggérer que « la pluralité des biens devrait être évidente dans la société moderne, si nous pouvions mettre de côté les œillères que notre métaéthique réductrice nous impose » [Traduction libre] (Taylor 1982, 142).

Dès lors, on pourra parler de « fautes » par rapport au discernement du bien commun lorsque les décideurs auront émis des « […] suppositions fausses sur la réalité, sur les lois de la nature ou sur des faits […] » (Spaemann 1997, 4). De fait, on peut reconnaître un certain « sens de la réalité », qui nous permet de déterminer si nos interlocuteurs se trompent, lorsque l'on échange (Taylor 1995, 154-155). Il en va de même pour les élus. Ces derniers pourront commettre des « erreurs » par rapport au bien commun s'ils avaient de la réalité « une fausse estimation » et dans ce cas, l'erreur équivaudrait à une « connaissance erronée ». Effectivement,

> […] nous pouvons nous tromper à propos de la connaissance de ce « en vue de » et des moyens de sa réalisation, et ainsi poursuivre un but « faux ». *Mais le faux dont il s'agit là n'a pas trait au fait que l'agir ne serait pas à la hauteur d'une norme « supérieure » et extérieure à lui*. Comment en effet une norme qui ne serait pas immanente à l'aspiration elle-même pourrait-elle avoir une signification pour cette aspiration? Le faux consiste en ce que nous tenons pour le « en vue de » dernier de l'aspiration ne l'est au fond pas. Mais nous tombons alors en contradiction avec nous-mêmes. Nous voulons ce que nous ne voulons pas. (Spaemann 1997, 5)

Conséquemment, la communauté politique pourra juger des actions de ses décideurs politiques concernant l'éthique face aux technologies d'intelligence artificielle en fonction de l'adéquation

[51] Un « nous » qui « comprend ensemble », selon Taylor (1997a, 139) [Traduction libre]

de ces actions à la finalité du bien commun (Spaemann 1997, 9). Politiquement parlant, cette idée est intéressante puisqu'elle touche directement à la responsabilité que portent les politiciens devant leurs électeurs, qu'ils sont censés représenter.

c) Une sensibilité profonde au contexte

Ma critique d'une éthique de la règle présente sa contrepartie, qui est une prise en compte profonde du contexte. Cette dernière découle par ailleurs de la vertu de prudence, comme la téléologie dont j'ai fait l'apologie. Alasdair MacIntyre soutient que l'être humain, quand il cherche « ce qui est bon pour lui en tant qu'humain », doit faire des choix, et que

> de tels choix exigent un jugement et l'exercice des vertus [, qui] requiert donc une capacité de juger et de faire la bonne chose au bon endroit, au bon moment et de la bonne manière. L'exercice d'un tel jugement *n'est pas une application routinière des règles*. D'où peut-être l'absence la plus évidente et la plus étonnante de la pensée d'Aristote pour tout lecteur moderne : *il est relativement peu fait mention des règles* dans l'Éthique. [Traduction libre, je souligne] (MacIntyre 1984, 150)

La réflexion éthique doit être sensible au contexte et reconnaître qu'« […] une technologie qui favorise l'épanouissement humain dans un contexte social […] peut le compromettre dans un autre […] » [Traduction libre] (Vallor 2018, 15). Ainsi, la cible qu'est le bien commun est mouvante, « […] parce que la technologie *elle-même* remodèle continuellement le contexte social et les personnes qui le composent » [Traduction libre] (Vallor 2018, 15). On trouve un écho de cette idée dans ce qu'affirme Postman, à savoir que « le changement technologique n'est pas additif, il est écologique. […] Un nouveau média n'ajoute pas quelque chose, il change tout » [Traduction libre] (Postman 1998, 4). Il s'agit en fin de compte d'une transformation holiste. L'avènement d'une technologie remodèle le contexte dans lequel les agents humains évoluent.

C'est pour cette raison que les règles systématiques ont tendance à être trop rigides lorsqu'il est question de raisonnement éthique dans des circonstances imprévues ou changeantes. Taylor affirme à ce sujet que « toute règle générale, dérivée d'un ensemble de cas, devra être reconsidérée et réajustée finement en fonction de situations différentes » (1997b, 166). Flyvbjerg va plus loin pour soutenir que « le contexte détermine et est déterminé par la compréhension que les chercheurs

ont d'eux-mêmes » [Traduction libre] (2001, 33). Conséquemment, le raisonnement hors contexte — celui qui mobilise des règles formelles ou des procédures — présente le risque d'une incohérence avec le raisonnement de l'agent qui sera inscrit dans ce contexte particulier (Flybjerg 2001, 42). Évidemment, des lois et des règlements doivent être adoptés pour la pratique et la vie politique. C'est seulement que ces mécanismes, difficilement évitables, ne sont pas les guides optimaux d'un dialogue politique en amont de leur adoption.

Par exemple, on a vu, au chapitre dernier, que la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle stipule que les dix principes qu'elle met de l'avant « […] doivent être interprétés de manière cohérente, en tenant compte de la spécificité des contextes sociaux, culturels, politiques et juridiques de leur application » (Comité d'élaboration de la Déclaration de Montréal IA responsable 2018, §6). Certes, des lois sont nécessaires pour régir des champs d'action précis. Ce n'est pas ce qui pose problème, et j'y reviendrai un peu plus bas. Plutôt, la raison pour laquelle l'exercice de la sagesse pratique serait entravé dans ce cas précis est que le champ de ce qui est possible d'accomplir dans l'action est rétréci par cette codification qui est marquée de l'impératif de la cohérence. La réalité pratique montre que parfois, cette « cohérence » n'est pas entièrement atteignable. Déjà, le dialogue a été conçu et cadré comme se devant d'atteindre cette cible. Par ailleurs, l'importance des problèmes dans une même situation peut varier (Taylor 1997b, 163) et alors, le poids accordé aux éléments de la « règle » devra changer, sans que l'on puisse le conceptualiser à l'avance. En effet, « […] il y a des différences d'importance non seulement entre les biens, mais aussi entre les occasions où ce que nous appelons le même bien est invoqué » (Taylor 1997b, 163).

Un avantage de tenir compte du contexte est d'éviter ce qui est actuellement décrié en éthique de l'IA, soit la haute voltige théorique et le manque de portée pratique des principes et éthiques. Même une charte de principes pluraliste pourrait être interprétée sans sensibilité au contexte et ce serait une erreur. En réalité, quand il s'agit d'éthique de l'IA, « […] très souvent, les questions de valeurs que nous devons considérer sont beaucoup plus locales et contextualisées — et donc, dans cette mesure, plus faciles à aborder » [Traduction libre] (Boddington 2017, 111). Une approche dialogique nourrie par des questions, comme celles de Postman évoquées plus haut, permet que la réponse soit teintée des particularités du contexte duquel elle émerge. Un dialogue

guidé par des règles à appliquer ou des principes définis à respecter serait appauvri et n'aurait pas la même ouverture à la raison pratique. Plusieurs des modalités seraient décidées d'avance, au lieu de permettre à la situation et aux interlocuteurs d'être surpris par les particularités de la réalité.

Des décideurs politiques sur un continent ou un autre amèneront des éléments différents dans leur réflexion et leur délibération. Ce que certains appellent des « mécanismes de délibération », je le comprends comme un dialogue orienté vers le bien commun et motivé par des questions qui permettent de tâter le pouls de la diversité des points de vue. Ce serait une interprétation du propos de Jobin, Ienca et Vayena, dans leur analyse de contenu de directives éthiques pour l'IA, qui soutiennent qu'

> un défi fondamental pour l'élaboration d'un programme global pour l'IA consiste à trouver un équilibre entre la nécessité d'une harmonisation transnationale et le respect de la diversité culturelle et du pluralisme moral. Ce défi nécessitera la mise en place de mécanismes de délibération pour régler les désaccords concernant les valeurs et les implications des avancées de l'IA entre les différentes parties prenantes des différentes régions du monde. [Traduction libre] (Jobin, Ienca et Vayena 2019, 16)

Un exemple de démarche politique qui prend en compte le contexte peut être fourni, en ce qui a trait aux décideurs politiques confrontés aux enjeux éthiques de l'IA. À des fins de clarté, on pourrait reprendre la catégorisation fournie par Miles Brundage et Joanna Bryson des politiques touchant à l'IA. Premièrement, on pourrait parler des « politiques d'IA directes », qui visent précisément des SIA, comme la voiture autonome; deuxièmement, des « politiques d'IA indirectes », c'est-à-dire des politiques technologiques qui englobent l'IA, comme des lois sur la protection des renseignements personnels; troisièmement, des « politiques pertinentes en matière d'IA », soit des politiques qui ne s'adressent pas à l'IA, « […], mais dans lesquelles la connaissance d'avenirs plausibles en matière d'IA profiterait aux décideurs politiques, tels que l'éducation, l'urbanisme et les politiques sociales » [Traduction libre] (Brundage et Bryson 2016, 5).

Les décideurs politiques, lorsqu'ils discutent d'éthique de l'IA, ont donc un large champ à couvrir. La question est de savoir si leur manière de procéder s'inspirera d'une théorie éthique moniste, qui préconise le respect de certaines règles ou procédures formelles pour leur dialogue, ou une sorte d'appel à une forme de la raison instrumentale; ou s'ils se tourneront plutôt vers un raisonnement pratique. Nous avons vu que ce raisonnement pratique peut être moniste orthodoxe,

dans le contexte moniste de l'éthique de la vertu, ou encore pluraliste, dans le constat de la fragmentation irrémédiable des valeurs qui exclut un bien commun (contrairement à l'éthique de la vertu). La démarche peut aussi être moniste non orthodoxe, c'est-à-dire viser l'unité ou la réconciliation, qui est possible en principe, tout en sachant que dans la pratique, elle comporte un risque d'échec. L'approche pluraliste est reprise par le professeur de droit Ryan Calo, pour qui l'un des défis pour l'élaboration des politiques concernant l'IA se retrouve précisément dans les conflits de valeurs. Par exemple, lorsque des décideurs politiques se mettront d'accord, en partenariat avec le secteur privé, pour développer des systèmes accordant la priorité à la transparence, la précision s'en trouverait touchée négativement. Les conséquences de ce compromis sont importantes en ce que de tels SIA peuvent être employés dans des instances judiciaires, par exemple (Calo 2017, 12).

La « politique d'IA indirecte » qu'est la Loi sur la protection des renseignements personnels et les documents électroniques (LPRPDE), au Canada peut servir de cas de figure. Le Commissariat à la protection de la vie privée du Canada a lancé une consultation de la population sur ses propositions touchant à la réglementation de l'IA (Commissariat à la protection de la vie privée du Canada 2018).[52] Il a en effet élaboré des propositions de changement de la LPRPDE de façon à réglementer plus avant l'emploi de l'IA, qui ne contient, au moment de la consultation, aucune disposition particulière la concernant. Les dilemmes qui se posent quant aux SIA et à la protection des données sont, par exemple, que la nécessité de limiter la collecte des données pourrait diminuer l'efficacité et la qualité des résultats des systèmes employant l'intelligence artificielle, tension semblable à celle que relevait Calo (Commissariat à la protection de la vie privée du Canada 2018). La Commission a lancé un appel à la discussion incluant des questions, bien documentées, qui permettent aux citoyens et aux experts d'amener leur contribution à la discussion des décideurs politiques en la matière.[53]

[52] Depuis, en novembre 2020, le Commissariat à la protection de la vie privée du Canada a publié un « cadre réglementaire pour l'IA », dans le but de réformer la LPRPDE (Commissariat à la protection de la vie privée du Canada 2020).

[53] Le Montreal AI Ethics Institute (MAIEI) a repris la balle au bond et organisé une consultation citoyenne sur la question en février 2020, à laquelle j'ai participé. Nous étions divisés en équipes et discutions des questions en formulant des recommandations concrètes les concernant. Le MAIEI a par la suite agrégé les réponses en ajoutant les siennes, et préparé un rapport présenté à la Commission (voir Snyder Caron 2020)

Il est intéressant de noter que, dans le contexte de la consultation citoyenne, la question du RGDP européen comme étant ou non adapté au contexte canadien a été soulevée. Il s'agirait là d'un dialogue éthique prenant en considération des particularités du contexte. De ce que l'on peut documenter du processus décisionnel (de ce qui n'est pas, autrement dit, conduit à huis clos), une telle approche, si elle peut déboucher sur recommandations de règles particulières, s'approche tout à fait d'une démarche éthique dialogique pratique. La Commission aurait tout aussi bien pu engager un éthicien déontologiste pour qu'il élabore un cadre kantien de règles à intégrer dans la loi existante, ou pour qu'il suggère aux élus une procédure décisionnelle sûre. Il est vrai que la présente analyse ne se base que sur le processus visible de consultation. À savoir si les décideurs politiques devront, comme le soutenait Calo, trancher entre des valeurs s'excluant mutuellement, il est difficile de le dire. Pour que la démarche participe entièrement de la raison pratique, il faudrait éviter de tracer une démarcation claire entre le dialogue citoyen, puis la prise de décision qui, seule, serait politique. Selon Blattberg, cette ligne de démarcation devrait être pointillée et non solide (Blattberg 2009, 41).

À ce sujet, il importe de réitérer que l'argumentaire ici étayé ne soutient pas qu'aucune règle éthique (ou aucune loi) ne puisse être adoptée par des décideurs politiques pour réglementer le développement, le déploiement et l'usage de l'intelligence artificielle. Simplement, des approches éthiques de la règle formelle ne sont pas adaptées pour guider la délibération de ces décideurs politiques en la matière, pour guider, autrement dit, l'orientation de la flèche vers la cible. La différence est là. Berlin et Williams le reconnaissent quand ils affirment qu'

> il est vrai qu'il existe certains conflits, notamment sur des questions de politiques publiques, qui sont mieux tranchés par des règles simples et publiables que par des jugements individuels d'importance. De même, il y a d'autres questions qu'il vaut mieux laisser aux jugements d'importance. *De plus, il n'existe pas, inévitablement, de procédure mécanique pour décider lesquels sont lesquels*. [Traduction libre, je souligne] (Berlin et Williams 1994, 308)

Williams et Berlin rappellent que « […] la décision pratique ne peut pas, en principe, être rendue complètement algorithmique et qu'une conception de la raison pratique qui vise un idéal algorithmique doit être erronée » [Traduction libre] (Berlin et Williams 1994, 308). Le raisonnement éthique, pour quelque question politique touchant à l'IA que ce soit, ne pourrait être abandonné à un système préprogrammé, comme on l'a vu proposé dans les démarches

déontologiques touchant à l'IA au chapitre quatre.[54] Cette façon de procéder ne tient pas compte de toutes les potentialités de la raison et que, dans l'éventualité où des décisions politiques doivent être prises, le dialogue facilité par des questions ouvrant sur les particularités du contexte serait souhaitable.

d) La reconnaissance des dilemmes insolubles

Un quatrième aspect qui est essentiel à toute approche éthique politique est tiré des postulats centraux du pluralisme en éthique, soit l'inévitabilité des dilemmes éthiques potentiellement insolubles. L'une des raisons de cette insolubilité peut être l'incommensurabilité de valeurs, lorsque prises dans un contexte particulier. On dit des valeurs qu'elles sont incommensurables lorsqu'elles « ne peuvent être réduites à une mesure commune » (Hsieh 2016, §1). Les valeurs sont ainsi irréductibles les unes aux autres ou encore à un principe quelconque. Une autre explication, distincte de l'incommensurabilité des valeurs, est leur potentielle incompatibilité. En effet, le fait que certaines valeurs, prises dans une compréhension holiste, soient incommensurables les unes aux autres, ne signifie pas automatiquement qu'elles soient incompatibles. Un moniste pourrait soutenir que des valeurs habituellement incommensurables, comme la liberté et l'égalité, sont compatibles dans une théorie politique comme le libéralisme.[55] Sur un autre plan, la multiplicité des biens (qu'ils soient conçus comme des valeurs à promouvoir ou encore des vertus) de pair avec la reconnaissance de leur incommensurabilité fragilise les démarches kantienne et utilitariste en éthique politique. De fait,

> […] aucune procédure de réflexion unique, qu'elle soit celle de l'utilitarisme, ou une théorie de la justice basée sur un contrat idéal, ne peut rendre justice à la diversité des biens que nous devons peser ensemble dans la pensée politique normative. [Traduction libre] (Taylor 1982, 142)

[54] Cette idée que le raisonnement éthique ne peut être réduit à une formule ou un calcul se trouve aussi dans les écrits de Joseph Weizenbaum (1976, 13-14), qui de son côté cite Hannah Arendt (1972, 11 et suivantes). Merci à Charles Ess pour la suggestion et la référence.

[55] En suivant les continuums métaphysiques que met de l'avant Charles Blattberg (2018), soit le spectre méréologique et le spectre hiburologique, on peut distinguer entre, d'une part, l'*incommensurabilité* des valeurs relevant du holisme sur le spectre méréologique; et d'autre part, l'*incompatibilité* des valeurs, qui relève du pluralisme, sur le spectre hiburologique. Le monisme ne reconnaît donc pas l'incompatibilité potentielle des valeurs, mais peut reconnaître leur incommensurabilité. Il n'est toutefois pas nécessaire d'explorer plus en profondeur cette distinction aux fins de cette section.

Ce quatrième élément de ma proposition fait que cette dernière ne peut être catégorisée comme relevant uniquement de l'éthique de la vertu ni, en conséquence, comme étant moniste orthodoxe. Déjà, il a été question de ma conception de la prudence, qui diffère de la *phronesis* aristotélicienne sur la base de la téléologie. Cela étant dit, il est évident qu'il y a des similarités entre mon approche et celle de l'éthique de la vertu, notamment la sensibilité profonde au contexte, en plus d'une forme de prudence et de téléologie (« douce », dans mon cas). Ce serait une forme non orthodoxe de l'éthique de la vertu en politique, si l'on veut.

J'ouvre la porte à la reconnaissance de dilemmes insolubles, car l'éthique de la vertu, comme approche moniste orthodoxe, n'accorde pas une place suffisante à la possibilité des dilemmes indécidables. Si elle peut reconnaître l'incommensurabilité des valeurs, l'incompatibilité de biens (ou de vertus) ne fait partie de ses postulats, en raison de son monisme. Le monisme de l'éthique de la vertu, on l'a vu au deuxième chapitre, consiste en l'unité du bien en une personne et conséquemment en l'unité de ces vertus, qui sont des capacités à faire le bien. L'agent qui possède une vertu possède nécessairement les autres et donc la *phronesis*, car il est « […] impossib[le] d'être prudent sans être vertueux » (Aristote 2014, Livre VI, 13, 1144a34-35). Plus encore, « […] il n'est pas possible d'être homme de bien au sens strict, sans prudence, ni prudent sans la vertu morale » (Aristote 2014, Livre VI, 13, 1144b30-35 -1145 a1). La *phronesis*, dans son acception aristotélicienne, scelle le monisme de l'approche en assurant l'unité des vertus. Pour le *phronimos*, toutes les vertus sont compatibles dans la pratique. Or, je pense que la prudence est une vertu à recommander aux décideurs politiques, et la proposition que j'étaierai au chapitre suivant les aidera à la mettre en pratique s'ils le désirent. Ainsi, affirme Aristote, on peut

> […] réfuter l'argument dialectique qui tendrait à établir que les vertus existent séparément les unes des autres, sous prétexte que le même homme n'est pas naturellement le plus apte à les pratiquer toutes, de sorte qu'il aura déjà acquis l'une et n'aura pas encore acquis l'autre. Cela est assurément possible pour ce qui concerne les vertus naturelles; par contre, en ce qui regarde celles auxquelles nous devons le nom d'homme de bien proprement dit, c'est une chose impossible, car en même temps que la prudence, qui est une seule vertu, toutes les autres seront données. (Aristote 2014, Livre VI, 13, 1144b30-35 -1145 a1)

Un autre aspect moniste orthodoxe de l'éthique de la vertu que j'ai de la difficulté à intégrer est l'idée selon laquelle si l'agent agit selon la vertu requise par la situation, il a agi globalement de manière vertueuse : conséquemment, même si l'on peut en regretter quelques conséquences, cet

agent n'a rien à se reprocher au plan moral. Autrement dit, il n'y a pas de « mains sales », chose difficile à éviter totalement dans le monde politique. Plus encore, les situations de la vie politique sont si complexes qu'elles ne requièrent pas forcément l'exercice d'une seule vertu ou d'un ensemble de vertu pour être correctement appréhendées dans leur entièreté.

À ce sujet, Vallor reconnaît que l'incommensurabilité de certaines valeurs au plan culturel est un défi pour sa défense d'une éthique de la vertu globale face à la technologie (2016, 52). Néanmoins, elle croit qu'en réussissant à s'entendre sur les biens que l'on cherche à atteindre, il sera possible de juger sur les moyens d'y parvenir (Vallor 2016, 54). Vallor fait toutefois l'impasse sur la possibilité que les décideurs politiques, représentant la communauté politique, ne puissent s'entendre effectivement sur ces biens qui, pourrait-on dire, entrent dans la compréhension du bien commun. Il ne semble donc pas être question de la possibilité de l'*incompatibilité* de certaines valeurs dans la pratique et conséquemment, de l'indissolubilité réelle de certains dilemmes. Coeckelbergh affirme que les dilemmes, en éthique de l'IA, sont incontournables. De fait, explique-t-il,

> […] si nous prenons l'éthique de l'IA au sérieux et mettons en œuvre ses recommandations, nous pourrions être confrontés à certains compromis, en particulier à court terme. L'éthique peut avoir un coût : en termes d'argent, de temps et d'énergie. Cependant, en réduisant les risques, l'éthique et l'innovation responsable soutiennent le développement durable à long terme des entreprises et de la société. [Traduction libre] (2020a, 174)

La réalité de ces désaccords profonds ne peut être palliée par une simple « formule » (Taylor 1994c, 213-214) ou encore une théorie systématique. C'est dans ce sens qu'abonde Paula Boddington lorsqu'elle discute de la viabilité d'une éthique de la vertu face à l'intelligence artificielle :

> […] pour Aristote, prendre la décision éthique adéquate était compris comme obtenir *la réponse appropriée*, comme toucher une cible aussi précisément que possible. *Il n'avait certainement pas l'intention de permettre le pluralisme éthique*. L'appel à la *phronesis* comme desiderata dans les codes d'éthique pour les technologies en développement rapide peut en fait ne pas apporter de réponse, mais indiquer au contraire la profondeur du problème. [Traduction libre, je souligne] (Boddington 2017, 102)

Autrement dit, il ne suffit pas d'appeler à la vertu de prudence pour éviter les dilemmes éthiques insolubles. Non seulement elle pourrait n'être pas pratiquée, mais elle peut aussi se révéler dans une certaine mesure « impuissante » face aux dilemmes éthiques. Assurément,

> *il existe des conflits de devoir*. Il y a des cas où il est juste de ne pas tenir une promesse parce qu'une chose plus urgente ou plus importante le justifie. Il est aisé de savoir ce qu'on doit faire dans des cas d'école simple. Mais *la plupart des situations dans lesquelles nous nous trouvons sont complexes*. (je souligne) (Spaemann 1999, 115)

Une façon de voir les choses serait de dire que le monisme de l'éthique de la vertu a besoin de certains des postulats du pluralisme des valeurs pour être viable politiquement, c'est-à-dire pour exhiber un certain réalisme politique. La cohérence interne dont je reproche la carence aux directives en tension métaéthique se trouve dans l'adaptation de l'approche éthique au monde de la pratique, qui est foncièrement imparfait. Il s'agit là d'une idée qui s'aligne très bien avec le monisme non orthodoxe. De fait,

> certains souhaits sont tout simplement *inconciliables entre eux*. De même qu'il y a en moi-même des souhaits opposés de rang différent, de même les souhaits de différentes personnes peuvent être de rang différent. […] Une solution acceptable par les deux ne peut exister que s'il y a un critère commun possible, c'est-à-dire un critère susceptible de vérité pour évaluer les souhaits. [je souligne] (Spaemann 1999, 23-24)

Il importe d'ajouter que les biens incommensurables ne sont pas pour autant incomparables entre eux, du moins selon les pluralistes des valeurs non décisionnistes (Blattberg 2015, 2). Leur combinaison pourrait et devrait faire l'objet d'un raisonnement pratique prudentiel, à la lumière du bien commun politique. La raison a quelque chose à dire dans les dilemmes de valeurs incommensurables (Blattberg 2018, 158). Ce raisonnement pratique prudentiel, dans la comparaison des valeurs, se ferait de manière optimale dans un dialogue guidé par des questions préalablement pensées, mais qui peuvent évidemment déboucher sur d'autres questions imprévues. Cela fait partie de la sensibilité au contexte et à une approche éthique non procédurale. Il serait souhaitable que cela arrive, d'ailleurs.

1. Des exemples de dilemmes potentiellement insolubles

On peut illustrer le propos en prenant la tension entre la valeur du respect de la vie privée (*privacy*) et celle de la sécurité : elles sont parfois irréconciliables de manière absolue. Ces deux valeurs ne sont pas toujours parfaitement compatibles dans la pratique. Autrement dit, pour que leur cohabitation soit possible, une perte de quelque genre que ce soit devra être acceptée. Par exemple, l'innovation que représentent les algorithmes de reconnaissance faciale pour la détection d'individus dangereux pose des problèmes de profilage et de « catégorisation » erronée ou faussée (au sens de « biaisée ») d'individus (Müller 2020, §2.4). Si l'on désire adopter un système d'intelligence pour assurer la protection nuit et jour de sa propriété, il faut souvent accepter un compromis avec l'accumulation de données personnelles par ce système. Plus généralement, les décideurs politiques pourraient être confrontés entre « d'une part, les résultats d'un raisonnement conséquentialiste sur le bien commun, d'autre part, les exigences de la fidélité à soi » (Taylor 1997b, 164). Il s'agit de dilemmes éthiques qui ne peuvent être résolus sans engendrer une perte. Un compromis doit être fait : le mieux qu'un décideur politique peut faire, dans une telle situation, est de limiter ses dommages potentiels. Taylor explique, à la suite d'une réflexion de Bernard Williams, que de

> […] choisir l'action qui aura les meilleures conséquences peut, dans certains cas, entrer en conflit avec les exigences de [son] intégrité; les exigences de la bienveillance à l'égard d'autrui peuvent contredire celles de [son] propre épanouissement, ou les exigences de la justice celles de la pitié et de la compassion. (Taylor 1997b, 156)

La question bien connue de l'alignement des valeurs (par exemple IEEE 2016; IEEE 2017; Müller 2020), en éthique de l'intelligence artificielle, devient ici cruciale.

Un autre cas de figure serait celui des élus politiques qui, dans un contexte de pandémie, pourraient être approchés par des compagnies du milieu industriel ou encore des chercheurs universitaires avec des moyens technologiques pour aider à enrayer la propagation du virus. Ce type d'innovation pourrait prendre la forme d'un programme opérant sur les téléphones cellulaires des citoyens, les prévenant de leur proximité et de leurs interactions avec des porteurs de la maladie en pleine expansion (Bengio et Dilhac 2020). Une application de la sorte pourrait « […] avertir les utilisateurs potentiellement infectés avant même qu'ils développent des symptômes de la maladie,

au moment où ils sont justement les plus contagieux sans le savoir » (Bengio et Dilhac 2020, §4). De tels systèmes pourraient aussi servir aux gouvernements à envoyer de l'aide dans les endroits les plus touchés par le virus (Expert Advisory Group on Society, Technology and Ethics in a Pandemic [STEP] 2020, 7).

Certes, il existe des manières d'anonymiser des données, mais la réalité demeure que la plupart du temps, des compromis entre ces deux valeurs doivent être acceptés pour « maximiser » l'une des deux. Par exemple, l'application « COVI » proposée par le Mila, l'Institut québécois d'intelligence artificielle, dans le contexte de la pandémie de la COVID-19, aurait été régie par la Charte canadienne du numérique et la Charte canadienne des droits et libertés, selon Yoshua Bengio et Marc-Antoine Dilhac (2020, §6). Elle aurait aussi respecté les principes de la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle, par exemple celui d'autonomie, en ne contraignant pas les utilisateurs à respecter les recommandations fournies par le système, et celui de respect de la vie privée, en anonymisant et en cryptant les données, qui seraient gérées par un organisme à but non lucratif (Bengio et Dilhac 2020, §8-9). Ces précautions sont excellentes et il faut les recommander. On peut quand même se demander si, dans la pratique, l'algorithme fonctionne de manière optimale indépendamment du nombre d'utilisateurs de l'application en question. Dans un tel contexte, on peut déceler un certain dilemme entre la décision de déployer des applications de la sorte, tout en connaissant les limites techniques, en termes de précision, de ces systèmes (Expert Advisory Group on Society, Technology and Ethics in a Pandemic [STEP] 2020, 7). La valeur de l'autonomie viendrait ici poser problème en compromettant la performance du système. Ce dilemme, toutefois, se pose pour une panoplie de SIA. Ou encore, étant donné le travail accru et les coûts liés à l'anonymisation des données (Müller 2020, s.p.), il faudrait voir si les lois du marché viendraient contrecarrer les bonnes intentions des entreprises exploitant ces systèmes innovateurs et freiner l'adoption de telles applications.

Plus généralement, le développement même des technologies employant l'intelligence artificielle peut constituer un dilemme éthique à lui seul. Si l'on est d'accord pour dire, avec Neil Postman, que « les avantages et les inconvénients des nouvelles technologies ne sont jamais répartis uniformément dans la population » [Traduction libre] (Postman 1998, 2), le simple fait qu'une majorité de pays industrialisés travaillent à l'élaboration et au déploiement de SIA très

pointus implique *de facto* un conflit avec le principe d'équité ou d'égalité, sur le seul plan de l'économie. Il y a lieu de se demander si les bénéfices des innovations dans le domaine de la santé qui sont permises par des algorithmes (voir par exemple Agence France-Presse [AFP] 2020a; Daouda 2018; Nguyen 2018) toucheront toutes les couches de la population d'un pays donné, ou encore tous les pays de manière équitable.

En réalité, c'est une poignée de gouvernements, dans le monde, qui peuvent se permettre de financer par centaines de millions de dollars la recherche en IA. Malgré les bénéfices indéniables de cette recherche, le ravin économique séparant les mieux nantis des plus pauvres est inexorablement amplifié. Plus encore, l'argent étant une ressource limitée, il y a un déséquilibre entre ceux qui bénéficient de telles subventions (au plan de l'emploi, ou encore du financement de la recherche) au détriment d'autres individus ou centres de recherche. Il semble difficile de penser comment aborder de tels problèmes sans la notion d'accommodement, qui est souvent associée à la position pluraliste. En effet, il est ardu de penser de tels scénarios sans prévoir des pertes d'un côté ou d'un autre, puisque les ressources impliquées ne sont pas infinies.

Il est loin d'être aisé, voire impossible, de proposer une approche éthique qui puisse éviter, dans l'absolu, toutes les compromissions de valeurs et ce, peu importe leur sérieux ou leur niveau de gravité. Étant donné les limites des ressources, des contextes, des personnes et surtout, en tenant compte de la liberté des agents dans leur processus de décision éthique, il me semble que d'admettre l'insolubilité de certains conflits est plus près de la réalité pratique politique. De toute façon, dans une perspective moniste non orthodoxe, même si on vise un bien commun,

> la plupart du temps une action meilleure que celle que l'on fait est encore possible. Et il serait tout à fait faux de dire que l'on a toujours le devoir de faire la meilleure des actions possibles. C'est tout à fait impossible. (Spaemann 1999, 116)

Si le conflit entre les valeurs peut être réconcilié, tant mieux. Toutefois, il faut être prêt à ce que parfois, cette réconciliation ne soit pas possible, et que l'on doive effectuer un compromis entre nos valeurs. La proposition concernant le dialogue des décideurs politiques pour les enjeux de l'IA, que je présenterai dans le chapitre suivant, ne fera pas l'impasse sur cette réalité.

4. Sur la viabilité de ma proposition moniste non orthodoxe

J'ai exposé, au fil de ce chapitre, en quoi les approches monistes orthodoxes à l'éthique que sont l'utilitarisme et le déontologisme, en raison de l'unification actuelle ou potentielle du réel, postulent qu'à tout dilemme éthique se trouve une solution. C'est aussi le cas de l'éthique de la vertu qui, sans mettre de l'avant la raison théorique ou instrumentale, ou encore une éthique de la règle, participe de cette vision unifiée du réel. Les solutions monistes orthodoxes (la pratique de la vertu, le respect du devoir ou la maximisation de l'utilité) permettent à l'agent de s'en tirer « les mains propres » moralement, même si la situation finale laisse à désirer sous quelques aspects. Par ailleurs, faire découler la pratique de l'éthique du raisonnement théorique peut participer d'une sensibilité naturaliste qui tend à traiter l'éthique comme une discipline des sciences naturelles (Taylor 1982, 139-141). Il s'agit d'une tentation puisque l'entrée en tension de valeurs irréductibles est difficile à systématiser au plan épistémologique. On peut aussi craindre que ce conflit ne nous mène vers une forme de scepticisme moral (Taylor 1982, 139). C'est ce qui explique que, dans la modernité

> […] nous avons été entraînés dans *une définition restrictive de l'éthique*, qui tient compte de certains des biens que nous recherchons, par exemple l'utilité, et du respect universel de la personnalité morale, tout en excluant d'autres, à savoir les vertus et les objectifs comme ceux mentionnés ci-dessus, en grande partie au motif que les premiers font l'objet d'une contestation moins embarrassante. [Traduction libre] (Taylor 1982, 140)

C'est ici qu'il m'apparaît que le monisme non orthodoxe permet une ouverture en ce qu'il reconnaît que, dans certains cas, la « solution » sera hors de portée, soit en raison des manquements de la théorie, soit à cause de la réalité pratique. Il est plus réaliste de le reconnaître que d'esquiver l'inévitabilité des mains sales. Cette reconnaissance « réfléchie » n'atteint pas la cohérence interne de l'approche : l'unité ou la réconciliation est toujours visée. Cela dit, davantage de réflexion serait nécessaire sur la qualification de ces mains sales, et leur relation avec la notion de culpabilité morale. Les manquements de la théorie ou du monde de la pratique ne sont pas, après tout, toujours tributaires de l'acteur aux prises avec un problème éthique donné.[56]

[56] C'est la raison pour laquelle j'aimerais — dans un futur plus ou moins lointain, mais pas dans cette thèse — m'atteler à esquisser les différences entre la *responsabilité* morale des décisions prises et la *culpabilité* morale de l'agent ayant pris la ou les décisions éthiques, dans la perspective du monisme non orthodoxe. Une chaîne d'erreurs peut être

Lorsque des décideurs politiques devront se prononcer quant à l'éthique d'un SIA, il pourra arriver que toutes les solutions que l'on peut recenser impliquent une forme de perte morale, même si cette dernière est minime. Face à cette possibilité, le monisme non orthodoxe ne verse pas dans la tragédie, puisque telle n'est pas l'issue inéluctable de chaque situation. La posture de l'inévitabilité de la perte constitue la faiblesse du pluralisme des valeurs « pur et dur ». Rien ne dit que, moyennant un dialogue entre décideurs politiques, scientifiques, citoyens, ou autre personne intéressée, une solution réconciliant les intérêts ne puisse être découverte. La tragédie demeure toutefois une probabilité. À cet égard, un pluraliste critiquerait le penseur moniste non orthodoxe pour son manque de réalisme, voire pour son utopisme. Le moniste non orthodoxe lui, nous l'avons vu, reproche au pluraliste des valeurs son pessimisme.

En ce qui concerne la métaéthique, la combinaison d'influences en provenance d'approches éthiques différentes peut fonctionner si ce mélange est réfléchi et transparent. Si tel n'est pas le cas, non seulement l'approche éthique risque d'être incohérente au plan métaéthique, mais les décideurs politiques ne sauront pas si, en adoptant une telle approche, ils vont au-devant de compromissions de valeurs. On l'a vu, des impératifs parfois incohérents sont mis de l'avant dans une même démarche et peuvent se révéler incompatibles dans la pratique. Des éléments monistes comme la primauté absolue de certaines valeurs ou principes peuvent cohabiter avec la volonté d'affirmer toutes les valeurs possibles, sans échelle pour les ordonner ni de réflexion sur le potentiel tragique de l'impossibilité de les harmoniser toutes dans la pratique. Probablement, un penseur moniste orthodoxe verrait dans cette ouverture à la perte un manque de rigueur ou une exagération des différends. Le moniste non orthodoxe le voit tout simplement comme un réalisme accru, dénué de distorsions dues à la théorie mal arrimée à la pratique.

La proposition métaéthique que j'ai décrite ici trouve son ancrage dans le monisme non orthodoxe. Le présent chapitre constitue donc un exercice de prise de conscience et de transparence par rapport à l'origine de ces postulats, puisque l'absence de cet exercice est précisément ce que je

commise en amont de la décision de l'agent, qui se retrouve devant une réalité éthique fragmentée et peut-être impossible à réunifier, sans qu'il en soit directement coupable.

reproche aux directives éthiques laissant transparaître des tensions.[57] La cohabitation des quatre éléments qui forment ma proposition sera développée, concrétisée et mise en application dans le chapitre suivant. De fait, j'exposerai en quoi ils forment les piliers du dialogue politique pour l'éthique de l'IA, dialogue qui prendra sa source dans la question, plutôt que le principe, la règle ou la théorie, suivant en cela certaines notions de la philosophie herméneutique.

[57] La différence de ma proposition éthique avec le « cadre d'interprétation » que représente le « pluralisme éthique *pros hen* [EP(ph)] », tel que le systématise Charles Ess (2020), devrait apparaître plus manifestement ici. D'une part, je ne travaille pas avec la notion de « cadre », qui m'apparaît trop rigide; d'autre part, j'ai distingué ma compréhension de la vertu de prudence de la *phronesis* tout aristotélicienne. Il est vrai, cependant, que nos propositions sont similaires au regard de l'importance de la vertu, de la prise en compte des différents points de vue et d'une certaine « équivocité » des termes (Ess 2020, 561-562), ainsi que de la centralité de l'acte d'interpréter ou de juger (Ess 2020, 554, 562). Cette notion d'interprétation sera développée davantage dans le chapitre suivant. De mon point de vue, le « pluralisme éthique *pros hen* » que propose Ess est très proche du monisme non orthodoxe. L'auteur le rapproche par ailleurs, de lui-même, au propos de Charles Taylor (Ess 2020, 561), que j'ai identifié comme un penseur moniste non orthodoxe.

Chapitre 7 – Aux décideurs politiques

« […] le prudent, disons-nous, a pour œuvre principale de bien délibérer; mais on ne délibère jamais sur les choses qui ne peuvent être autrement qu'elles ne sont, ni sur celles qui ne comportent pas quelque fin à atteindre, fin qui consiste en un bien réalisable. » — Aristote (2014, Livre VI, 8, 1141b)

« *Contrairement à l'opinion généralement répandue, il est plus difficile de questionner que de répondre* […]. » — Hans Georg Gadamer (1990, 385-386)

Introduction

Retour sur les questions de recherche

Deux interrogations ont orienté ma recherche jusqu'à présent dans cette thèse. La première, « Quelle(s) tradition(s) ou approche(s) éthique(s) informent les directives éthiques sur l'intelligence artificielle parues entre 2016 et 2020? », appelait une réponse plus descriptive et analytique. La section « Portrait », composée des chapitres 4 et 5, de même que le début du chapitre 6, mettent en lumière que les démarches exhibant une tension métaéthique dominent dans la littérature. Mon hypothèse de départ était que les démarches procédurales et pluralistes allaient être les plus nombreuses. Je ne me doutais pas de l'existence de démarches présentant une tension sur le plan de la métaéthique, et encore moins de leur récurrence.

Je l'ai expliqué dans le chapitre précédent : le reproche que j'adresse à de telles directives est l'« inconscience » (ou la non-prise en compte) de cette tension, qui a son lot de conséquences sur le dialogue des décideurs politiques. En effet, ces derniers ne pourront estimer s'ils devront ou non se salir les mains, s'ils feront face à des dilemmes insolubles ou non. Puis, la forme de dialogue à adopter sera différente selon l'approche éthique favorisée : dialogue-t-on en tenant pour acquis qu'il y aura des valeurs à compromettre, ou cherchons-nous plutôt une solution qui peut unifier ce qui est divisé? Plus encore, la tension métaéthique peut allier, inconsciemment, les faiblesses des approches éthiques procédurales ainsi que du pluralisme des valeurs que j'ai décrites dans le

chapitre précédent. C'est la raison pour laquelle j'ai aussi, dans ce chapitre, mis de l'avant ma proposition métaéthique en explicitant tous mes postulats et leurs conséquences. Ils ouvrent la porte à une potentielle compromission des valeurs en raison des limites du monde de la pratique.

Ma seconde question de recherche, « Quelle approche éthique permet de favoriser un dialogue optimal des décideurs politiques en ce qui a trait à l'éthique de l'IA? », nécessite une réponse beaucoup plus normative. C'est pour cette raison qu'au chapitre dernier, j'ai présenté ce que je considérais comme les fondements éthiques optimaux pour favoriser le dialogue de ceux qui prendront les décisions au plan politique. Cependant, je ne me contenterai pas de fournir l'approche éthique qui me semble la meilleure pour favoriser le dialogue. J'aimerais aller un peu plus loin et proposer, dans ce dernier chapitre de la thèse, une démarche dialogique pour les décideurs politiques. C'est donc au fil des prochaines pages que je terminerai de répondre à la seconde question de recherche. Mon hypothèse au départ était que l'approche à offrir aux politiciens devait se composer d'une sorte de fusion entre l'éthique de la vertu et la philosophie herméneutique, soit une « éthique de la vertu herméneutique ». Le lecteur est à même de constater, grâce au chapitre 6, que mon point de vue s'est nuancé à ce sujet — du moins en ce qui concerne l'éthique de la vertu. Si je m'en inspire encore, je m'en éloigne à certains niveaux.

Plan du chapitre

Il convient de réitérer les quatre éléments que j'estime optimaux pour approcher l'éthique, à savoir 1) la prudence comme vertu intellectuelle, 2) une orientation téléologique « douce » vers le bien commun, 3) une sensibilité profonde au contexte et 4) une reconnaissance des dilemmes insolubles. La proposition que je développe ici résulte d'une rencontre entre ces quatre éléments et une approche herméneutique au dialogue, soit l'échange d'interprétations. L'idée centrale de ce chapitre est qu'une démarche par questions, en éthique et en politique, sera plus fructueuse qu'une approche par principes. Dans cette optique, j'approfondirai en premier lieu le raisonnement qui sous-tend la thèse selon laquelle la métaéthique peut influencer la pratique du dialogue. Je développerai ensuite, dans une première section de ce chapitre, une présentation de ce qu'est le courant herméneutique, de son lien avec l'éthique, avec la politique et avec la primauté du questionnement. J'en profiterai pour clarifier où je me situe dans la compréhension de la

philosophie herméneutique et en quoi je la réconcilie avec une démarche appelant la raison pratique, plutôt que la raison théorique. Dans la deuxième section, je mettrai en relation l'herméneutique telle que définie dans la première section, avec les quatre éléments d'une démarche éthique optimale. En résulteront des questions que les décideurs politiques devraient se poser à eux-mêmes et entre eux, pour leur dialogue sur des enjeux touchant à l'éthique de l'IA, quels qu'ils soient. Enfin, ces questions seront insérées dans un « parcours herméneutique » que les décideurs politiques sont invités à emprunter pour dialoguer sur l'éthique de l'IA. L'illustration du parcours, à la Figure 1, ainsi que les questions qui y sont associées, dans le Guide herméneutique, se trouvent à la fin de ce chapitre.

1. Métaéthique et modes de dialogue

Je l'ai mentionné d'entrée de jeu dans la thèse : les approches éthiques à l'intelligence artificielle pourront avoir une incidence sur le type de dialogue des décideurs politiques. La politique étant la réponse au conflit par le dialogue, et le conflit en question pouvant renvoyer à des interprétations potentiellement divergentes du bien commun en ce qui a trait à l'usage des SIA, l'ancrage métaéthique a une importance capitale. Plus encore, le dialogue des législateurs entraînera des conséquences pratiques différentes, selon que l'on tient pour acquis que les intérêts sont irrémédiablement fractionnés d'avance (pluralisme), ou non (monisme). Cela revient à savoir si les décideurs politiques devront aborder les enjeux en étant prêts à se salir les mains en compromettant leurs valeurs, ou bien si la théorie formelle les en sauvera. Gibert (2020) affirme que justement, la multiplicité des valeurs en jeu dans le domaine de l'éthique de l'IA justifie que ce soit les décideurs politiques qui puissent se saisir du thème de l'éthique de l'intelligence artificielle. Il est vrai qu'en tant que représentants de la population, les élus en politique sont à même de connaître et de faire valoir cette multiplicité.

D'un côté, certaines démarches éthiques préconisent des approches binaires comme des listes de contrôle, qui classifient les exigences en deux catégories : respectées ou non. Cette façon de voir peut faire l'impasse sur le fait que les enjeux éthiques de l'IA ne sont pas purement techniques (Madaio et al. 2020, 3). Plus encore, avancer de manière procédurale en vue de respecter

des biens éthiques peut conduire à concevoir ces derniers de manière procédurale également (Madaio et al. 2020, 8). Même une procédure éthique très sophistiquée ne pourrait englober l'entièreté des possibilités de concevoir un dilemme éthique, puisque selon Dignum, chaque personne l'évaluerait de manière différente étant donné sa subjectivité (2019, 36). Penser le contraire renverrait à l'un des écueils du monisme orthodoxe tel que pourraient l'exprimer ses formes utilitaristes ou déontologiques. Certes, il peut être facilitant pour un programmeur de recevoir un « code éthique » à programmer de manière binaire. En revanche, Dignum estime que les SIA ne peuvent raisonner comme les humains dans des situations de dilemmes éthiques, du moins pas pour le moment. Un humain présentera toujours une richesse accrue dans la résolution de problèmes éthiques (Dignum 2019, 43-44), précisément parce qu'il a la capacité de raisonner plus complètement que simplement par l'entremise d'une liste de contrôle prépensée.

Un autre aspect de la prise de décision éthique qui est passé sous silence avec des approches trop « manichéennes » est précisément la sensibilité au contexte, une manière de procéder au « cas par cas » (franzke et al. 2020, 4). L'être humain qui, rappelle Dignum, a cette capacité de raisonner qui est plus « entière » et riche qu'une liste de contrôle ne pourrait le permettre, est aussi un être habilité au dialogue, aux rencontres d'horizons culturels différents. C'est du moins l'avis d'aline shakti franzke, Anja Bechmann, Michael Zimmer et Charles Ess dans leur rapport sur l'éthique de la recherche Internet (franzke et al. 2020, 4-5). Pour ces derniers, cocher des cases ou établir des recettes une fois pour toutes n'est pas fructueux en éthique (*Ibid*., 6). Ainsi, une approche purement procédurale en éthique est mal adaptée à la réalité que l'on cherche à évaluer.

D'un autre côté, si l'on aborde la réalité éthique comme autant de valeurs fractionnées entre différentes parties prenantes, l'objectif du dialogue peut être de chercher à « atténuer les risques » (*mitigating risks*) sans pour autant les éliminer complètement ou garantir que cette opération sera un succès (Madaio et al. 2020, 3; voir aussi Blattberg 2018, 158-159). Il est clair qu'une telle approche au dialogue est informée par la métaéthique du pluralisme des valeurs et que le type de dialogue privilégié sera la négociation, qu'on espère de bonne foi (Blattberg 2004, 28, 34, 77). L'écueil pluraliste ici est de postuler une fragmentation des intérêts d'entrée de jeu, en ne faisant appel qu'au mode de dialogue de la négociation.

Il est vrai que les élus politiques ne peuvent s'attendre à ce que l'interprétation des biens en jeu dans le dialogue soit parfaite (Blattberg 2009, 231). Ils souscriraient alors à une vision moniste orthodoxe de la réalité, qui est mal adaptée au monde éthique et politique. Ce dernier ne répond pas à des standards de perfection, dans ses formes ou son contenu, comme je l'ai expliqué au chapitre précédent. La pratique du dialogue souffrirait d'un même formalisme et de décalages analogues avec le contexte et ses particularités. En revanche, il existe une approche au dialogue qui n'est ni moniste, ni pluraliste, et qui peut s'avérer intéressante de privilégier avant un recours à la négociation. Pour la comprendre, il importe de s'attarder à la distinction entre deux modes de dialogue : la négociation, mentionnée plus haut, et la conversation.

a) Négociation et conversation

Adaptée à la métaéthique pluraliste, la négociation, lorsqu'on la considère comme réussie, débouche sur un accommodement, un compromis (Blattberg 2004, 8, 28). En effet, d'expliquer Blattberg,

> la négociation est un aspect central du pluralisme, lequel suppose l'existence, dans le monde, d'une pluralité de valeurs et de façons de vivre parfois incommensurables qui, sans que quiconque n'en soit nécessairement responsable, entrent parfois en conflit. (2004, 28)

Au fond, étant donné cette incommensurabilité, la dynamique du dialogue ne peut être que celle d'un jeu à somme nulle (Blattberg 2004, 29) puisqu'il est impensable que tous en sortent gagnants sur tous les plans.

La conversation diffère de la négociation en ce que d'abord, sa vision de la communauté politique n'est pas fragmentée dans son essence de manière irrémédiable. Plutôt, elle renverrait à « […] un tout dont les parties ne constituent pas une pluralité d'éléments indépendants, mais une diversité de caractéristiques plus ou moins intégrées » (Blattberg 2004, 39). Dans ce contexte, les différences émergeant de cette diversité sont peut-être conciliables (2004, 41), voire *réconciliables.* Le dialogue sera animé par ce désir commun d'« […] arriver à un accord allant dans le sens de la réalisation du bien commun […] » (Blattberg 2004, 41). La conversation donc, contrairement à la négociation, présente des interlocuteurs qui ont le désir d'apprendre et se placent à l'écoute les uns

des autres, avec un effort de politesse (Blattberg 2004, 42),[58] car ils savent que d'apprendre, et de ce fait élargir leurs horizons, consiste en un gain. Plutôt que de chercher à minimiser les pertes, comme dans la négociation, les participants au dialogue visent à

> […] partager une même interprétation du sens et des implications réelles de la question dont ils discutent (même si, encore là, cette compréhension et ce partage ne se feront pas exactement de la même façon pour chacun d'entre eux). (Blattberg 2004, 41-42)

Conséquemment, il est aisé de voir que le mode de dialogue qui parvient à harmoniser tous les biens en jeu est la conversation, et que cette dernière permet de mieux réaliser le bien commun que la négociation ou, du moins, de manière plus entière. En effet, la négociation impliquera toujours des pertes et donc, pour des décideurs politiques confrontés à des enjeux éthiques, le fait de « se salir les mains ». On peut déjà entrevoir en quoi certaines métaéthiques et certains modes de dialogue peuvent présenter des affinités métaphysiques. On verra aussi dans les pages à suivre que la conversation est par ailleurs très fidèle à une compréhension herméneutique du dialogue. J'aimerais simplement proposer une illustration de ces notions avant de développer plus amplement la philosophie herméneutique.

b) Illustration : l'incidence de la métaéthique sur le mode de dialogue

Une démarche éthique élaborée à l'intention des décideurs politiques, comme le « Projet de recommandation sur l'éthique de l'intelligence artificielle de l'UNESCO » (2020, 5)[59] peut exiger une unité dans l'interprétation des éléments mis de l'avant. Le Groupe d'experts ad hoc (GEAH) de l'UNESCO, mis sur pied pour l'élaboration de ce document, affirme que

> la présente Recommandation doit s'entendre comme un tout, et les valeurs et principes fondamentaux comme étant complémentaires et interdépendants. Chaque principe doit être considéré dans le contexte des valeurs fondamentales. (Groupe d'experts ad hoc 2020, 21).

[58] En 2004, Blattberg parlait de « tact » : à présent, il lui préfère le terme « politesse ».
[59] À titre informatif, l'UNESCO a consulté plusieurs « parties prenantes » (en provenance de plus d'une cinquantaine de pays) dans le cadre de l'élaboration de ce rapport, dont le Mila et l'Algora Lab, de l'Université de Montréal (Dilhac et al. 2020, 4). Des membres de ces derniers ont élaboré leur contribution à la consultation (Dilhac et al. 2020).

Même si, dans le document, les rédacteurs reconnaissent l'éventualité de conflits de valeurs, la logique opérante est unifiante. Certes, l'UNESCO admet que la mise en pratique des dispositions de sa réflexion sera « […] conforme aux pratiques institutionnelles et aux structures de gouvernance de chaque État […] » (Groupe d'experts ad hoc 2020, 3), ce qui démontre un véritable souci de tenir compte des particularités du contexte dans la mise en application. Toutefois, le « cadre » et son contenu sont les mêmes pour tous et les États sont tenus de s'y conformer (2020, 6, 3).

On retrouve une logique similaire dans l'analyse que fait Ess du Règlement général sur la protection des données (RGPD) européen. Le cadre proposé est moniste (influencé par l'éthique de la vertu et l'éthique déontologique kantienne, explique-t-il), mais l'on reconnaît que les mises en application, dans la pratique, pourront être guidées par des interprétations diverses (Ess 2019, 81). Le fait demeure que le cadre est le même pour tous. En effet,

> bien que [le RGPD] serve de « cadre réglementaire *unique* pour tous les États membres », au moins certaines « législations nationales complémentaires sont autorisées », car « une marge de manœuvre discrétionnaire a été accordée et/ou préservée à cette fin ». [Traduction libre, je souligne] (Ess 2019, 82)

À mon sens, il s'agit d'une forme de monisme qui exhibe une certaine sensibilité à la diversité des contextes. De fait, même si on peut ajouter de la législation à celle qui existe déjà, on ne peut changer le cadre moniste en tant que tel. Cela fait aussi écho au pluralisme éthique *pros hen*, qui comprend un « […] engagement commun envers une norme, une valeur ou une ligne directrice fondamentale — mais l'interprétation ou l'application de cette norme diffère selon les contextes » [Traduction libre] (Ess 2020, 560). J'opère cette distinction sur la base de la nuance apportée par Blattberg (2018) entre le monisme orthodoxe et non orthodoxe, nuance peu employée dans la littérature, mais reprise dans ma réflexion. Ainsi, avec cette manière de penser, je verrais ici une forme de monisme plus orthodoxe que non orthodoxe, au sens où le contenu de ce qui est commun est déjà déterminé d'avance, au moyen de la valeur ou de la ligne directrice en question.

Conséquemment, dans le cas du projet de l'UNESCO, le « dialogue interculturel mondial » des « parties prenantes » (Groupe d'experts ad hoc 2020, 6) sera un mode de dialogue oscillant entre la théorie et la pratique, puisqu'on trouve, dans ce document comme dans bon nombre des

directives à l'étude pour cette thèse, des éléments métaéthiques en tension les uns avec les autres. Devant une telle directive, il est difficile pour les décideurs politiques de savoir comment s'y prendre concrètement. On leur présente une liste de valeurs et de principes, qui sont différents, mais qui doivent tous être respectés de manière à former une unité systématique. Sans question pour les guider, mais avec une certaine insistance sur l'importance des différents contextes étatiques, qui pourraient justifier des interprétations différentes des principes énoncés, et sans conscience que la réconciliation de toutes les valeurs pourrait échouer dans la pratique, il est en somme difficile de voir comment les élus devraient faire sens de telles recommandations, aussi profondes et recherchées soient-elles. De même, il est ardu de voir comment, par exemple, un comité parlementaire, formé de membres de différents partis politiques rivaux, pourrait arriver à une entente éthique satisfaisant le bien commun (le mieux possible), si cet objectif de viser un bien commun n'est pas explicité. Les intérêts partisans auront certainement une part dans le dialogue des décideurs, et il convient de voir comment les canaliser.

Quelque chose d'analogue s'observe dans la sphère privée. Les géants tels que Google, qui adoptent des principes éthiques assez généraux pour guider leur production, ne mentionnent pas explicitement que justement, « les […] principes sont ouverts à l'interprétation. Et ils sont contrôlés par les dirigeants qui doivent également protéger les intérêts financiers de l'entreprise » (Metz 2019, §15). Puisque les principes et valeurs éthiques devront être interprétés selon les situations et les personnes impliquées dans ces conversations, l'on devrait donc faire appel à la forme interprétative de la raison pratique. Gadamer (1982, 332) explique à juste titre que pour Aristote, « cela semble avoir été une évidence […] que […] le savoir en général ne peut prétendre être autonome et qu'il implique toujours d'être appliqué concrètement à un cas particulier ».

D'un autre côté, si la fonction publique se voit confier une directive comme celle du Groupe d'experts ad hoc de l'UNESCO pour sa mise en pratique, la question des intérêts partisans ne devrait plus se poser (au Canada, du moins). En revanche, la question de la mise en pratique des valeurs et principes énoncés demeure entière. Si la conversation permet que les biens en jeu, de même que leur compréhension, puissent être transformés, un cadre formel devient alors trop rigide pour guider l'action éthique. Si le cadre éthique lui-même ne laisse pas de place à sa transformation dans le dialogue, il lui manque la souplesse du monde de la pratique. Peut-être un document

éthique, tel que celui de l'UNESCO, se veut-il « neutraliste » en vue d'être « universel » : la réalité est qu'il est difficile, à mon sens impossible, d'atteindre de tels objectifs qualitatifs dans des domaines requérant l'usage de la raison pratique, qui fonctionne justement en étant pétrie de son contexte particulier.

C'est ainsi que le positionnement métaéthique de la démarche adoptée par rapport à l'IA aura une incidence sur le type de dialogue possible pour des décideurs politiques qui reçoivent des directives éthiques. Il s'agit de la raison pour laquelle les traditions éthiques invoquées dans les documents émis par les entreprises privées, la société civile, les organisations à multiples partenaires de même que les institutions internationales est une question à caractère politique. Cela explique aussi pourquoi je propose une approche alternative à celles qui ont été mises au jour dans la section « Portrait » de la thèse. Logiquement, les fondements éthiques que j'ai suggérés au chapitre 6 seront mis à contribution pour développer le mode de dialogue qui me semble optimal pour les décideurs politiques, soit un dialogue herméneutique. Le résultat sera une approche au dialogue éthique ancrée dans la question. Avant de détailler le contenu de cette approche alternative, il convient de préciser ce à quoi je fais référence par « herméneutique ».

2. L'herméneutique comme démarche dialogique

a) Brève esquisse d'une démarche herméneutique

L'herméneutique est, pour Hans Georg Gadamer — le philosophe contemporain qui l'a pensée avec le plus de rigueur — une expression de la philosophie pratique telle que la conçoit Aristote (Berti 2000, 347). C'est Gadamer qui a réconcilié l'éthique et l'herméneutique, en ce que sa philosophie représente précisément une « éthique herméneutique » (Tarantino 2017, 1, 10, 13). À l'origine, elle renvoyait à la manière d'interpréter des textes donnés, le plus souvent religieux ou légaux. Des philosophes du 20^{e} siècle — dont Martin Heidegger — en ont fait un courant philosophique à part entière. L'herméneutique devient ainsi une « doctrine de la vérité dans le domaine de l'interprétation » (Grondin 2010, 4) et, à mon sens, en sciences sociales, une *recherche* d'interprétations justes. Cette démarche philosophique présente plusieurs nuances, contient

différentes approches et a été employée dans toutes sortes de disciplines. C'est l'herméneutique gadamérienne qui a retenu mon attention en ce qu'elle consiste en une approche de la réalité plutôt que d'une méthode à proprement parler (Gadamer 1990, 11–15). En ce sens, elle s'éloigne des démarches positivistes et même post-positivistes en sciences humaines et sociales pour leur préférer des « philosophies de l'interprétation ». Par ailleurs, elle n'est pas procédurale, à la différence des théories éthiques comme l'éthique déontologique kantienne ou encore l'utilitarisme.

Il faut mentionner de prime abord que la notion d'« horizon » est centrale à la démarche herméneutique de Gadamer. Elle contribue à la situer dans une façon de penser que l'on pourrait qualifier de « holiste » (par opposition à « atomiste »). Gadamer reconnaît l'avoir empruntée à la phénoménologie d'Edmund Husserl (1990, 266). Dans cette façon de voir, chaque être humain entrant en dialogue avec un autre est forcément situé à l'intérieur de son « horizon de compréhension », une sorte d'arrière-plan préréflexif qu'il est impossible de formuler entièrement à l'aide de propositions théoriques ou de représentations, étant formé d'« habitudes et de coutumes » (Dreyfus 2014, 132). Cette façon de voir marque, entre autres, la pensée de Charles Taylor, ou encore celle de Hubert L. Dreyfus (se référer, par exemple, à Dreyfus et Taylor 2015; Dreyfus 1980, 7-8, 19). Ces horizons de compréhension permettent d'expliciter l'importance de la sensibilité profonde à tout contexte d'action, en ce que chaque personne voit, entend ou comprend les choses différemment à un moment, dans un endroit et à l'intérieur d'une culture donnés.

Évidemment, cet horizon de compréhension ne s'étire pas le long d'une ligne rigide ou fixe. Il s'agit plutôt de quelque chose qui voyage avec soi et qui est susceptible de s'élargir lors de fusions d'horizons avec d'autres interlocuteurs (Gadamer 1990, 266). Taylor affirme sans équivoque qu'« un horizon aux contours immuables est une abstraction » [Traduction libre] (2002, 136). L'élargissement de l'horizon de compréhension est souhaitable, surtout si l'on s'achemine tous, dans une communauté politique, vers un bien commun partagé, soit la vérité visée par des interprétations. C'est dans cet ordre d'idées que l'on peut dire que « notre activité a pour horizon la Polis et notre choix du faisable s'élargit […] au point de prendre place dans la totalité de notre être extérieur et social » (Gadamer 1982, 324). La « Polis » ou communauté politique, composée de ses différents membres, peut permettre de concilier toutes sortes d'intérêts représentés par des groupes différents, par une « conciliation délibérée » (Crick 1993, 17 — 19).

1. L'herméneutique comme démarche pratique : quelques ambiguïtés

Avant de poursuivre avec l'adoption de l'herméneutique comme approche au dialogue éthique, il importe de clarifier quelle place accorde Gadamer à la raison théorique aux côtés de la raison pratique. Puis, à l'intérieur de la raison pratique, comment il différencie la raison productive (à laquelle on associe la vertu de *techne*) et la raison prudentielle (à laquelle on associe la vertu de *phronesis*). Compte tenu de l'importance que j'ai accordée à la raison pratique à la base du raisonnement éthique, et étant donné l'impact de l'approche éthique sur le type de dialogue favorisé, je tenterai de clarifier ces notions pour que mon positionnement soit le plus limpide possible.

a. Raison théorique versus raison pratique

Gadamer affirme clairement que l'herméneutique s'oppose à la théorie « moderne », qui serait en porte à faux avec « l'application pratique » (Gadamer 1982, 312). Selon lui,

> la réflexion morale philosophique que renferme l'activité philosophique propre à l'éthique n'est pas une théorie qu'il faut mener jusqu'à l'application pratique. Elle n'est absolument pas un savoir en général, un savoir à distance qui ne pourrait que dissimuler l'exigence concrète de la situation […]. (1982, 325).

Cela étant dit, Gadamer reconnaît une « universalité du concept » dans « […] la conscience tout à fait non théorique, moyenne et générale, de la norme, dans la réflexion pratico-morale de chacun », que ce soit pour chaque personne prise individuellement, ou encore pour « […] l'homme d'État qui agit pour tous » (Gadamer 1982, 325). L'universel entre dans le particulier dans l'herméneutique gadamérienne, et suit en cela de près l'éthique aristotélicienne, qui « […] a assigné pour tâche centrale, à l'éthique philosophique comme au comportement moral, la concrétisation de l'universel et son application à chaque situation » (Gadamer 1982, 326).

En éthique politique, on l'a vu au chapitre précédent, il en va différemment que dans une éthique personnelle. À mon sens, il est plus aisé pour une personne de s'orienter éthiquement grâce

à un « universel » (que cela soit une théorie éthique ou une série de principes ou de valeurs jugées universelles), et de l'appliquer concrètement aux situations de sa vie. En politique, étant donné la multiplicité des acteurs et des points de vue valables sur une question, l'éthique politique requiert une approche différente. Pour Gadamer, certainement, un appel à la raison spéculative, en plus de la raison pratique, interviendra dans la pratique individuelle.

C'est peut-être ce que Gadamer entend lorsqu'il ne conçoit pas la même opposition entre théorie moderne (positiviste) et pratique, qu'entre la *theoria* d'Aristote et la *praxis*, la première étant en réalité « une praxis suprême » (Gadamer 1982, 311-312). Cette façon de voir corroborerait l'idée selon laquelle l'*eudaimonia* aristotélicienne comprend à la fois la contemplation du bien, mais aussi l'exercice des vertus dans la pratique (Roochnik 2009, 70). Chez Aristote, la *theoria* ne renvoie pas seulement à la contemplation de ce qui est immuable. Elle peut aussi intervenir dans la pratique, pour « théoriser ce qui est contingent », par exemple (Roochnik 2009, 73), non au sens d'une théorie moderne, mais au sens large du raisonnement appliqué à ce qui est contingent. La *theoria* est une manière de raisonner, et non une théorie positiviste formulée comme telle. En ce sens, on pourrait dire que « […] nous sommes par nature théoriques, que nous "vérifiions" une situation pratique ou les étoiles » [Traduction libre] (Roochnik 2009, 78). C'est donc l'objet de notre réflexion qui déterminera à quelle forme de la raison nous faisons appel.

Pour une communauté politique donnée dialoguant sur le bien commun, la démarche herméneutique doit être plus pratique que théorique, en ce sens que le raisonnement est mené collectivement. Un individu aura un accès différent à la *theoria* que deux ou plusieurs personnes l'auront dans leur conversation, puisque chacun arrive avec son horizon de compréhension et que la conversation elle-même peut faire changer ces horizons en cours de route. Il est certain que de ce point de vue, le dialogue est — ou devrait être — plus riche que la réflexion personnelle. Une manière de le comprendre est de justement se pencher davantage sur cette notion d'horizons de compréhension, qui fait de l'herméneutique gadamérienne une approche holiste du réel. Peut-être qu'il sera plus clair, à la suite de cette digression, de voir en quoi je reproche à Gadamer une sorte d'ambiguïté, de flou non entièrement résolu entre la raison théorique et la raison pratique, qui pose problème quand vient le temps de dialoguer au sujet d'enjeux éthiques politiques. Même si Gadamer (selon la lecture que Georgia Warnke fait de sa vision de l'éthique) désire vraiment faire

comprendre que son approche est éminemment pratique et non théorique (Warnke 2002, 82), le doute peut parfois s'installer chez le lecteur de ses œuvres. Gadamer reconnaît lui-même le fait « qu'un mot comme herméneutique oscille entre un sens pratique et un sens théorique […] » (1982, 330).

Hubert L. Dreyfus, qui a approfondi cette question, voit en Gadamer un holiste théorique plutôt que pratique, en ce qu'il ne parvient pas totalement à distinguer la théorie de la pratique, même s'il affirme le faire (Dreyfus 2014, 134). C'est entre autres dans la notion de « croyance », ou de « préjugé », centrale à l'herméneutique gadamérienne, que l'on peut retracer cette confusion. Selon l'interprétation de Dreyfus, les horizons de compréhension, dans la pensée de Gadamer, seraient plus près de croyances (qui peuvent être explicitées en propositions) que des pratiques ou habitudes (Dreyfus 2014, 134-135). Gadamer parlerait d'« idées normatives dans lesquelles nous avons été élevées et qui sont au fondement de l'ordre social » (Gadamer 1982, 347).

Il n'est pas faux de dire qu'il y a des idées explicites dans nos horizons de compréhension, mais ces derniers ne sont pas formés *que* de propositions explicites. En ce sens, sans souscrire entièrement à l'idée que Gadamer réduit les horizons de compréhension à des croyances, l'idée de rendre ces horizons explicites relève effectivement de la raison théorique. L'holisme pratique implique que le contexte est indissociable de la réflexion — dans ce cas-ci, la réflexion éthique. Les pluralistes (holistes) diraient que les valeurs ne peuvent être isolées hors de leur contexte, pensées à la manière des sciences naturelles qui s'en détachent pour les observer. J'y reviendrai plus bas, mais on voit clairement ici en quoi une approche herméneutique s'inscrivant dans un holisme *pratique* est toute indiquée pour une démarche éthique profondément sensible au contexte.

C'est dans l'optique où Gadamer s'inspire d'Aristote pour penser l'éthique comme une discipline pratique que je désire m'inspirer pour ma proposition, en prenant ce qui, dans sa pensée, s'arrime bien à la raison pratique prudentielle. Ainsi, la « connaissance éthique » n'est pas

> […] la connaissance objective que les observateurs ont des relations nécessaires ou constantes entre les objets. Au contraire, les situations auxquelles l'acteur doit répondre de manière éthique varient; elles sont multiples et substantiellement uniques. […] *La connaissance éthique n'est pas une connaissance qu'un spécialiste ou un théoricien peut découvrir pour les autres une fois pour toutes*; elle n'est pas la même chose qu'une théorie du bien ou qu'un récit d'un universel séparé et immuable […]. Une théorie du

> bien a plutôt un statut provisoire dans le schéma d'Aristote. Elle peut faire référence à une liste de vertus, mais elle ne le fait que comme une sorte de guide pour l'action. [Traduction libre, je souligne] (Warnke 2002, 82-83).

Conséquemment, c'est d'une acception pratique de l'herméneutique que je compte m'inspirer pour ma proposition aux décideurs politiques, mais pratique « prudentielle », et non « phronétique ». J'ai expliqué comment, au chapitre 6, ma conception de la vertu de prudence n'est pas entièrement aristotélicienne en ce qu'elle ne présuppose pas une cible entièrement prédéterminée, contrairement à la *phronesis*. Là se trouve peut-être une autre différence entre l'herméneutique de Gadamer et l'usage que je veux en faire, à savoir que Gadamer admet que toute sa philosophie « n'est que *phronesis* » (Gadamer dans Tarantino 2017, iv). Il faut admettre qu'il est possible qu'il entendît par là son interprétation de la *phronesis*, puisque pour lui, tout est interprétation. Si tel était le cas, nous aurions simplement des compréhensions un peu différentes de cette vertu.

Une dernière clarification est nécessaire avant d'exposer les détails de ma proposition, parce que la raison pratique elle-même peut porter à confusion, lorsqu'on la tire de la pensée aristotélicienne. À mon sens, dans certains passages des écrits de Gadamer, on retrouve parfois une distinction très claire entre la production, qui requiert la vertu intellectuelle de *techne*, et l'action, qualifiée par la vertu intellectuelle de *phronesis*. À d'autres endroits, ce contraste s'estompe, pouvant de ce fait donner l'impression que l'éthique (ou encore n'importe quel champ de la philosophie pratique) peut être rapprochée d'une compétence technique.

b. Raison pratique : techne versus phronesis

On l'a vu, Gadamer affirme que, comme l'éthique, l'herméneutique est un savoir pratique (1982, 311). Cela étant dit, à l'intérieur même de la raison pratique, Aristote établit une autre distinction, en disant que « les choses qui peuvent être autres qu'elles ne sont comprennent à la fois les choses qu'on fabrique et les actions qu'on accomplit. Production et action sont distinctes […] » (Aristote 2014, Livre VI, 1140a1-5). La production, comme l'action, chez Aristote, concerne ce qui peut ou ne pas être. Toutefois, à la différence de l'action, son « […] principe d'existence réside dans l'artiste et non dans la chose produite […] » (Aristote 2014, Livre VI, 1140a5-15). À l'activité de l'artisan ou de l'artiste, Aristote associe ainsi la vertu de *techne* (2014, Livre VI, 1139b-1140a).

Gadamer présente parfois l'herméneutique comme un « art », une « technique » au sens aristotélicien, et parle également de l'« art politique » qui « […] se révèle être celui de tisser, de manière à assembler les contraires en unité » (Gadamer 1982, 334-335, 336). Il affirme aussi que « la philosophie pratique ne se limite pas à un domaine particulier. Elle est certes capable de mettre au point des méthodes (ou plus exactement des règles de base) et son art, une fois maîtrisé, peut être porté à la perfection » (Gadamer 1982, 347). Dans son œuvre *Vérité et méthode*, Gadamer soutient qu'« il y a bien une correspondance véritable entre la perfection de la conscience morale et la perfection du pouvoir-faire qu'est la *techne*, mais il est clair qu'elles ne sont pas identiques » (Gadamer 1990, 337).

Si l'herméneutique est « employée » pour guider le dialogue des décideurs politiques sur des enjeux éthiques, ce dialogue doit être compris comme une action plutôt que comme une production. À cet égard, je me distancie quelque peu de l'herméneutique de Gadamer. En effet, les méthodes ou les « règles de base » qu'évoque Gadamer auront pour effet de fausser les biens inhérents à la pratique en voulant les sortir de leur contexte pour les ajuster à la théorie ou aux règles guidant la « production éthique ». La conversation elle-même sera déformée. Plus encore, l'idée d'une perfection du dialogue est inadéquate, quand on sait que les limites des acteurs et du monde de la pratique génèrent justement des tensions, des incompréhensions, voire des dilemmes insolubles. Ce serait donc plus dans le sens où l'herméneutique est une pratique prudentielle, en ce qu'elle est et conduit à l'action (et non à la production), que je l'adopte. Cette pratique est le lieu de rencontre de la pensée et de la parole, où les deux ne deviennent qu'une même chose (Blattberg 2009, 244). L'action dont il est question est le dialogue des élus politiques sur des questions éthiques, celles touchant aux enjeux posés par le développement et l'adoption de l'intelligence artificielle. Ce dialogue devrait être animé, à mon sens, par de bonnes questions.

b) L'herméneutique et la question

Une des forces de l'approche herméneutique à l'éthique et à la politique est de faire prendre conscience aux agents de leur « […] profonde dépendance à l'égard de réalités culturelles qui ne sont pas de [leur] fait » [Traduction libre] (Davey 2006, 9). Gadamer parle de la « priorité

herméneutique de la question » dans le dialogue, une « primauté qui fonde le concept de savoir » (1990, 388). En effet, exprime-t-il, « on ne fait pas d'expérience si on ne se met pas à questionner », et donc « seul a du savoir celui qui a des questions », puisqu'on acquiert ce dernier par la dialectique (Gadamer 1990, 385, 388). Conséquemment, un dialogue herméneutique devrait favoriser la question, dans le cadre d'une conversation alimentée par « la dialectique de la question et de la réponse » (Gadamer 1990, 410). Il est intéressant de noter que Gadamer lui-même précise que « l'art de questionner […] n'est même pas un art au sens où les Grecs parlent de *techne*, un pouvoir qui puisse s'enseigner et qui permettrait de s'emparer de la connaissance de la vérité » (Gadamer 1990, 390). La question se rattacherait à l'autre branche de la raison pratique, la branche prudentielle qui guide l'action.

Interroger permet de « particulariser » la réflexion éthique à la situation donnée ainsi qu'à son contexte (les personnes impliquées, les biens en jeu, la culture ou le moment historique, par exemple). C'est ce qui permet à Gadamer de dire que

> ce n'est pas dans les concepts universels de courage et de justice, etc., que s'accomplit le savoir moral, mais, au contraire, dans l'application concrète de ce qui détermine, à la lumière de ce savoir, ce qui est faisable ici et maintenant. (1982, 322)

Cela étant dit, il ne s'agit pas de poser n'importe quelle question, simplement dans le but d'alimenter une conversation. La question rhétorique est sans intérêt ici (Gadamer 1990, 387). L'interrogation est évidemment dirigée dans le sens de la discussion, elle est, autrement dit, orientée. On peut affirmer que « le sens de la question est donc la direction dans laquelle seule peut s'effectuer la réponse, si elle veut être une réponse sensée et pertinente » (Gadamer 1990, 385). Il s'agit du « sens de ce qui est juste », de ce qui est vrai et conséquemment, ce qui « correspond nécessairement à la direction frayée par une question » (1990, 387). La question devrait orienter l'action, elle devrait faire déboucher sur une action politique, et non sur la contemplation d'une réalité ou d'une idée, si elle se veut fidèle à la raison pratique prudentielle. En revanche, il ne faudrait pas tomber dans l'écueil opposé, soit celui de suivre à la lettre une liste de questions sans permettre que la conversation en fasse émerger d'autres.

Le fait de procéder en posant des questions comprend l'avantage de permettre de faire face à des dilemmes qui se présentent comme étant insolubles, ou même d'en faire émerger. Gadamer

estime que « ce qui fait le sens de l'interrogation, c'est qu'elle découvre tout ce qu'a de problématique ce que l'on interroge » (1990, 386). Plus encore, une telle démarche ouvre la porte à la réalité que la raison pratique (prudentielle) a encore un mot à dire devant le conflit, qu'elle n'est pas condamnée au décisionnisme (Berlin et Williams 1994). Concrètement, une manière de s'y prendre pourrait être de poser un dilemme, et de présenter des réponses possibles. C'est la façon de procéder qu'a retenue Charles Ess dans sa réflexion éthique sur les médias numériques, par exemple (2009, 170-171). L'idée de rechercher toutes les réponses possibles à une question, en plus de la poser, est de faire valoir des contenus que l'on aura jugés non pertinents au premier abord, sans les avoir d'abord bien entendus. Gadamer abonde dans ce sens lorsqu'il affirme qu'« être en dialogue, ce n'est pas réduire l'autre au silence par l'argumentation, c'est au contraire déterminer le poids réel de son opinion (1990, 390). La personne qui questionne se mettra en quête de « […] tous les arguments favorables à telle ou telle opinion. La dialectique ne consiste pas à trouver la faiblesse de ce qui est dit, mais à commencer […] par lui donner sa véritable force » (Gadamer 1990, 390-391).

Un exemple notable d'approche éthique procédant par questions est celle élaborée par aline shakti franzke pour l'Université d'Utrecht, le DEDA (Data Ethics Decision Aid for Researchers) (franzke et al. 2020, 8). Le DEDA vise à assister « […] les analystes de données, les chefs de projets et les décideurs à reconnaître les questions éthiques dans les projets de données, la gestion des données et les politiques en matière de données » [Traduction libre] (Utrecht Data School, Utrecht University, s.d.b, s.p.). Au moyen d'une spirale de questions classées par catégories d'enjeux et principes, en plus d'une application, plus de 230 questions sont pensées pour faire face à des enjeux éthiques touchant à la recherche Internet (franzke et al. 2020, 8; Utrecht Data School, Utrecht University s.d.a, s.p.; *Ibid*. s.d.c., s.p.). Les questions visent évidemment à orienter vers l'action (Utrecht Data School, Utrecht University s.d.a, s.p.).

De plus, Shannon Vallor, dans la « boîte à outils » qu'elle développe au profit des ingénieurs et concepteurs de SIA (mentionnée dans les chapitres un et quatre), énumère elle aussi une liste de questions que ces derniers devraient se poser, comme « Quels résultats voulons-nous obtenir grâce à cet outil? Quels risques voulons-nous que son utilisation atténue ou diminue? », ou encore « Quels intérêts, désirs, compétences, expériences et valeurs avons-nous simplement

supposés, plutôt que réellement consultés? Pourquoi avons-nous fait cela, et avec quelle justification? » [Traduction libre] (Vallor 2018, 2, 9). Bent Flyvbjerg, dans sa proposition pour une « science sociale phronétique », procède aussi en listant quatre questions, en précisant que

> […] les questions sont posées en admettant qu'il n'y a pas de « nous » unifié par rapport auquel les questions peuvent recevoir une réponse finale. Les chercheurs phronétiques ne considèrent *aucun terrain comme étant neutre*, *aucune vue comme émanant de « nulle part »,* pour leur travail. […] personne n'est assez expérimenté pour donner des réponses *complètes* aux quatre questions […]. Ce à quoi il faut s'attendre, cependant, ce sont des *tentatives* de la part des chercheurs en sciences sociales phronétiques pour développer leurs réponses partielles aux questions; ces réponses seraient un apport au dialogue social en cours sur les problèmes et les risques auxquels nous sommes confrontés et sur la façon dont les choses peuvent être faites différemment. [Traduction libre, je souligne] (Flyvbjerg 2001, 60-61)

Par ailleurs, le fait d'avoir des questions pour orienter la conversation et, par la suite, les actions qui en découleront, peut aider les décideurs politiques à se centrer sur les vrais enjeux éthiques que pose l'intelligence artificielle et les innovations dans ce domaine, et non sur des problèmes imaginaires créés par le sensationnalisme (Müller 2020, §3), malheureusement assez récurrents dans le domaine de la réflexion sociale sur l'IA.

En conséquence, ce que l'herméneutique apporte de crucial à la pratique du dialogue, c'est la nécessité d'écouter. Blattberg affirme à juste titre que « […] nous devons écouter, plutôt que de *chercher à voir*, la vérité » [Traduction libre, je souligne] (2009, 231). Cela renvoie à la nécessité d'écouter non seulement les questions, mais les réponses proposées. Les questions permettent aux personnes engagées dans le dialogue de « nouer avec le contexte », de ne pas chercher à s'en abstraire, mais plutôt de s'engager directement lui (Blattberg 2009, 243). La sensibilité au contexte est l'un des quatre éléments que j'ai présentés comme formant la base d'une approche éthique optimale. Cette sensibilité est encouragée, entre autres, par la vertu intellectuelle de la prudence, qui implique aussi une orientation téléologique « douce » vers un bien commun. Toutefois, il faut aussi tenir compte de la possibilité que malgré nos efforts, des compromis doivent être effectués en cours de route. C'est ainsi que dans la seconde section de ce chapitre, je mets en relation « l'herméneutique de la question » avec une approche éthique pour penser les enjeux de l'IA.

3. Une approche alternative ancrée dans la question

Les quatre éléments qui fondent ma proposition éthique alternative sont ici séparés pour assurer la clarté, mais sont en réalité indissociables. Ils ne forment pas un cadre formel ni une théorie systématique avec des composantes potentiellement distinctes, indépendantes les unes des autres. Pris isolément, aucun de mes éléments n'a totalement de sens. C'est entre autres parce que ma façon de concevoir la réalité pratique est holiste, plutôt qu'atomiste, que cette séparation de mes éléments me semble factice; je l'opère néanmoins pour ancrer la rigueur de l'analyse. Dans un sens, ils s'apparentent aux idéaux types de Max Weber, chacun étant « une construction *analytique* unifiée [qui] dans sa pureté conceptuelle […] ne peut être trouvée empiriquement nulle part dans la réalité » [Traduction libre] (Weber 1949, 90).

Les questions qui émergent de ma réflexion sont alimentées certes par ma réflexion éthique, mais aussi par mes lectures sur l'IA. Un des effets de ces recherches est que parfois, je constatais plusieurs points communs entre des questions qui surgissaient dans ma réflexion, et des questions ou points de réflexion que des auteurs avaient mis de l'avant eux-mêmes. En ce sens, elles ne surgissent pas de nulle part. Cependant, le fait que des questions que je me suis posées, en pensant au dialogue des décideurs politiques sur les enjeux éthiques de l'IA, aient trouvé un écho dans des travaux de penseurs et éthiciens de l'IA m'apparaît comme un signe de leur pertinence. Il faut préciser que ces auteurs n'évoluent pas dans la logique herméneutique que je privilégie, ou encore dans la compréhension — critique — de l'éthique que j'ai mise de l'avant dans cette thèse. Je mentionne ces penseurs par souci de transparence et les cite lorsque nos idées convergent, peu en importe l'origine, mais il n'en demeure pas moins que ma manière de procéder diffère plus ou moins des leurs, tout en exhibant certains points communs.

Une courte liste de ces sources inclurait des réflexions explicitées sous forme de questions, comme celles pose Flyvbjerg (2001) pour approcher de manière optimale les travaux en sciences sociales, celles qu'adresse Vallor (2018) aux concepteurs et développeurs de technologies, celles que propose le Ethical Explorer (s.d.b), du Omidyar Network, de même que celles que met de l'avant Mittelstadt pour évaluer les directives éthiques concernant l'IA (2019, 15) et finalement, les questions que propose Greene (2018) à l'intention des gouvernements municipaux. Dès le début

de ma réflexion sur l'intelligence artificielle, j'ai incorporé les cinq idées concernant la technologie avancées par Neil Postman (1998), dans le but de voir si ces idées pouvaient être transformées en questions, puis servir de base à une proposition éthique alternative.

De plus, au fil de mes recherches, quelques réflexions de Floridi (2019) et de Floridi et al. (2018a, 2018b) sur l'éthique de l'IA et ses défis, la distinction des enjeux à court et long terme que proposent Prunkl et Whittlestone (2020), l'analyse croisée de directives éthiques de Hagendorff (2019) et celle de Boddington sur la déclaration d'Asilomar (2017), de même que la place des enjeux politiques en éthique de l'IA que détaille Calo (2017), sont venues s'ajouter et enrichir le tout. Ces travaux ont tous nourri mon travail de questionnement quant à l'éthique politique de l'IA, d'une manière difficile à préciser ou mesurer avec exactitude. Parfois, c'était en amont, d'autres fois, en aval, quand je retrouvais chez eux des questions que j'avais déjà explicitées. Dans d'autres cas encore, j'ai cherché à voir les points communs de nos questionnements et à les fusionner. C'est la raison pour laquelle les questions que je propose devraient sembler assez familières à un lecteur averti de l'éthique de l'IA.

a) Un dialogue herméneutique prudent

Le dialogue des décideurs politiques constituera un exercice de sagesse pratique s'il vise le bien commun, s'il s'effectue (ou essaie de s'effectuer) en synergie avec le contexte et enfin, s'il admet que des dilemmes insolubles pourraient se présenter. La vertu intellectuelle de prudence (telle que je l'ai présentée au chapitre précédent) n'est, en aucune façon, un synonyme de perfection. Comme il ne s'agit pas de suivre une méthode (Gadamer 1982, 333), il importe de mener le raisonnement en déterminant les composantes de la cible (le bien commun) et les meilleurs moyens d'y parvenir : cette raison ordonnée à la finalité dans l'action est ce qu'on appelle la prudence. Ainsi, les deux éléments suivants de ma proposition, soit la sensibilité profonde au contexte et une orientation téléologique douce vers un bien commun, font partie intégrante de la vertu intellectuelle de prudence, et c'est en cela qu'il est difficile de les séparer absolument. Le dialogue des décideurs politiques, qui sera guidé par la logique de la primauté herméneutique de la question, traitera, dans chaque situation où un enjeu éthique concernant les SIA, de la détermination du contenu du bien commun et des moyens de l'atteindre.

Des questions faisant appel à la vertu intellectuelle de prudence porteraient sur les finalités, tant celles des SIA (Vallor 2018, 13) que celles des décideurs politiques dans leur dialogue. Elles inviteraient à une conversation au sujet des moyens à mettre en place pour atteindre ces finalités et permettre leur évaluation (Vallor 2018, 2; Ethical Explorer s.d.b, 9). D'autres pourraient porter sur les personnes ou entités à l'origine des directives éthiques ainsi que leurs intérêts (Mittelstadt 2019, 15). Dans l'optique des décideurs politiques, d'autres questions devraient faire surgir les priorités politiques de l'heure et y décider de la place des enjeux liés à l'IA (Floridi et al. 2018b, 514). Puis, il conviendra, pour des élus politiques prudents, de s'informer des personnes et entités touchées par le développement de certains SIA ou programmes (Vallor 2018, 9). La vertu de prudence est indispensable tout au long de la démarche herméneutique des décideurs politiques, puisqu'à la base, elle permet d'obtenir l'information nécessaire à la réflexion et au dialogue.

b) Un dialogue herméneutique orienté vers un bien commun

On l'a vu au chapitre précédent, le bien commun n'est pas un objet fixe et rigide, inflexible. Plutôt, selon Bernard Crick, il s'agit du

> […] processus de conciliation pratique des intérêts des diverses « sciences », agrégats ou groupes qui composent un État; il ne s'agit pas d'un adhésif spirituel externe et intangible, ni d'une prétendue « volonté générale » ou « intérêt public » objectif. Ce sont là des explications trompeuses et prétentieuses de la cohésion d'une communauté; pire, elles peuvent même justifier la destruction soudaine de certains éléments de la communauté au profit d'autres — il n'y a pas de droit d'entraver la volonté générale, dit-on. [Traduction libre] (Crick 1993, 24)

On comprend facilement le rôle du dialogue des décideurs politiques entre eux, représentants de ces différents intérêts et porteurs du désir d'un bien « pour tous », dans une telle conception du bien commun. Le dialogue n'implique pas de devoir se départir de ses aspirations. En effet,

> réaliser le bien commun […] ne peut être une question de *transcendance* des différences, de consensus sur une même chose, détachée, car comprendre — selon ce que nous enseigne l'herméneutique — c'est toujours comprendre *différemment*. [Traduction libre] (Blattberg 2009, 34)

Ainsi, lorsque les élus politiques converseront, ils ne tenteront pas de se dépouiller de leurs façons de voir, mais d'englober celles des autres : autrement dit, d'élargir la leur sans la déformer, et d'atteindre ensemble la vérité. C'est ce que Gadamer (1990, 409, 412, 419) comprend comme la fusion des horizons. Non seulement il n'est pas souhaitable de se défaire de sa façon de voir ou d'entendre, mais il ne s'agit pas non plus d'exécuter « transfert de soi vers l'autre » (Gadamer 1990, 405, 419), qui constitue l'idée centrale de l'herméneutique romantique.

C'est la vérité (qui est le bien commun) qui constitue le sujet de la conversation, ce « *logos* » « […] qui n'est ni le mien ni le tien, et dépasse donc l'opinion subjective des interlocuteurs […] » (Gadamer 1990, 391). Cela devient en revanche un « nôtre », quelque chose qui « se comprend ensemble » (Taylor 1997a, 138-139). Même si l'interprétation du bien commun, dans le dialogue des décideurs politiques, « […] doit s'accommoder de la situation herméneutique à laquelle elle appartient » (Gadamer 1990, 420), il n'en découle pas que cette interprétation devient relativiste. Les questions devraient permettre un acheminement commun vers la vérité. Aristote soutient à ce propos que

> […] les hommes sont naturellement aptes à recevoir une notion suffisante de la vérité; la plupart du temps ils réussissent à la saisir. Aussi, à l'homme en état de discerner sûrement le plausible, il appartient également de reconnaître la vérité. (Aristote s.d.b, Livre 1, 1, XI)

Charles Ess tient un propos semblable en parlant de l'exercice du jugement au moyen de la *phronesis*, qui tend à « s'autocorriger », et qui doit être pratiquée (2020, 562, 565).

Orientées vers le bien commun politique, les questions que je proposerais pourraient porter sur les besoins de la communauté politique, mais aussi aider à dévoiler ceux de l'économie (Postman 1986; Vallor 2018, 13; Boddington 2017, 56). La distinction entre les finalités à court et à long terme (Prunkl et Whittlestone 2020) d'un SIA ou d'un programme gagnerait à être explorée, de même que leur caractère « désirable » (Flyvbjerg 2001, 145). Dans un angle plus négatif, les risques et inconvénients prévisibles devraient être mis au jour (Greene 2018), ainsi que les « gagnants et les perdants » des changements qui auront lieu (Postman 1998, 1-3; Flyvbjerg 2001, 145; Ethical Explorer s.d.b, 7). La place et les bénéfices du secteur privé, de même que leur engagement éthique, sont au centre de telles questions (Boddington 2017, 54, 110; Floridi 2019,

188; Calo 2017, 6-7). L'information nécessaire pour dialoguer sur l'enjeu précis qui occupera les élus politiques, ainsi que la disponibilité de cette information, sont d'autres points d'interrogation pertinents (Ethical Explorer s.d.b, 3, 13, 11; Vallor 2018, 2; Ethical Explorer s.d.a, 11). En somme, il s'agit de données concrètes de la situation qui permettent un regard plus avisé sur le contenu du bien commun.

c) Un dialogue herméneutique profondément sensible au contexte

Comme je l'ai mentionné à plusieurs reprises, le seul fait de poser des questions exhibe et permet une plus grande sensibilité au contexte. Parce que, dans le cas de l'IA en particulier, « la politique technologique, à proprement parler, est difficile à planifier et à appliquer » [Traduction libre] (Müller 2020, §1.3). L'éthique et la politique n'étant pas des sciences naturelles, le contexte n'est pas externe à la « science », mais bien interne. En fait, « cela découle du fait que la théorie réussit à décontextualiser alors que les sciences humaines doivent s'occuper du contexte humain » [Traduction libre] (Dreyfus 2014, 142). Ce contexte « […] est si vaste et si profond qu'il ne peut être question de le suspendre simplement et d'agir en dehors de celui-ci. Le suspendre totalement serait ne rien comprendre du tout à l'être humain », exprime Taylor [Traduction libre] (2002, 131). Une approche herméneutique reconnaissant la primauté de la question permet de tenir compte de ce contexte, de l'incorporer totalement à la délibération et conséquemment, à la prise de décision qui s'ensuivra, et qui contiendra déjà des liens multiples au monde de la pratique. Flyvbjerg pourrait d'ailleurs sembler favoriser une approche par questions lorsqu'il écrit que

> […] le but des sciences sociales n'est pas de développer des théories, mais de contribuer à la rationalité pratique de la société *en élucidant où nous sommes, où nous voulons aller et ce qui est souhaitable en fonction de divers ensembles de valeurs et d'intérêts*. L'objectif de l'approche phronétique devient celui de contribuer à la capacité de la société à délibérer et à agir de manière rationnelle en fonction des valeurs. [Traduction libre, je souligne] (2001, 167-168)

Le contexte inclut les « doctrines politiques » à l'œuvre, les éléments de partisanerie qui influencent évidemment le jugement de chaque acteur politique, qui sont elles-mêmes tributaires de leur contexte (Crick 1993, 22). Gadamer admet en effet que « […] l'action politique et pratique de l'homme est déterminée par l'acteur et son savoir » (1982, 332). Le jugement est également

central à la démarche, puisque la morale et la politique, on l'a vu, ne fonctionnent pas comme les mathématiques : « nous ne pouvons pas retirer l'agent délibérant de l'éthique et de la politique, réduisant la *politikê* à une application passive [et procédurale] de principes universels à des circonstances particulières » [Traduction libre] (Abizadeh 2002, 270). Suivant cette logique, non seulement les questions — qui peuvent changer —, mais les réponses seront influencées par leur contexte d'émergence. Il va sans dire que les questions que je propose ici sont teintées par ma perspective éthique moniste non orthodoxe. Cela étant dit, malgré l'influence des doctrines politiques à l'œuvre chez les interlocuteurs, ce sont ces derniers qu'il faut tenter d'écouter, non pas les idéologies qui prédisent le contenu des interventions.

La question devrait faire déboucher les interlocuteurs vers l'action. De toute évidence, cette dernière sera particulière et concrète, non universelle et abstraite. La question contribue, au moins en partie, à pallier un problème tant décrié dans la littérature en éthique de l'IA, à savoir le niveau d'abstraction des listes de principes et de valeurs.[60] En effet,

> les déclarations reposant sur des concepts normatifs vagues cachent des points de conflit politique et éthique. Les concepts d'« équité », de « dignité », d'« épanouissement » et autres concepts abstraits de ce type sont des exemples de « concepts essentiellement contestés » ou de concepts dont la signification peut être conflictuelle et *qui nécessitent une interprétation contextuelle en fonction des convictions politiques et philosophiques de chacun*. Ces différentes interprétations, qui peuvent être rationnellement et sincèrement soutenues, conduisent à des exigences substantiellement différentes *dans la pratique*. [Traduction libre, je souligne] (Mittelstadt 2019, 5)

Une approche par questions, sensibles au contexte, devient un guide de l'action (*action-guiding*), moyennant la formulation de réponses : cela permet d'éviter le risque du relativisme moral (Mittelstadt 2019, 5). L'approche dialogique que je prône n'établit pas de distinction nette entre la réflexion et la pratique éthiques. Elles font partie de la même démarche et relèvent toutes deux de la raison pratique prudentielle. Idéalement, elles permettront d'arriver à ce que Flyvbjerg appelle « un riche dialogue en contexte » (2001, 139-140).

[60] Cela étant dit, on trouve parfois aussi des questions posées dans les directives éthiques analysées dans le cadre de cette thèse. Cependant l'approche ne se veut pas une approche par questions, pour les raisons que j'expose dans cette thèse, ni spécifiquement aux décideurs politiques en premier lieu. Par exemple, on peut retrouver des principes et des valeurs aux côtés de questions dans des directives comme celle du IEEE (2016).

Enfin, d'un point de vue plus « technique », il convient de mentionner que la sensibilité au contexte peut vouloir dire de modifier les questions selon les différentes étapes de développement et de déploiement d'un SIA. Bien que plusieurs des questions seront de nature technique et donc adressées aux ingénieurs et experts en conception (par exemple la boîte à outils de Vallor 2018), une conscience des différents enjeux au fil du développement d'un SIA peut être d'un grand secours pour les décideurs politiques et ceux qui participent au dialogue. Par exemple, Morley et al. (2019) ont cherché à voir comment opérationnaliser les principes éthiques qui foisonnent dans la littérature grise (en les réduisant préalablement à cinq) lors de sept étapes de développement d'un SIA, qu'ils identifient comme 1) « le développement des affaires et des cas d'utilisation », 2) « la phase de conception », 3) « la formation et l'acquisition des données d'essai », 4) « la construction de l'IA », 5) « le test de l'IA », 6) « le déploiement de l'IA » et 7) « le suivi des performances » [Traduction libre] (Morley et al. 2019, 7).

Un autre exemple se trouve dans l'« Ethical Explorer Pack » (l'« Explorateur éthique »), qui se veut un guide de questions pour les « responsables de la production, les designers, les ingénieurs ou encore les fondateurs » (Ethical Explorer s.d.a), qui vise à « soutenir les valeurs humaines », « créer une culture du questionnement » et « susciter le changement par le dialogue » (Ethical Explorer 2020, s.p.; Ethical Explorer s.d.a, 3). (Pierre Omidyar, le fondateur de eBay, à l'origine du Omidyar Network.) Les questions sont posées dans le cadre de possibles scénarios, par exemple :

> nous lançons une nouvelle fonctionnalité qui recommande des magasins en fonction de la localisation de nos utilisateurs. Favorisons-nous certains magasins en particulier? Quelles sont les protections de la vie privée en place? [Traduction libre] (Ethical Explorer s.d.a, 9)

Il s'agit d'un outil de questions pour les ingénieurs, mais plusieurs éléments peuvent être repris (parfois quelque peu amendés) pour les décideurs politiques.

Dans l'« Explorateur éthique », les types de risques que peut représenter une technologie sont catégorisés et se voient assigner des questions. Le format est celui d'un ensemble de cartes, un côté visant la compréhension de l'enjeu, et l'autre la possibilité de poser des questions en profondeur (Ethical Explorer s.d.b, 1). Par exemple, pour ce qui touche à la surveillance, on

demande : « Comment protégerons-nous la vie privée? » et plusieurs autres sous-questions plus précises, comme « est-ce ainsi que notre technologie est destinée à être utilisée? », ou encore « que pourraient faire les gouvernements ou organisations tierces avec les informations que nous recueillons? » [Traduction libre] (Ethical Explorer s.d.b, 2-3). D'autres enjeux tels que la désinformation, l'exclusion, les biais algorithmiques, la dépendance, le contrôle des données, les mauvais acteurs et le pouvoir démesuré sont listés (Ethical Explorer s.d.b, 4-17), et on peut même ajouter le sien.

Le fait de favoriser une approche par questions pour faciliter le dialogue est tout à fait semblable à ce que je propose. Cependant, cet outil n'est pas accompagné par une réflexion sur la métaéthique et sur le type de dialogue à favoriser. Des négociations comportant des pertes pourraient très bien être vues comme le type d'échange optimal dans le contexte des innovations technologiques. Cela serait plausible puisque l'« Explorateur éthique » m'apparaît comme évoluant dans la logique du pluralisme des valeurs, avec l'accent mis sur les valeurs et leur mise en balance, les compromis de valeurs et les différentes parties prenantes (par exemple, Ethical Explorer s.d.b, 17).

Dans le but que les décideurs politiques exhibent une profonde sensibilité au contexte dans leur dialogue, il serait bien de soulever des questions permettant d'identifier les différents acteurs et biens en jeu aux décideurs politiques. Autrement formulé, il s'agirait de se demander qui devraient être les interlocuteurs de ce dialogue (Floridi et al. 2018b, 524-525). Il s'agit évidemment de groupes au sein de la société, mais aussi des différents paliers et ordres gouvernementaux. Ce type de questionnements ouvrira sur celui de la responsabilité : si certaines entités ont des intérêts dans l'enjeu qui préoccupe les élus politiques, il convient de déterminer qui doit en porter la responsabilité (Floridi 2019, 191; Vallor 2018, 2) et qui en sera redevable.

d) La possibilité de dilemmes insolubles par le dialogue herméneutique

Les conflits sont inévitables quand il est question de politique. En effet, « plus on est impliqué dans des relations avec d'autres, plus il y aura de conflits d'intérêts, ou de caractère et de circonstance » [Traduction libre] (Crick 1993, 25). Ce qui importe principalement est la manière

d'y répondre et j'ai tâché de démontrer que de favoriser un dialogue visant un bien commun partagé, avec prudence et sensibilité au contexte, était la meilleure avenue possible pour les élus en politique. Cela étant dit, même avec une approche éthique herméneutique favorisant la question, des dilemmes peuvent s'avérer impossibles à résoudre sans subir quelque perte que ce soit. Il s'agit de véritables « problèmes », dont Gadamer retrace l'étymologie chez Aristote, comme étant

> […] les questions qui se présentent comme des alternatives en suspens, parce que des arguments de toute sorte plaident en faveur d'une possibilité comme de l'autre, et que nous ne croyons pas être en mesurer de trancher en invoquant des raisons décisives, parce qu'il s'agit de questions trop vastes (Gadamer 1990, 400)

ou même, de questions trop sensibles, voire carrément insolubles. Par exemple, la notion d'équité, en tant que concept « complexe et profondément contextuel » (Madaio et al. 2020, 15, 2), peut être comprise de plusieurs manières différentes pouvant s'opposer à quelques niveaux. Plus encore, « accorder la priorité à l'équité dans les systèmes d'IA signifie souvent faire des compromis en fonction des priorités concurrentes. Il est donc important d'être explicité et transparent sur les priorités et les hypothèses » (Madaio et al. 2020, 15). Plus largement, d'expliquer Müller, « la politique en matière d'IA pourrait entrer en conflit avec d'autres objectifs de la politique technologique ou de la politique en général » [Traduction libre] (2020, §1.3).

La raison pratique a quelque chose à dire dans des situations de dilemmes, justement en posant des questions. Même si, d'après les analyses de Neil Postman, « […] tout changement technologique constitue un compromis [et que] […] la technologie donne et la technologie reprend » [Traduction libre] (1998, 1), il est possible et souhaitable pour les décideurs politiques d'entretenir un dialogue sur ces conflits. Certes, dans quelques cas, l'on ne pourra éviter la compromission de valeurs (les « mains sales »), mais au moins, les décideurs politiques auront abordé le problème de manière optimale et profité de l'exercice qui, à coup sûr, aura été une source d'apprentissage. Le fait que les biens que l'on cherche à promouvoir entrent en conflit avec d'autres ne devrait pas générer de la peur, ou faire tomber dans le relativisme, puisque les conflits ne remettent pas en question la pertinence de ces biens et de ces valeurs (Taylor 1997b). Au contraire, ils la confirment. Nørskov et Rodogno ne voient pas, dans de telles situations, des raisons de sombrer dans un « pessimisme absolu » : plutôt, suggèrent-ils, « […] le fait que nous n'ayons pas

de réponses claires maintenant n'exclut pas que nous puissions, par des processus de dialogue, nous mettre d'accord sur certaines réponses à un moment ultérieur » [Traduction libre] (s.d., 10).

Comme dans les autres sous-sections, je propose ici un autre ensemble d'interrogations à l'intention des décideurs politiques. Cela étant dit, comme l'affirme Greene, je suis d'avis que

> les décideurs peuvent choisir un ensemble différent de questions et de facteurs de risque ou les pondérer différemment de ce que nous avons fait ici, et ils peuvent arriver à des conclusions différentes sur le risque éthique d'une application d'IA. Ces variations sont des sous-produits prévisibles des différences politiques et culturelles, mais le plus important est [de commencer] à poser ces questions et à réfléchir soigneusement aux risques potentiels de l'IA. [Traduction libre] (2018, s.p.)

Les questions qui permettront de mettre au jour la possible insolubilité de certains dilemmes seront certes délicates. Elles toucheront évidemment aux biens et aux intérêts en jeu et leur compatibilité pratique avérée (Ethical Explorer s.d.b, 1. 13, 15), et les procédures décisionnelles en cas de conflit de valeurs. En effet, Mittelstadt reconnaît que des interprétations (des biens, des intérêts, des besoins, des finalités ou encore des moyens) mises de l'avant pourront être conflictuelles (2019, 15). De même, la question des effets des SIA, programmes ou politiques adoptés devra être spécialement soignée, tant dans la réflexion sur l'aspect prévisible (Ethical Explorer s.d.b, 5) ou définitif des changements causés (Postman 1998, 4), que dans la redéfinition du contexte que ces derniers peuvent amener (Vallor 2018, 9).

4. Ma proposition : un parcours herméneutique formé de questions

En définitive, la philosophie herméneutique de Gadamer, qui propose une primauté de la question dans le dialogue, me semble tout indiquée comme approche face au défi immense que pose l'intelligence artificielle aux décideurs politiques. J'ai exposé en quoi une telle démarche s'harmonise bien avec les quatre éléments éthiques que j'ai mis de l'avant au chapitre dernier et quelles interrogations pourraient en émerger. À présent, j'aimerais présenter une série de questions aux décideurs politiques, qui pourront aussi servir à ceux qui les assistent, par exemple dans la fonction publique. Mon but, avec le « Parcours herméneutique », est de guider leur dialogue sur n'importe quel enjeu éthique de l'IA par de bonnes questions. Ces dernières permettront, en

principe, l'exercice de la prudence, la sensibilité au contexte, tout en visant le bien commun et en reconnaissant que parfois, des compromis devront être effectués en cours de route. Comme je l'ai exposé un peu plus haut, si on évite ces compromis en arrivant à une interprétation partagée des biens en jeu et de la finalité visée, l'on saura que les élus auront *conversé* avec succès. Si la dynamique du dialogue a dû requérir des accommodements, elle relevait de celle d'un jeu à somme nulle, et l'on saura que les décideurs politiques auront *négocié* pour minimiser les pertes, mais qu'au final, ils auront dû « se salir les mains », du moins dans une certaine mesure.

Ma proposition aux décideurs politiques (le « Parcours herméneutique ») se trouve un peu plus bas. La Figure 1 donne une idée globale du parcours grâce à l'illustré. Les questions qui suivent sont le détail de ce parcours que j'invite les législateurs à emprunter, avec l'ordre proposé. Le nombre d'étapes ou d'escales dépendra, comme le lecteur le verra, de l'état et du contenu du dialogue, ainsi que de la situation. Mon objectif, en m'adressant aux décideurs politiques, est de leur présenter des questions à caractère politique, en ayant fait en amont le travail métaéthique de prise en compte du continuum distinguant le monisme du pluralisme. Ce travail de « défrichage métaéthique » est en somme la thèse, en particulier dans ses deux premières sections.

a) L'élaboration du parcours herméneutique

Les étapes et les questions que je propose sont le résultat du croisement entre les interrogations élaborées dans ce chapitre et une série de cinq dialogues (fictifs) de décideurs politiques tous azimuts. Cet exercice complémentaire à la thèse a été mené sur des enjeux de l'IA aussi variés que la sécurité internationale, les systèmes autonomes, l'éthique des machines, l'attribution de la responsabilité, la surveillance et les biais et discriminations des SIA, tels que décrits au chapitre premier. Mon but était de présenter et de structurer, pour les législateurs, des questions que j'avais déjà plus ou moins cernées. Il m'apparaissait optimal de leur fournir un point de départ de même qu'un point d'arrivée.[61] De même, il était manifeste qu'au point d'arrivée, les

[61] En ce sens ici aussi, un parallèle peut être esquissé entre la démarche de mon parcours herméneutique et la spirale du DEDA mentionnée précédemment (Utrecht Data School, Utrecht University s.d.c, s.p.), sans que je m'en sois inspirée au préalable pour élaborer le parcours.

décideurs seraient face à une réconciliation ou un accommodement par rapport à l'enjeu d'IA traité. De plus, je désirais que ce guide ne soit pas inflexible.

La récurrence de certaines questions au fil des dialogues, voire de quelques nœuds auxquels se heurtaient les interlocuteurs supposés, a permis de circonscrire non seulement les questions, mais les étapes du parcours entre les points de départ et d'arrivée. Ainsi, après chaque dialogue, j'ai extrait les questions récurrentes. J'ai remarqué, par exemple, que la définition très précise de l'enjeu était absolument incontournable, tant au plan politique que technologique. Pour ce faire, des informations sont nécessaires, dans la législation existante de même que dans les expériences accumulées au fil des années dans des situations analogues. La question de l'accès à cette information prend également un certain relief. Bref, définir l'enjeu et l'état des choses dans lesquels les décideurs politiques se trouvent au moment de leur discussion renvoie précisément à la vertu intellectuelle de prudence, et fait appel à la sensibilité profonde au contexte. Il s'agit de prendre le pouls de la réalité concrète du terrain, plutôt que de plaquer un principe déterminé sur ce dernier. Cette ouverture au contexte fait aussi place au besoin de cerner ce qui est différent dans l'enjeu concret, ainsi que de nouvelles données qui peuvent fluctuer au cours du processus de dialogue et de légiférération.

D'autres points communs aux sujets de conversations distinctes ont surgi. La place et le rôle du gouvernement en est un, qui se rattache ensuite aux questions de représentation et de responsabilité. La finalité de la discussion est clarifiée, si elle ne l'était pas d'emblée. De même, les intérêts et les biens en jeu se doivent d'être déclarés au cours de la conversation, si l'on recherche leur harmonisation ou une tentative dans ce sens. Les citoyens seront forcément touchés par les technologies et les décisions de leurs gouvernants. Il importe par ailleurs de tenir compte des multiples points de vue sur des défis et opportunités concrets que posent les SIA. De ce fait, la question de la définition des interlocuteurs devenait saillante. On l'a vu, les incidences positives et négatives sont mentionnées à répétition dans la littérature. Dans le cas du parcours herméneutique, elles revêtent la même importance. Plus encore, il existe la possibilité que l'enjeu tel que circonscrit en début de dialogue puisse sembler tellement énorme qu'il devienne judicieux de le scinder en sous-enjeux. C'est donc pour cette raison qu'une des questions tient compte de cette éventualité.

J'ai rassemblé ces questions récurrentes au sein des dialogues dans des catégories : la définition de l'enjeu, de l'état des choses, de la finalité, du rôle du gouvernement ainsi que des intérêts et biens impliqués. Ces catégories sont devenues les escales sur le parcours. J'ai ensuite tracé des parallèles entre ces catégories de questions et les quatre éléments à la base de mon approche éthique : la prudence, qui implique la sensibilité au contexte, une orientation téléologique « douce » vers le bien commun, mais aussi la reconnaissance de dilemmes potentiellement insolubles. De cette façon, un lien demeurait entre la métaéthique et l'approche au dialogue mise de l'avant. J'ai par la suite inscrit ces groupes de questions dans une continuité « temporelle », dans une sorte de chemin dialogique à suivre pour arriver à la décision politique. Il m'apparaissait important d'inclure des possibilités de retour en arrière[62], pour éviter que ce parcours ne souffre de la rigidité que je reproche aux démarches éthiques procédurales. Le chemin se devait de garder la flexibilité et l'imperfection du dialogue.

b) Comment emprunter le « parcours herméneutique »?

Peut-être faut-il spécifier que le but du « parcours herméneutique » n'est pas de rendre les décideurs sensibles à des questions d'ordre psychologique pendant leur discussion, comme ce que des sociologues comme Erving Goffman appellent « l'analyse de la conversation » (voir Goffman 1983 dans Heritage 2001). Il s'agit bien plutôt d'une démarche normative pour des décisions politiques portant sur le contenu de la discussion, à savoir un problème concret lié à l'éthique de l'IA. Donc, le parcours débute avec la définition de l'enjeu dont il est question, par exemple le SIA ou programme qui est amené, ou la question générale qui se pose sur le plan de l'éthique. Le degré d'urgence, l'information disponible, les priorités de l'heure et les dynamiques sociétales révélées par cet enjeu font partie de l'exploration à ce premier arrêt. La deuxième étape du parcours est un premier exercice de précision de la démarche, à savoir la détermination claire du contexte concret et actuel qui entoure l'enjeu défini à la première escale. Si les décideurs

[62] Dans le contexte d'une démarche inspirée par la philosophie herméneutique, la possibilité de revenir en arrière n'est pas sans rappeler le « cercle herméneutique ». Gadamer explique que le cercle peut faire référence à un retour « aux choses elles-mêmes », à la possibilité que la décision prise (ou la compréhension donnée) soit révocable (1990, 287-288). En effet, suggère-t-il, « [q]uiconque cherche à comprendre est exposé aux erreurs suscitées par des pré-opinions qui ne résistent pas à l'épreuve des choses mêmes » (Gadamer 1990, 288). C'est parce que nos interprétations sont précisément situées dans un contexte qu'elles peuvent être révisées, enrichies (Warnke 2002, 80-81). Je remercie Charles Ess pour cette suggestion bienvenue.

politiques ne savent pas, par exemple, quelle législation existe par rapport au SIA identifié, ni son potentiel ou ses manquements, il sera difficile de déterminer ce qui doit être fait. Les ressources et la prise de position sur l'enjeu sont aussi des questions qui sont posées à cette étape. La prudence, qui implique la sensibilité au contexte fait partie des éléments métaéthiques principaux qui sont mobilisés et auxquels on fera appel lors de ces deux premiers arrêts dialogiques.

Après avoir franchi les deux premières escales du parcours, les interlocuteurs peuvent choisir d'aller à la troisième ou à la quatrième étape, selon ce qu'ils jugeront qui sera le plus fructueux dans leur dialogue. La troisième escale leur permettra de déterminer les biens en jeu, qui ne sont pas nécessairement incompatibles à la base, mais qui pourraient se révéler l'être dans la pratique. C'est ici que l'ouverture à l'insolubilité de certains dilemmes s'ajoute à la prudence et à la sensibilité au contexte pour nourrir un dialogue optimal. Les questions devraient outiller les décideurs à cerner les acteurs, les points de vue, les biens, les intérêts, ainsi que les préjudices et bénéfices qui pourraient entrer en ligne de compte dans le cas qu'ils auront circonscrit. Ils seront également confrontés au besoin de déterminer une stratégie d'action en cas de conflits de ces données de la situation, ou de ces biens.

La quatrième étape du parcours vise à déterminer la cible, soit la finalité du dialogue des décideurs politiques. Une précision s'impose ici, à la lumière de l'analyse métaéthique proposée dans la thèse. Il est évident que les interlocuteurs devraient avoir une idée générale de la finalité de leur dialogue avant de l'entreprendre, ou même de se rendre à la quatrième étape. S'ils l'avaient probablement en tête avant de commencer à dialoguer donc, le contexte, il importera à ce moment de l'expliciter clairement, de manière à ce que l'on puisse savoir quel type de politique attendre de leur dialogue. Il s'agit d'un exercice de précision de la finalité, à la lumière de ce qui aura été discuté auparavant.

Néanmoins, il faut aussi rappeler que, dans une démarche prudentielle inspirée (mais non entièrement calquée) d'Aristote, la fin et les moyens ne sont pas séparés de façon manichéenne. L'on ne retrouve pas une distinction typiquement moderne entre les deux, comme on peut l'observer dans une démarche utilitariste. Les fins et les moyens seront nécessairement interconnectés, de manière difficile à exprimer hors du contexte ou de la culture dans lesquels les

décideurs se trouveront au moment de leur dialogue. Par exemple, si ces derniers cherchent à aborder l'enjeu d'IA avec prudence — ce que le parcours herméneutique les enjoint à faire — la prudence est considérée ici à la fois comme une finalité, mais également comme un moyen de l'atteindre.

Puis, la conversation sur la ou les finalités devrait être orientée « téléologiquement », de manière « douce », vers le bien commun de la communauté politique. Les personnes ou entités à consulter, la distribution de la responsabilité et l'évaluation des finalités font partie des questions que les décideurs auront à se poser à cette escale, ce qui leur permettra, une fois de plus, d'aborder leur dialogue avec prudence. Ces quatre étapes franchies, il sera plus facile pour les interlocuteurs de déterminer précisément leur rôle, ainsi que celui d'autres potentiels acteurs, dans l'enjeu qui fait l'objet de leur dialogue. Ils seront invités, à la cinquième étape, à voir s'il serait avisé de diviser l'enjeu et, si cela s'avérait plus réaliste, il pourra être judicieux de revenir par le quatrième point pour s'assurer que la finalité de l'enjeu redéfini, ou mieux circonscrit, est claire. Enfin, c'est à cette étape qu'il leur sera plus aisé de voir si des compromis ou des accommodements seront nécessaires (ce qui implique des pertes), ou si, au contraire, une harmonisation des intérêts préalablement cernés est envisageable.

La dernière étape du parcours est celle de l'adoption ou de la formalisation d'une entente sur la base de l'issue du dialogue (compromis et/ou réconciliation), puis l'invitation à sa mise en application. Cela étant dit, j'ai inclus dans le parcours la possibilité, après la mise en application de la résolution, quelle que soit sa forme (règlement, loi, création d'une entité, amendement d'un texte existant, etc.) de revenir à la deuxième étape pour reprendre le dialogue. Cela pourrait s'avérer nécessaire, par exemple, si l'adoption des mesures identifiées par le gouvernement entraîne des changements importants dans la société, qui font que le contexte dans lequel la technologie ou le SIA avait été pensé ne cadre plus avec la réalité des choses. Évidemment, ce retour à la seconde escale peut se faire à n'importe quel moment jugé nécessaire, pendant le parcours herméneutique. Les interlocuteurs pourraient aussi entièrement reconsidérer l'ordre des escales, selon les données et informations découvertes pendant leur conversation, bien entendu.

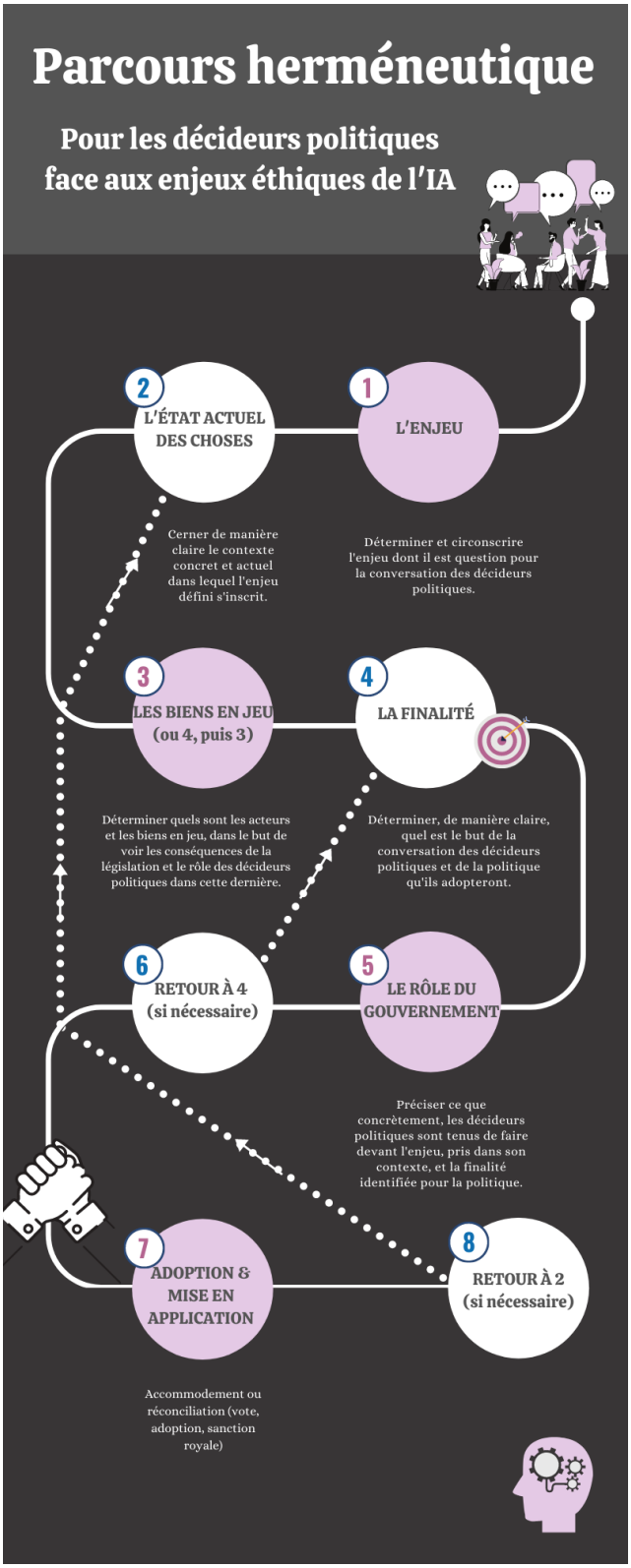
Si les interlocuteurs de ce dialogue suivent la progression que je propose, ils rencontreront certes des questions plus larges que d'autres, mais, en bout de piste, ces dernières devraient leur permettre de circonscrire l'enjeu qui les occupe de plus en plus précisément et donc de cerner quel est leur rôle concret face à cet enjeu. De plus, je pense qu'un tel parcours permettra de réduire les tensions métaéthiques que j'ai documentées tout au long de la thèse, pour élaborer une politique souhaitable face à la multiplicité des enjeux qui sont générés actuellement par l'IA. Ainsi, à chaque escale du parcours, je propose aux décideurs entre cinq et dix questions, comprenant pour la plupart différents volets. Ils n'ont pas à y répondre s'ils ne le désirent pas, ou si, dans le contexte de la conversation qu'ils tiennent, elles ne leur semblent plus pertinentes. De même, d'autres interrogations peuvent tout à fait être soulevées, selon les réponses fournies (et j'espère qu'elles le seront). De même, l'ordre que je propose est flexible et des occasions de revenir sur leurs pas sont clairement explicitées dans le guide de questions. Il faut donc comprendre ma proposition comme une suggestion, que j'estime valide, mais certainement pas absolue.

Comme toute démarche prudentielle en éthique ou en politique, l'issue de la réflexion et du dialogue devrait mener à l'action, à la mise en pratique de ce qui a été discuté. C'est là où le parcours herméneutique s'arrête, puisque c'est là où la politique publique commence. Ce que je m'étais proposé, pour répondre à ma seconde question de recherche, était de mettre de l'avant l'approche éthique à privilégier pour un dialogue optimal des décideurs politiques face aux enjeux de l'IA. Je crois que ce but est atteint avec ces deux documents. J'ai par ailleurs effectué un « test » (individuel) de ce parcours avec un enjeu éthique concret touchant à l'IA (la dissémination des fausses nouvelles sur les réseaux sociaux, en contexte de pandémie et d'élections américaines) et je crois que le modèle, malgré toute la subjectivité que comprend cette affirmation, joue son rôle de manière satisfaisante. Cela étant dit, il ne s'agit pas d'une proposition finale et rigide. Des propositions pour l'amender permettront de l'améliorer. Ce dont je suis convaincue, c'est que si les décideurs politiques empruntent réellement ce parcours herméneutique, ils pourraient, par leur expérience, grandement bonifier mon modèle et le compléter, de manière à ce qu'il reflète davantage la réalité pratique.

Enfin, le lecteur se doute bien que le fonctionnement optimal d'un tel guide à la conversation dépend de la collaboration de ceux qui l'emploient. S'il est compatible avec

l'inclusion des intérêts politiques, économiques ou sociétaux des participants au dialogue, il va sans dire qu'un recours cynique à ce guide sera sans doute décevant. De même, certains pourraient objecter que des acteurs pourraient mal employer des guides éthiques mis à leur disposition, dans le but d'en faire mauvais usage. Il m'apparaît, comme Ess, qu'il n'y a pas là matière suffisante pour les rejeter (2020, 565). Je l'ai mentionné dans les chapitres précédents : je ne crois pas que l'éthique politique puisse *exiger* des acteurs politiques la vertu. J'estime toutefois qu'une ouverture sincère à la compréhension de l'autre (Taylor 2002, 126) ainsi qu'un désir profond d'atteindre le bien commun politique seront les meilleures garanties du succès de ces parcours herméneutiques. La partisanerie politique pourrait faire en sorte que ce ne soit pas dans l'intérêt de tous que la conversation aboutisse à une harmonisation des biens. Néanmoins, c'est un risque à prendre. Après tout, le monde de la pratique n'est pas parfait, mais on peut œuvrer dans le sens de son changement pour le mieux.

Figure 1. – Parcours herméneutique illustré



Questions pour le parcours herméneutique des décideurs politiques

Destinataires : Les décideurs politiques et ceux qui les assistent (en comité, en caucus, en commission, en réunion, ou encore dans la fonction publique).
Objectif : Guider le dialogue concernant les enjeux relevant de l'éthique de l'IA.
Conseils :

1. Prendre connaissance du parcours en entier au moyen de l'illustré (Figure 1.).
2. Suivre le parcours dans l'ordre pour un résultat optimal. Il est toutefois possible de revenir sur ses pas à tout moment de la conversation.
3. Les astérisques (*) sont facultatifs et relatifs à l'état de la conversation.
4. D'autres questions peuvent surgir au cours de la conversation; toutes n'ont pas besoin d'être répondues nécessairement.
5. Ceci n'est pas un manuel d'exigences. Les points 1 – 4 ci-dessus reflètent plutôt quelque chose à lire avant de s'engager avec d'autres dans un dialogue ouvert destiné à répondre à des questions comme celles ci-dessous. La lecture devrait *contribuer* (sans plus) à préparer le dialogue et à le guider au fur et à mesure qu'il se déroule. L'objectif est d'aider les interlocuteurs à se faire une idée de ce à quoi s'attendent, à réfléchir aux questions auxquelles il importerait de répondre en premier, à ce qui pourrait survenir et comment on pourrait y répondre, etc. La conversation est un mode de dialogue extrêmement fragile, donc tout doit être fait pour aider à réussir avant de devoir l'abandonner et de passer à la négociation.

1. L'ENJEU

Objectif : Déterminer et circonscrire l'enjeu dont il est question pour la conversation des décideurs politiques. Ces définitions et précisions concernent aussi bien le SIA que le type de politique à adopter à son égard.

Élément(s) de mon approche métaéthique mobilisé(s) : 1) la prudence, qui entraîne 2) la sensibilité profonde au contexte

Questions proposées :

a. Quel est le but de notre conversation?
b. De quel SIA est-il question ici? Quel est l'enjeu concret?
c. Comment formuler en une question l'enjeu qui nous occupe?
d. Est-ce que tous les termes de la question sont clairs, ou devons-nous obtenir des définitions?
e. Quel est le degré d'urgence de cet enjeu, ou quelle est sa place sur notre échelle de priorités actuelles?
f. Qui sont les experts liés à cet enjeu? A-t-on obtenu leur avis sur l'enjeu en question?
g. Qui développe ce SIA? Dans quel(s) but(s)?
h. À quelle information avons-nous accès? À quelle information la population a-t-elle accès?
i. Est-ce que cet enjeu révèle un problème sociétal plus large que le problème circonscrit ici?

2. L'ÉTAT ACTUEL DES CHOSES

Objectif : Cerner de manière claire le contexte concret et actuel dans lequel l'enjeu maintenant déterminé s'inscrit. Il s'agit du contexte politique et technologique.

Élément(s) de mon approche métaéthique mobilisé(s) : 1) la prudence, qui implique 2) la sensibilité profonde au contexte

Questions proposées :

a. Quelles sont les priorités politiques actuelles? Commentaire : cette question aura peut-être été trouvé réponse au point 1.e.
b. Quel est l'état de la législation à l'égard de l'enjeu identifié, si législation il y a? Existe-t-il des enjeux similaires auxquels on peut se référer pour en examiner la législation et en tirer des expériences, voire apprendre d'erreurs passées?
c. La législation existante est-elle suffisante? Sinon, que faut-il ajouter ou changer? Quel est l'élément de « rupture technologique » justifiant ce changement?
d. Quel est notre point de vue sur cet enjeu?
e. Quelles sont nos ressources? Que peut-on obtenir de plus, au besoin?

Commentaire : Les décideurs peuvent choisir d'aller à l'étape 4 avant d'aller à l'étape 3, s'ils estiment que le fait de clarifier la finalité de la démarche peut aider à cerner les biens en jeu.

3. LES BIENS EN JEU

Objectif : Maintenant que l'enjeu et le contexte politique sont clarifiés, c'est vers les acteurs et les biens en jeu qu'il faut se tourner, dans le but de voir les conséquences de la législation et le rôle des décideurs politiques dans la législation.

Élément(s) de mon approche métaéthique mobilisé(s) : 1) l'ouverture aux dilemmes insolubles, 2) la sensibilité profonde au contexte et 3) l'orientation vers le bien commun, 2 et 3 faisant partie de 4) la prudence

Questions proposées :

a. Qui sont les acteurs (personnes, entités) touchés par cet enjeu?
b. Parvenons-nous à comprendre les points de vue de ces acteurs, surtout s'ils proviennent d'origines interdisciplinaires?
c. Quels sont les biens valorisés par ces acteurs que cette situation fait ressortir (par exemple l'autonomie, l'égalité, le droit à la vie privée, le progrès technologique, la sécurité)? Est-ce que nous les comprenons tous de la même façon? Sinon, comment?
d. Quels sont nos intérêts (partisans, électoraux et économiques surtout) dans cette situation? Ces intérêts sont-ils mieux servis en échouant à nous réconcilier avec nos opposants?
e. Quels sont les intérêts du secteur privé dans cette situation et quelle est leur relation avec le secteur public?
f. Peut-on prévoir des préjudices à certains des acteurs ou biens identifiés avec nos projets politiques? De quelle nature? Seraient-ils irréversibles? Peut-on les prévenir?
g. Peut-on prévoir des bénéfices à certains des acteurs ou biens identifiés avec nos projets politiques? De quelle nature? Sont-ils rendus publics?
h. Quelle stratégie prévoyons-nous adopter dans le cas de conflits de biens en jeu? Pouvons-nous envisager une compréhension commune et enrichie de ces biens? Sinon, quel accommodement viser?

4. LA FINALITÉ

Objectif : À présent que les décideurs politiques connaissent l'enjeu politique et technologique, le contexte dans lequel ils évoluent ainsi que les acteurs et les biens qui le caractérisent, ils peuvent (re)définir de manière claire le but de leur conversation et de la politique qu'ils cherchent à adopter.

Élément(s) de mon approche métaéthique mobilisé(s) : 1) l'orientation téléologique douce vers un bien commun, comprise dans 2) la prudence

Questions proposées :

a. Quelle est la finalité des SIA ou instruments technologiques en question ici?
b. Quel est le but de notre conversation? (Revenir sur la réponse à la question 1.a. et la préciser.) Vers quelle politique désirons-nous qu'elle aboutisse (règlement, amendement ou rédaction d'une loi, politique publique, mise sur pied d'un organisme)?
c. Qui est-ce que cette politique vise?
d. Quel est notre échéancier pour la développer?
e. Qui devons-nous consulter? Commentaire : Si les décideurs ont choisi d'aborder les questions de l'étape 4 avant celles de l'étape 3, ce serait un bon moment ici de consulter celles du point 3.
f. Qui est responsable de mener à bien cette/ces finalité(s)? Commentaire : cette question amène déjà celles de l'étape 5. On peut décider de suspendre notre réponse pour la reposer à la cinquième étape.
g. Comment évaluer si notre finalité/nos finalités seront atteintes?
h. Quelles sont les conséquences liées à notre politique que nous pouvons déjà anticiper? Commentaire : cette question se répond mieux si on a répondu aux questions de l'étape 3 préalablement.

5. LE RÔLE DU GOUVERNEMENT

Objectif : Préciser encore davantage ce que les décideurs politiques sont tenus de faire concrètement devant l'enjeu, pris dans son contexte, et la finalité identifiée pour la politique à adopter.

Élément(s) de mon approche métaéthique mobilisé(s) : 1) l'orientation téléologique vers un bien commun, qui fait partie de 2) la prudence

Questions proposées :

a. Quel(s) palier(s) de gouvernement la politique identifiée concerne-t-elle et nécessite-t-elle? Commentaire : cette question est formulée en tenant compte du contexte canadien, mais elle inclut aussi les gouvernements municipaux, qui peuvent se positionner quant aux dispositifs des villes intelligentes, par exemple.
b. Quel est le rôle des conseillers et acteurs politiques non élus dans la poursuite de la finalité identifiée (par exemple la fonction publique)?
c. Quels moyens pouvons-nous établir pour mener à bien les finalités identifiées?
d. Quelle est la responsabilité concrète du gouvernement dans cette politique et quelle est la responsabilité des autres acteurs?
e. Peut-on régler tous les problèmes que nous percevons? Sinon, a-t-on élaboré un plan à court et à long terme?
f. Devrions-nous diviser cet enjeu en plusieurs parties et nous y pencher séparément? Commentaire : si la réponse à cette question est « oui », il pourrait être intéressant de revenir à l'étape 4 pour redéfinir la finalité de l'enjeu plus circonscrit ou visant le court terme.
g. Que devons-nous aux citoyens par rapport à cet enjeu? Que pourrait-on nous reprocher si on le négligeait?
h. Est-il possible d'éviter de « se salir les mains » dans cet enjeu? Autrement dit, est-ce que nous sommes capables de trouver une solution qui harmonise les biens identifiés à l'étape 3 avec la/les finalité(s) de l'étape 4? Sinon, comment trancher?
i. Si la conversation (impliquant l'harmonisation de nos biens) échouait, comment pourrions-nous les hiérarchiser, et quels accommodements pourrions-nous proposer? Des idéologies politiques spécifiques (libéralisme, féminisme, écologisme) peuvent-elles nous aider dans notre exercice de hiérarchisation de nos biens?
j. Quelles sont les modalités de la mise en application de notre politique?

6. *RETOUR À LA FINALITÉ (4)

Commentaire : cette étape est facultative et dépend des réponses données à l'étape 5. Une fois le retour à l'étape 4 terminé, il serait bien de poursuivre avec l'étape 5, car elle fait déboucher la délibération politique sur l'action.

7. ADOPTION ET MISE EN APPLICATION

Objectif : Formaliser une entente issue de la conversation guidée par les étapes 1 à 7.

Élément(s) de mon approche métaéthique mobilisé(s) : 1) la prudence, qui implique 2) une téléologie « douce » vers un bien commun, tout en manifestant une 3) ouverture à des dilemmes insolubles.

Modalités : Selon la réponse à la question 5.h., les décideurs seront placés devant une possibilité d'accommodement ou de réconciliation, en adoptant la politique identifiée au point 4. La réponse à la question 5.i. les orientera vers la mise en application de la politique adoptée.

8. *RETOUR À L'ÉTAT DES CHOSES (2)

Commentaire : ce point est proposé dans le cas où le contexte ou les informations sur ce dernier (étapes 1 à 3 surtout) venaient à changer de manière importante pendant le parcours herméneutique des décideurs politiques, ou après l'adoption de leur politique et sa mise en pratique.

Si le contexte politique, technologique ou sociétal change de manière si importante que cela justifie une nouvelle conversation sur la politique, on pourrait reprendre le parcours à l'étape 2. Cela peut se faire à n'importe quel moment pendant le parcours. L'idée est la même si la politique adoptée fait émerger de nouveaux enjeux imprévus, de manière à poursuivre le parcours jusqu'à la cinquième étape. Ce changement holiste serait plausible, car la transformation du sens moral qu'entraîne la conversation ouvrira à de nouvelles perspectives, ainsi qu'à des problèmes inédits.

Conclusion

On l'a vu au chapitre dernier : deux questions de recherche ont orienté le travail présenté ici. La première, « Quelle(s) tradition(s) ou approche(s) éthique(s) informent les directives éthiques de l'IA parues entre 2016 et 2020? », a obtenu sa réponse dans les chapitres quatre et cinq. Ce sont les directives pluralistes, de même que celles comportant une tension métaéthique entre le monisme et le pluralisme qui dominent dans ce type de littérature. Ainsi, des éléments éthiques, voire des injonctions incompatibles, cohabitent dans une même démarche à plusieurs reprises, et ce, que le document émane d'une entreprise privée, de la société civile ou d'une organisation internationale. La seconde question de recherche, « Quelle approche éthique permet de favoriser un dialogue optimal des décideurs politiques en ce qui a trait à l'éthique de l'IA? », a trouvé des propositions de réponses dans les chapitres six et sept. Le monisme non orthodoxe, formé de la prudence et de la reconnaissance de potentiels dilemmes insolubles, constitue la base éthique sur laquelle est fondé un dialogue herméneutique informé par des questions pouvant servir de guide aux décideurs politiques.

Contrairement à ce que j'attendais en amont de la thèse, ce n'est pas le pluralisme des valeurs ou le « procéduralisme » pris isolément qui domine en éthique de l'IA actuellement. Peut-être en est-il ainsi en raison du grand nombre d'études en éthique appliquée, qui ont recours au procéduralisme. Je l'ai exposé tout spécialement au chapitre deux : le procéduralisme implique une métaéthique moniste orthodoxe, comme l'éthique de la vertu, que je privilégiais au départ. Au fil de mon étude, j'ai constaté que cette dernière comporte des faiblesses qu'une démarche moniste *non* orthodoxe et une compréhension pratique prudentielle de l'herméneutique peuvent pallier pour assister le dialogue des législateurs.

La thèse, au fil des trois sections, a permis d'arriver à ces conclusions qui ne sont certes pas définitives, mais utiles comme issue d'un premier défrichement. Il a fallu commencer par une revue thématique du champ de l'éthique de l'intelligence artificielle au chapitre un. J'ai présenté les débats saillants en éthique de l'IA, allant des systèmes autonomes et de la sécurité internationale aux enjeux de la singularité de même que du risque existentiel, en passant par la protection de la

vie privée, les enjeux économiques ou le risque de biais et de discriminations dans les algorithmes, pour n'en nommer que quelques-uns. J'ai aussi souligné le nombre croissant d'études croisées sur les directives éthiques publiées dans les dernières années, pour emprunter à mon tour un chemin similaire à quelques égards. En effet, mon propos ne s'est pas inscrit en éthique appliquée, mais en métaéthique, c'est-à-dire dans l'analyse comparée des traditions à l'œuvre dans les documents.

En vue de ce travail analytique, j'ai clarifié, dans la première section, les fondements des traditions monistes orthodoxes que sont l'éthique de la vertu, l'utilitarisme et l'éthique déontologique, au chapitre deux. J'ai présenté quelques-unes de leurs forces et de leurs faiblesses, pour ensuite traiter du pluralisme des valeurs au chapitre trois. Cette section, quoique volumineuse, est indispensable à l'interprétation des démarches éthiques parues entre 2016 et 2020 en provenance de divers secteurs de la société. C'est une fois que le continuum métaéthique s'étirant entre le monisme et le pluralisme est clarifié que l'identification des traditions à l'œuvre dans les réflexions contemporaines peut s'opérer. L'utilitarisme et l'éthique déontologique sont des approches éthiques procédurales qui impliquent que l'agent ne se salit pas les mains, puisque le respect de la théorie éthique n'implique pas de faute morale. En cherchant à maximiser l'utilité, ou en accomplissant son devoir pour lui-même, par exemple, même si des compromis doivent être faits en cours de route, ces derniers n'impliquent pas de culpabilité de la part des agents. Étant donné que la réalité éthique n'est pas fractionnée de manière irréparable, on parle donc de monisme. Au sein de l'éthique de la vertu, l'unité des vertus garantie par la *phronesis* est également une marque de monisme. Un agent vertueux ne peut pas être vicieux simultanément, car en possédant réellement une vertu, il les possède toutes.

Le pluralisme des valeurs est assez critique de ce postulat de l'unité que l'on retrouve chez les monistes. Il semble faire l'impasse sur les difficultés du monde de la pratique éthique, qui implique la réalité de dilemmes absolument insolubles. Je suis consciente que ma compréhension du pluralisme peut être discutée, puisque le terme est hautement équivoque dans la littérature. Néanmoins, le *pluralisme des valeurs*, tel qu'il s'exprime par Isaiah Berlin et Bernard Williams principalement, forme effectivement un courant éthique à part entière, qu'on remarque largement à l'œuvre dans les directives éthiques concernant l'intelligence artificielle. Ce qui caractérise cette approche éthique est la notion de perte qui engendre le « réalisme moral » (Hall 2020). Par ailleurs,

le pluralisme, tel que je l'ai appréhendé dans la thèse, est assez proche de certains aspects de la pensée aristotélicienne, en ce qui a trait à l'importance de la raison pratique et de la prudence en éthique. Malgré cela, j'ai reproché au pluralisme des valeurs une importance trop marquée au compromis, ancrée dans la présupposition du fractionnement de tous les intérêts des parties impliquées, sans possibilité d'harmonisation sans perte. De plus, le pluralisme des valeurs a du mal à se défendre de manière complètement convaincante contre des allégations de relativisme éthique, quoique Berlin et Williams rejettent catégoriquement ce type d'insinuations. Ce qui apparaissait manifestement, mais implicitement, à l'issue de la première section, c'est que le « monisme orthodoxe » ainsi que le pluralisme des valeurs comprennent des écoles éthiques tout à fait différentes, voire en opposition, et que chacune d'entre elles possède d'évidentes faiblesses.

La deuxième section de la thèse a elle aussi regroupé deux chapitres. On peut les voir comme une sorte de contribution à la sociologie de la connaissance, si l'on voulait recenser les écoles à l'œuvre dans les réflexions éthiques sur l'IA parues dans les dernières années. Dans le chapitre quatre, après avoir exposé au lecteur l'échantillon avec lequel j'ai choisi de travailler, j'ai présenté les documents que je qualifierais de monistes ou quasi monistes, parce que certaines démarches sont difficiles à classifier hors de tout doute. Ce constat est loin d'être inquiétant puisque, comme je l'ai répété au long de ces pages, l'éthique n'est pas une discipline analogue aux sciences naturelles. Comme le dit d'ailleurs Aristote dans son *Éthique*, il convient de

> […] ne pas chercher une égale précision en toutes choses, mais au contraire, en chaque cas particulier tendre à l'exactitude que comporte la matière traitée, et seulement dans une mesure appropriée à notre investigation. (2014, Livre I, 7, 1098a25)

Pour chacune des traditions monistes en éthique, j'ai fourni un ou plusieurs exemples — une esquisse — de ce dont aurait l'air une approche à l'éthique de l'IA s'inspirant de leur théorie.

Le chapitre cinq a suivi une structure semblable à celle du chapitre quatre, pour ce qui est des démarches pluralistes en éthique de l'intelligence artificielle. C'est là que j'ai aussi exposé au lecteur que des tensions métaéthiques existaient — et même surabondaient — dans ce type de littérature. En effet, dans plusieurs directives, peu importe leur provenance, force était de constater que le monisme et le pluralisme cohabitaient de manière malaisée, impliquant de ce fait des contradictions internes. Ces incohérences ont une portée non négligeable pour les décideurs

politiques. Ces derniers aborderaient les enjeux moraux de l'IA, à la lumière de démarches exhibant une tension métaéthique, sans savoir s'ils doivent conformer la réalité à la théorie, ou simplement accepter la tragédie des mains sales résultant de la compromission inévitable de leurs valeurs.

Évidemment, d'autres influences pourraient s'être glissées dans les documents que je lis à la lumière des traditions monistes et pluralistes identifiées dans la thèse. C'est d'ailleurs fort probable, puisque je n'ai choisi que quatre écoles éthiques, majoritairement occidentales. Néanmoins, je crois que la tension métaéthique que j'ai relevée dans plusieurs de ces directives est réelle et, dans une certaine mesure, problématique. Plus encore, la combinaison inconsciente du monisme « orthodoxe », tel que décrit dans le chapitre deux, de même que du pluralisme des valeurs, explicité au chapitre trois, est d'autant plus précaire en raison de leurs failles respectives.

C'est la raison pour laquelle il importait que je fournisse à mon tour quelques propositions en éthique de l'IA. Je les ai conçues à l'intention des décideurs politiques explicitement. La troisième section de la thèse s'est présentée comme les fondements et l'expression de ma proposition alternative. Dans le chapitre six, j'ai présenté une critique plus étayée des points faibles du monisme orthodoxe des théories éthiques procédurales, ainsi que du pessimisme du pluralisme des valeurs. Cela m'a permis de montrer au lecteur pourquoi je n'allais pas emprunter la voie tracée par ces écoles pour ma proposition. J'ai par la suite avancé au lecteur les quatre éléments, interconnectés dans un tout holiste, de ce que je pensais être une approche éthique optimale, pouvant être qualifiée de moniste « non orthodoxe ». La vertu de prudence, qui se distingue à quelques égards de la *phronesis* classique aristotélicienne, est le premier élément qui implique les deux suivants : une sensibilité profonde au contexte de décision et d'action, arrimée à une orientation téléologique « douce » vers le bien commun. De pair avec une influence pluraliste qui permet la reconnaissance de la possibilité (mais non, j'ajouterais, de l'inévitabilité) de dilemmes éthiques insolubles, ces quatre éléments fondent la base éthique de ma contre-proposition aux législateurs. Comme ce type de prise de position n'est pas inédite, mais se retrouve dans la pensée — très cohérente — de philosophes éminents comme Charles Taylor et Martha Nussbaum, j'estime que cette approche minimise les tensions métaéthiques que j'ai observées dans mon analyse des documents.

Dans le chapitre sept, j'ai élaboré un guide pour le dialogue des décideurs politiques, en combinant mon approche éthique moniste non orthodoxe avec des influences de la philosophie herméneutique de Hans Georg Gadamer, au sein de laquelle le fait de questionner a une importance prépondérante. C'est en approchant les problématiques éthiques par l'entremise de la question, prudente, sensible au contexte, orientée vers le bien commun et ouverte à l'insolubilité de certains dilemmes, que les législateurs pourront entretenir une conversation de qualité en ce qui concerne l'intelligence artificielle. Mon avis est que cette approche est plus fructueuse, en politique, qu'une approche énonçant des principes, une procédure, ou encore une théorie à respecter. Elle se meut de manière harmonieuse avec la raison pratique, plutôt que la raison théorique ou spéculative. J'ai donc esquissé ce que j'ai appelé le « parcours herméneutique » pour les décideurs politiques, que j'estime provisoire — car il peut grandement être bonifié par la pratique —, mais pertinent comme point de départ au dialogue. Cette façon de voir n'est pas sans rappeler ce que le journaliste américain Howard Fineman suggère, à savoir que les fondements mêmes de son pays reposent sur la capacité d'argumenter et que « le processus qui [les] rend si fragiles [les] rend également durables » [Traduction libre] (2008, 14).

La contribution de cette thèse à l'avancement des connaissances en éthique de l'IA est certes limitée. Le champ d'études est en pleine expansion, dans la pratique comme dans la réflexion. Je crois pourtant que la présente thèse fournit quatre apports importants à la discipline. Le premier est le portrait que j'ai esquissé, aux chapitres quatre et cinq, de l'échantillon de directives en éthique de l'IA. À ma connaissance, une analyse métaéthique des documents n'a pas été menée de manière aussi profonde, bien que Charles Ess (2019) ait lui-même exploré la métaéthique de quelques directives éthiques de l'IA et que, comme je l'ai démontré au premier chapitre, les analyses croisées et recensements de ce type de documents ne manquent pas. Ma compréhension du pluralisme est toutefois différente de ces études. La deuxième contribution est liée à la première : il s'agit de l'omniprésence silencieuse du pluralisme des valeurs dans la réflexion sur l'éthique de l'intelligence artificielle. La troisième renvoie à la notion de « tension métaéthique », que j'ai développée aux fins de ce travail, sur la base de ce que j'ai observé dans mon analyse. Néanmoins, je crois que ce phénomène serait potentiellement observable dans d'autres applications de l'éthique, surtout si sa portée sociétale est importante, puisque les valeurs impliquées seraient

sûrement très variées. Enfin, l'apport plus « appliqué » de la thèse réside dans le parcours herméneutique élaboré à l'intention des décideurs politiques.

Cela étant dit, il est vrai que le parcours herméneutique n'a pas été mis à l'épreuve dans la pratique, avec des législateurs en chair et en os. Il s'agit d'une avenue de développement du propos qui serait fort intéressante pour poursuivre dans le sillage de cette thèse. De même, j'imagine que certains prétendront que ma proposition comporte une certaine faiblesse — d'autres diront la naïveté —, celle de reposer sur la bonne volonté des décideurs pour son fonctionnement optimal. Mais, comme je l'ai dit au chapitre dernier, un recours cynique au parcours herméneutique ne permet pas le plein exercice de la vertu de prudence, qui implique de viser le bien commun en toute sincérité.

Je suis également sensible à la réalité « pluraliste » de la compréhension des écoles éthiques que j'ai exposées au fil de ces pages. Évidemment, le lecteur peut être en désaccord avec mon interprétation des fondements des traditions éthiques, et ma lecture de leur expression dans les directives. Nous pourrions avoir des regards divergents sur ces mêmes documents. À mon sens, cela contribuerait à nourrir une discussion encore plus riche sur l'éthique de l'intelligence artificielle. De même, je ne m'attends pas à ce que le monisme non orthodoxe soit embrassé par tous comme étant l'approche éthique la meilleure. En éthique politique, peut-être est-ce la « moins mauvaise », ou l'une des moins mauvaises, pour parler comme une pluraliste. Il n'en demeure pas moins que l'envers d'un propos normatif, c'est évidemment son caractère non universel et potentiellement équivoque. En effet, je n'ai pas la prétention de pouvoir prouver l'unification ou le fractionnement ultime de la réalité. Je peux seulement me positionner de façon inéluctablement métaphysique en ce qui concerne ce point.[63]

Il faudrait entendre directement les décideurs politiques sur la viabilité du parcours herméneutique et le compléter à la suite de leurs commentaires. Il serait très enrichissant d'explorer et de comparer les nuances et les couleurs que prendrait le parcours selon le contexte des législateurs qui l'emploient. Une telle analyse comparative serait intéressante à documenter et

[63] Même John Rawls ne peut entièrement éviter une discussion sur la métaphysique dans les questions de pensée politique, lui qui était connu pour l'orientation « politique, mais non métaphysique » de son travail (voir Rawls 2005, 29, note 31).

pourrait nous informer sur ce qui importe réellement à certaines cultures, à des groupes en particulier, ou encore à des époques ou moments de l'histoire qui ont leurs spécificités propres. Forcément, le parcours herméneutique ne serait jamais totalement le même pour tous les groupes, voire pour toutes les situations. Cela étant dit, sa grande flexibilité n'empêche pas qu'il soit bonifié à l'avenir.

Le pluralisme des valeurs est souvent confondu avec la reconnaissance de la multiplicité et des différences. Cela est compréhensible puisque l'absence de rigidité de cette approche éthique permet la prise en compte de plusieurs valeurs, mais pas leur harmonisation sans heurt. Le pluralisme des valeurs a quelque chose à apporter à la réflexion éthique contemporaine, en particulier en intelligence artificielle, où le contexte est en fluctuation constante, surtout en raison de la vitesse des percées technologiques. Il doit toutefois être nommé et présenté pour ce qu'il est, avec son insistance sur le potentiel tragique des décisions, ce qui n'est pas le cas actuellement dans la littérature en éthique de l'IA. Ce faisant, les citoyens qui « pensent l'éthique » pourraient remarquer, souligner et tenter de résoudre les tensions métaéthiques contemporaines entre le monisme et le pluralisme. Cette thèse aura peut-être permis un premier pas en ce sens.

Références bibliographiques

Abel, Alain. 2019. « Deepfake. Le vrai du faux ». *Enquête (Radio-Canada)*. https://ici.radio-canada.ca/tele/enquete/site/episodes/424051/video-deep-fake-imitation-visage-technologie-numerique-web-immigration.

Abizadeh, Arash. 2002. « The Passions of the Wise: Phronêsis, Rhetoric, and Aristotle's Passionate Practical Deliberation ». *Review of Metaphysics*, nº 56 : 267-96. https://www.jstor.org/stable/20131817.

Access Now. 2020. « Access Now Resigns from the Partnership on AI ». *Access Now* (blog). 13 octobre 2020. https://www.accessnow.org/access-now-resignation-partnership-on-ai/.

Adam, Alison. 1998. *Artificial Knowing: Gender and the Thinking Machine*. Psychology Press. https://books.google.ca/books?id=IVLmW1C7oGEC.

Agence France-Presse [AFP]. 2020. « Élections américaines : Facebook et Google mettent en avant les infos fiables ». *La Presse*, 13 août 2020. https://www.lapresse.ca/affaires/techno/2020-08-13/elections-americaines-facebook-et-google-mettent-en-avant-les-infos-fiables.php.

AI Global. 2017. « AI Global. Catalyzing Responsible Artificial Intelligence ». AI Global. 2017. https://ai-global.org/.

Alexander, Larry, et Michael Moore. 2016. « Deontological Ethics ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/.

AlgorithmWatch, M. Spielkamp, M. Matzat, K. Penner, K. Thummler, V. Thiel, S. Gießler, et A. Eisenhauer. s. d. « AI Ethics Guidelines Global Inventory ». AI Ethics Guidelines Global Inventory. Consulté le 13 mai 2020. https://inventory.algorithmwatch.org.

Allo, Patrick, Mariarosaria Taddeo, Sandra Wachter, Luciano Floridi, et Brent Daniel Mittelstadt. 2016. « The Ethics of Algorithms: Mapping the Debate ». *Big Data & Society* 3 (2) : 1-21. http://journals.sagepub.com/doi/pdf/10.1177/2053951716679679.

Ananny, Mike. 2016. « Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness ». *Science, Technology, & Human Values* 41 (1) : 93-117. https://doi.org/10.1177/0162243915606523.

Anderson, Mark Robert. 2017. « After 75 Years, Isaac Asimov's Three Laws of Robotics Need Updating ». *The Conversation*, 2017. http://theconversation.com/after-75-years-isaac-asimovs-three-laws-of-robotics-need-updating-74501.

Anderson, R. Lanier. 2017. « Friedrich Nietzsche ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Summer 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/nietzsche/.

Annas, Julia. 1995. « Virtue as a Skill ». *International Journal of Philosophical Studies* 3 (2) : 227-43. https://doi.org/10.1080/09672559508570812.

Annas, Julia. 2004. « Being Virtuous and Doing the Right Thing ». *Proceedings and Addresses of the American Philosophical Association* 78 (2) : 61-75. https://doi.org/10.2307/3219725.

Anscombe, G.E.M. 1958. « Modern Moral Philosophy ». *Philosophy* 33 (124) : 1-16. https://www.pitt.edu/~mthompso/readings/mmp.pdf.

Apfel, Lauren J. 2011. *The Advent of Pluralism: Diversity and Conflict in the Age of Sophocles*. New York, NY : Oxford University Press. https://books.google.ca/books?id=fimQDwAAQBAJ&hl=fr.

Aquin, Thomas d'. 1928. *Somme théologique. L'âme humaine. 1a, Questions 75-83*. Paris, Tournai, Rome : Société Saint Jean L'Évangéliste; Desclés & Cie.

Aquin, Thomas d'. 2008. *Commentaire de l'Éthique à Nicomaque d'Aristote*. Docteur angélique. http://docteurangelique.free.fr/livresformatweb/philosophie/commentaireethiquenicomaque.htm.

Aquin, Thomas d'. s. d. « Summa Theologiae ». New Advent. Consulté le 27 novembre 2019. https://www.newadvent.org/summa/.

Aristote. 1990. *Les politiques*. Paris : GF Flammarion.

Aristote. 2014. *Éthique à Nicomaque*. Les Échos du Maquis. https://philosophie.cegeptr.qc.ca/wp-content/documents/%C3%89thique-%C3%A0-Nicomaque.pdf.

Aristote. s. d. *La Rhétorique*. Œuvre numérisée par J. P. MURCIA. Consulté le 20 mai 2020. http://www.documentacatholicaomnia.eu/03d/-384_-322,_Aristoteles,_Rhetorique,_FR.pdf.

Asimov, Isaac. 2013. *I, Robot*. Londres : Harper Voyager.

Assemblée nationale et Sénat de France, Claude De Ganay, et Dominique Gillot. 2017. « Rapport au nom de l'Office parlementaire d'évaluation des choix scientifiques et technologiques. Pour une intelligence artificielle maîtrisée, utile et démystifiée. Tome I : Rapport ». N° 4594 Assemblée nationale / N° 464 Sénat. Office parlementaire d'évaluation des choix scientifiques et technologiques. https://www.senat.fr/rap/r16-464-1/r16-464-11.pdf.

Association for Computing Machinery (ACM). 2018. « ACM Code of Ethics and Professional Conduct ». 2018. https://www.acm.org/code-of-ethics.

Astor, Maggie. 2017. « Your Roomba May Be Mapping Your Home, Collecting Data That Could Be Shared ». *The New York Times*, 25 juillet 2017, sect. Technology. https://www.nytimes.com/2017/07/25/technology/roomba-irobot-data-privacy.html.

Bacciarelli et al. s. d. « The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems ». Édité par Amnesty International et Access Now. Amnesty International & Access Now. Consulté le 25 juillet 2018. https://www.accessnow.org/cms/assets/uploads/2018/05/Toronto-Declaration-D0V2.pdf.

Beauchamp, Tom L. 2001. *Principles of Biomedical Ethics*. 5th ed.. New York, N.Y. : Oxford University Press. http://www.myilibrary.com?id=83526.

Beauchamp, Tom L., et Oliver Rauprich. 2016. « Principlism ». Dans *Encyclopedia of Global Bioethics*, édité par Henk ten Have, 2282— 93. Cham : Springer International Publishing. https://doi.org/10.1007/978-3-319-09483-0_348.

Bengio, Yoshua. 2017. « La communauté de l'intelligence artificielle a bien fait ses devoirs ». *Le Devoir*, 2017. https://www.ledevoir.com/opinion/libre-opinion/514491/la-communaute-de-l-ia-a-bien-fait-ses-devoirs.

Bengio, Yoshua, et Marc-Antoine Dilhac. s. d. « Une application de suivi de contacts «intelligente et éthique» contre la COVID-19 ». *Le Devoir*. Consulté le 25 mai 2020.

https://www.ledevoir.com/opinion/idees/579485/coronavirus-une-application-de-suivi-de-contacts-intelligente-et-ethique.

Bentham, Jeremy. 2000. *The Works of Jeremy Bentham*. Electronic edition. Past Masters Series (InteLex Corporation). Charlottesville, Va : InteLex Corporation. http://pm.nlx.com/xtf/view?docId=bentham/bentham.00.xml.

Berger, Peter L. 2005. *L'impératif hérétique : les possibilités actuelles du discours religieux*. Paris : Van Dieren Editeur.

Berlin, Isaiah. 1964. « Rationality of Value Judgements ». Dans *Rational Decision*, édité par Carl J. Friedrich, 221— 23. New York — London : Atherton Press. http://berlin.wolf.ox.ac.uk/published_works/singles/RationalityofValueJudgements.pdf.

Berlin, Isaiah. 1988. « Deux conceptions de la liberté ». Dans *Éloge de la liberté*. Paris : Calman-Lévy. http://www.institutcoppet.org/wp-content/uploads/2013/10/Eloge-de-la-liberté-Isaiah-Berlin.pdf.

Berlin, Isaiah. 1990. « The Pursuit of the Ideal ». Dans *The Crooked Timber of Humanity: Chapters in the History of Ideas*, édité par Henry Hardy, 277. New York : Knopf : Distributed by Random House, New York. http://assets.press.princeton.edu/chapters/s9983.pdf.

Berlin, Isaiah, et Bernard Williams. 1994. « Pluralism and Liberalism: A Reply ». *Political Studies* 42 (2) : 306–9. https://doi.org/10.1111/j.1467-9248.1994.tb01914.x.

Berti, Enrico. 2000. « Gadamer and the Reception of Aristotle's Intellectual Virtues ». *Revista Portuguesa de Filosofia* 56 (3/4) : 345— 60. https://www.jstor.org/stable/40337581.

Bieber, Friedemann, et Jakob Moggia. 2020. « Risk Shifts in the Gig Economy: The Normative Case for an Insurance Scheme against the Effects of Precarious Work ». *Journal of Political Philosophy* n/a (n/a). https://doi.org/10.1111/jopp.12233.

Blattberg, Charles. 2004. *Et si nous dansions? Pour une politique du bien commun au Canada*. Les Presses de l'Université de Montréal. https://umontreal.on.worldcat.org/oclc/243566508.

Blattberg, Charles. 2008. « From Moderate to Extreme Holism ». Dans *Patriotic Elaborations : Essays in Practical Philosophy*. Montreal & Kingston: McGill-Queen's University Press. https://www.academia.edu/25746672/From_Moderate_to_Extreme_Holism.

Blattberg, Charles. 2009. « Patriotic, Not Deliberative, Democracy ». Dans *Patriotic Elaborations : Essays in Practical Philosophy*, 26 à 42. Montreal and Kingston: McGill-Queen's University Press.

Blattberg, Charles. 2013. « Welfare: Towards the Patriotic Corporation », 1-12. https://www.academia.edu/3277458/Welfare_Towards_the_Patriotic_Corporation.

Blattberg, Charles. 2015. « Playing Political Philosophy ». *The Review of Politics* 78 (02) : 307-8. https://www.academia.edu/11747339/Playing_Political_Philosophy.

Blattberg, Charles. 2016. « The Ironic Tragedy of Human Rights », 1-19. https://www.academia.edu/2067337/The_Ironic_Tragedy_of_Human_Rights.

Blattberg, Charles. 2018. « Dirty Hands: The One and the Many ». *The Monist* 101 (2) : 150-69. https://doi.org/10.1093/monist/onx040.

Boddington, Paula. 2017. *Towards a Code of Ethics for Artificial Intelligence*. Artificial Intelligence: Foundations, Theory, and Algorithms. Oxford, England : Springer. https://link.springer.com/content/pdf/10.1007%2F978-3-319-60648-4.pdf.

Boden, Margaret A. 2018. *Artificial Intelligence: A Very Short Introduction*. New York : Oxford University Press.

Bordeleau, Stéphane. 2020. « La firme montréalaise Element AI vendue à l'américaine ServiceNow | Radio-Canada.ca », 30 novembre 2020. https://ici.radio-canada.ca/nouvelle/1753370/intelligence-artificielle-element-ai-vendue-service-now.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. First edition.. MyiLibrary. Oxford, United Kingdom: Oxford University Press. http://lib.myilibrary.com?id=627724.

Bostrom, Nick, Allan Dafoe, et Carrick Flynn. 2020. « Policy Desiderata for Superintelligent AI: A Vector Field Approach (V. 4.3) ». Dans *Ethics of Artificial Intelligence*, édité par S Matthew Liao. New York : Oxford University Press. https://nickbostrom.com/papers/aipolicy.pdf.

Bostrom, Nick, et Eliezer Yudkowsky. 2014. « The Ethics of Artificial Intelligence ». Dans *The Cambridge Handbook of Artificial Intelligence*, édité par Keith Frankish et William M. Ramsey, 316— 34. New York : Cambridge University Press. https://www.cambridge.org/core/books/cambridge-handbook-of-artificial-intelligence/3DCB2E04739722A99EDE86B7A34A30E3.

Brashier, Nadia M., et Daniel L. Schacter. 2020. « Aging in an Era of Fake News ». *Current Directions in Psychological Science* 29 (3) : 316-23. https://doi.org/10.1177/0963721420915872.

Brundage, Miles, et Joanna Bryson. 2016. « Smart Policies for Artificial Intelligence ». https://arxiv.org/pdf/1608.08196.pdf.

Brzozowski, Alexandra. 2019. « No Progress in UN Talks on Regulating Lethal Autonomous Weapons ». *Www.Euractiv.Com* (blog). 22 novembre 2019. https://www.euractiv.com/section/global-europe/news/no-progress-in-un-talks-on-regulating-lethal-autonomous-weapons/.

Burton, Emanuelle, Judy Goldsmith, et Nicholas Mattei. s. d. « Teaching AI Ethics Using Science Fiction ». *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*. Consulté le 22 septembre 2020. https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10139/10129.

Business Wire. 2019. « $1M Leaders Prize to Be Awarded to Solution Combating “Fake News” Using Artificial Intelligence ». *Financial Post*. 19 juin 2019. https://business.financialpost.com/pmn/press-releases-pmn/business-wire-news-releases-pmn/1m-leaders-prize-to-be-awarded-to-solution-combating-fake-news-using-artificial-intelligence.

Calo, Ryan. 2017. « Artificial Intelligence Policy: A Primer and Roadmap ». SSRN Scholarly Paper ID 3015350. Rochester, NY : Social Science Research Network. https://papers.ssrn.com/abstract=3015350.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, et Kate Crawford,. 2017. « AI Now 2017 Report ». AI Now Institute. https://ainowinstitute.org/AI_Now_2017_Report.pdf.

Cavello, B. 2020. « PAI Launches Interactive Project To Put Ethical AI Principles into Practice ». The Partnership on AI. 21 janvier 2020. https://www.partnershiponai.org/pai-launches-interactive-project-to-put-ethical-ai-principles-into-practice/.

Childs, Martin. 2011. « John McCarthy : Computer Scientist Known as the Father of AI ». *The Independent*, 1 novembre 2011. http://www.independent.co.uk/news/obituaries/john-mccarthy-computer-scientist-known-as-the-father-of-ai-6255307.html.

Clark, Jack. 2020. « COVID-19 Surveillance Strengthens Authoritarian Governments ». CSET Foretell. *CSET Foretell Blog* (blog). 28 juillet 2020. https://www.cset-foretell.com/blog/surveillance-creep.

Coeckelbergh, Mark. 2020a. *AI Ethics*. Cambridge, MA : The MIT Press. https://www.amazon.ca/-/fr/Mark-Coeckelbergh/dp/0262538199.

Coeckelbergh, Mark. 2020b. « How To Use Virtue Ethics for Thinking About the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance ». *International Journal of Social Robotics*, octobre, (Pas de # de pages). https://doi.org/10.1007/s12369-020-00707-z.

Coeckelbergh, Mark, Janina Loh, Michael Funk, Johanna Seibt, et Marco Norskov, éd. 2018. *Envisioning Robots in Society - Power, Politics, and Public Space. Proceedings of Robophilosophy 2018 / TRANSOR 2018. February 14-17, 2018, University of Vienna, Austria*. Amsterdam-Berlin-Washington, DC : IOS Press.

Cohen, Allison, et Abhishek Gupta. 2020. « Report prepared by the Montreal AI Ethics Institute In Response to Mila's Proposal for a Contact Tracing App ». *arXiv:2008.04530 [cs]*, août. http://arxiv.org/abs/2008.04530.

Comité d'élaboration de la Déclaration de Montréal IA responsable. 2018. « La Déclaration de Montréal pour un développement responsable de l'intelligence artificielle ». https://docs.wixstatic.com/ugd/ebc3a3_28b2dfe7ee13479caaf820477de1b8bc.pdf?index=true.

Commissariat à la protection de la vie privée du Canada. 2018. « Consultation sur les propositions du Commissariat visant à assurer une réglementation adéquate de l'intelligence artificielle ». 26

janvier 2018. https://priv.gc.ca/fr/a-propos-du-commissariat/ce-que-nous-faisons/consultations/consultation-ai/pos_ai_202001/.

Commissariat à la protection de la vie privée du Canada. 2020. « Un cadre réglementaire pour l'IA : recommandations pour la réforme de la LPRPDE ». 12 novembre 2020. https://priv.gc.ca/fr/a-propos-du-commissariat/ce-que-nous-faisons/consultations/consultations-terminees/consultation-ai/reg-fw_202011/.

Commission européenne. 2019. « Ethics Guidelines for Trustworthy AI ». Text. Shaping Europe's Digital Future — European Commission. 8 avril 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Commission mondiale d'éthique des connaissances scientifiques et des technologies. 2017. « Report of COMEST on Robotic Ethics ». SHS/YES/COMEST-10/17/2 REV. https://unesdoc.unesco.org/ark:/48223/pf0000253952.

Conn, Ariel. 2018. « European Parliament Passes Resolution Supporting a Ban on Killer Robots ». Future of Life Institute. 14 septembre 2018. https://futureoflife.org/2018/09/14/european-parliament-passes-resolution-supporting-a-ban-on-killer-robots/.

Copeland, Eddie. 2018. « 10 Principles for Public Sector Use of Algorithmic Decision Making ». Nesta. 2018. https://www.nesta.org.uk/blog/10-principles-for-public-sector-use-of-algorithmic-decision-making/.

Crawford, Kate. 2016. « Artificial Intelligence's White Guy Problem ». *The New York Times*, 25 juin 2016, sect. Opinion. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

Crawford, Matthew B. 2019. « Algorithmic Governance and Political Legitimacy ». *American Affairs Journal* III (2). https://americanaffairsjournal.org/2019/05/algorithmic-governance-and-political-legitimacy/.

Crick, Bernard. 1993. *In Defence of Politics*. 4th ed.. Chicago : University of Chicago Press.

Crisp, Roger, éd. 1996. *How Should One Live? Essays on the Virtues*. Oxford ; New York : Oxford University Press.

Cuthbertson, Anthony. 2020. « Elon Musk Claims AI Will Overtake Humans “in Less than Five Years” ». *The Independent*, 27 juillet 2020. https://www.independent.co.uk/life-style/gadgets-and-tech/news/elon-musk-artificial-intelligence-ai-singularity-a9640196.html.

Danaher, John, et Shannon Vallor. 2018. « Philosophical Disquisitions: Episode #45 — Vallor on Virtue Ethics and Technology ». *Philosophical Disquisitions* (blog). 18 septembre 2018. https://philosophicaldisquisitions.blogspot.com/2018/09/episode-45-vallor-on-virtue-ethics-and.html.

Daniels, Craig. 2018. « Ethical AI: People Write Algorithms. People Can Fix Them ». *Communitech News*, 5 décembre 2018. http://snip.ly/yqv45d#http://news.communitech.ca/ethical-ai-people-write-algorithms-people-can-fix-them/.

Darwall, Stephen. 2003a. « Introduction ». Dans *Consequentialism*, édité par Stephen Darwall, 1— 6. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Darwall, Stephen. 2003b. « Introduction ». Dans *Virtue Ethics*, édité par Stephen Darwall, 1— 3. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Davey, Nicholas. 2006. *Unquiet Understanding. Gadamer’s Philosophical Hermeneutics*. New York : State University of New York Press.

Davies, Jim. 2016. « Program Good Ethics into Artificial Intelligence ». *Nature* 538 (7625) : 291. https://www.nature.com/news/program-good-ethics-into-artificial-intelligence-1.20821.

De Luca-Baratta, Anthony. 2019. « Autonomous Vehicles, Social Robots, and the Mitigation of Risk: A New Consequentialist Approach ». Montreal AI Ethics Institute. 26 août 2019. https://montrealethics.ai/ai-design-social-robots-and-the-mitigation-of-risk-toward-a-new-consequentialist-approach-to-ai-safety/.

DeepMind. 2017. « DeepMind Ethics and Society ». DeepMind. 2017. https://deepmind.com/about/ethics-and-society.

Deglise, Fabien. 2020. « Ottawa dit non à Mila et son application COVI de recherche de contacts de personnes contaminées ». *Le Devoir*, 10 juin 2020. https://www.ledevoir.com/societe/580507/ottawa-dit-non-a-mila.

Della Foresta, Josiah. 2020. « Consequentialism and Machine Ethics - Towards a Foundational Machine Ethic to Ensure the Ethical Conduct of Artificial Moral Agents ». Montreal AI Ethics Institute. 31 mars 2020. https://montrealethics.ai/consequentialism-and-machine-ethics-towards-a-foundational-machine-ethic-to-ensure-the-ethical-conduct-of-artificial-moral-agents/.

Denyer, Nicholas. 1981. « Chess and Life: The Structure of a Moral Code ». *Proceedings of the Aristotelian Society* 82 : 59-68. https://www.jstor.org/stable/4544979.

Dignum, Virginia. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms Ser. Cham : Springer. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5972650.

Dilhac, Marc-Antoine. 2018. « Quels risques éthiques ? » *Courrier de l'UNESCO. Intelligence artificielle : promesses et menaces*, 2018. https://fr.unesco.org/courier/2018-3/ethical-risks-ai.

Dilhac, Marc-Antoine, Camylle Lanteigne, Carl-Maria Mörch, Lucia Flores Echaiz, Pauline Noiseau, Fatima Gabriela Gomez Salazar, et Vincent Mai. 2020. « The Open Dialogue on AI Ethics: Contribution to the UNESCO Recommendation on the Ethics of Artificial Intelligence ». AlgoraLab — Mila (Université de Montréal). https://opendialogueonai.com/wp-content/uploads/2020/08/RAPPORT_DialogueOnAIethics_ENG_v2.pdf?fbclid=IwAR3bS54aJOdw33a068FO8g4fpzhFQ-9tdbREgP1ehnv28q7NnWv3dazEG8M.

Dreyfus, Hubert L. 1980. « Holism and Hermeneutics ». *The Review of Metaphysics* 34 (1) : 3-23. https://www.jstor.org/stable/20127455.

Dreyfus, Hubert L. 1999. *What Computers Still Can't Do. A Critique of Artifical Reason*. 6e éd. Cambridge, Massachusetts; London, England: The MIT Press.

Dreyfus, Hubert L. 2014. *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*. *Skillful Coping*. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199654703.001.0001/acprof-9780199654703.

Dreyfus, Hubert L., et Stuart Dreyfus. 1988. *Mind Over Machine*. New York : Free Press.

Dreyfus, Hubert L., et Charles Taylor. 2015. *Retrieving Realism*. Cambridge, MA : Harvard University Press.

https://books.google.ca/books?id=UL_FCQAAQBAJ&printsec=frontcover&hl=fr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

Ducas, Isabelle. 2020. « Pas de reconnaissance faciale par les policiers sans l'accord des élus ». *La Presse*, 22 septembre 2020, sect. Actualités. https://www.lapresse.ca/actualites/2020-09-22/montreal/pas-de-reconnaissance-faciale-par-les-policiers-sans-l-accord-des-elus.php.

Dutton, Tim. 2018a. « AI Policy 101: An Introduction to the 10 Key Aspects of AI Policy ». *Medium* (blog). 5 juillet 2018. https://medium.com/politics-ai/ai-policy-101-what-you-need-to-know-about-ai-policy-163a2bd68d65.

Dutton, Tim. 2018b. « An Overview of National AI Strategies ». *Medium*, 25 juillet 2018. https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

Dutton, Tim. 2018c. « Politics + AI Reading List ». *Politics + AI* (blog). 26 juin 2018. https://medium.com/politics-ai/politics-ai-reading-list-839ced8a408b.

dvanw. 2010. « L'ameublement du cerveau : 146 — Paradoxe de Moravec ». *L'ameublement du cerveau* (blog). 29 mars 2010. http://dvanw.blogspot.com/2010/03/146-paradoxe-de-moravec.html.

Éditions Larousse. s. d. « Définitions : optimal — Dictionnaire de français Larousse ». Larousse. Consulté le 17 septembre 2020. https://www.larousse.fr/dictionnaires/francais/optimal/56252.

Element AI. 2020. « Les Fiducies de Données : Une Gouvernance Renforcée Des Données Qui Responsabilise Le Public ». Element AI. 2020. https://hello.elementai.com/les-fiducies-de-donnees.html.

Elish, Madeleine Clare. 2020. « Who Is Responsible When Autonomous Systems Fail? » *Centre for International Governance Innovation*, 15 juin 2020. https://www.cigionline.org/articles/who-responsible-when-autonomous-systems-fail.

Engelmann, Severin, Mo Chen, Felix Fischer, Ching-yu Kao, et Jens Grossklags. 2019. « Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior ». Dans *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 69–78. FAT* '19. New York, NY, USA : ACM. https://doi.org/10.1145/3287560.3287585.

Ess, Charles Melvin. 2006. « Ethical pluralism and global information ethics ». *Ethics and Information Technology*, nº 8 : 215-26. https://link.springer.com/article/10.1007/s10676-006-9113-3.

Ess, Charles Melvin. 2009. *Digital Media Ethics*. Digital Media and Society Series. Cambridge, UK : Polity.

Ess, Charles Melvin. 2019. « Intercultural Privacy: A Nordic Perspective ». Dans *Privatsphäre 4.0 : Eine Neuverortung des Privaten im Zeitalter der Digitalisierung*, édité par Hauke Behrendt, Wulf Loh, Tobias Matzner, et Catrin Misselhorn, 73— 88. Stuttgart : J.B. Metzler. https://doi.org/10.1007/978-3-476-04860-8_5.

Ess, Charles Melvin. 2020. « Interpretative Pros Hen Pluralism: from Computer-Mediated Colonization to a Pluralistic Intercultural Digital Ethics ». *Philosophy & Technology*, nº 33 (juillet) : 551–569. https://doi.org/10.1007/s13347-020-00412-9.

Ethical Explorer. s.d.a « Field Guide ». Ethical Explorer: Tools to help navigate the future impact of today's technology. (Consulté le 22 septembre 2020) https://ethicalexplorer.org/.

Ethical Explorer. s.d.b « Tech Risk Zones ». Ethical Explorer: Tools to help navigate the future impact of today's technology. (Consulté le 22 septembre 2020) https://ethicalexplorer.org/.

Etzioni, Amitai, et Oren Etzioni. 2016a. « AI Assisted Ethics ». *Ethics and Information Technology* 18 (2) : 149-56. https://link.springer.com/content/pdf/10.1007%2Fs10676-016-9400-6.pdf.

Etzioni, Oren, et Amitai Etzioni. 2016b. « Designing AI systems that obey our laws and values ». *Communications of the ACM* 59 (9) : 29-31. http://delivery.acm.org/10.1145/2960000/2955091/p29-etzioni.pdf?ip=132.204.9.239&id=2955091&acc=ACTIVE%20SERVICE&key=FD0067F557510FFB%2EA58F811D2973983A%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1529633140_a44bc8246e88140ca6e6acc61573c420.

Etzioni, Oren. 2017. « How to Regulate Artificial Intelligence ». *The New York Times*, 1 septembre 2017, sect. Opinion. https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html.

Eubanks, Virginia. 2017. *Automating Inequality : How High-Tech Tools Profile, Police, and Punish the Poor*. New York : St Martin's Press.

European Commission. 2018a. « Communication From the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe ». https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

European Commission. 2018b. « Declaration of cooperation on Artificial Intelligence ». https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence.

European Commission. 2018c. « Statement on artificial intelligence, robotics and "autonomous" systems ». https://publications.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1.

European Commission. 2019. « Communication From the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Building Trust in Human-Centric Artificial Intelligence ». https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence.

European Parliament. Directorate-general for Internal Policies. Policy Department. Citizens' Rights and Constitutional Affairs. 2016. « European Civil Law Rules on Robotics ». PE 571.379. http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf.

European Union. 2018. « General Data Protection Regulation ». Intersoft Consulting. 2018. https://gdpr-info.eu.

Expert Advisory Group on Society, Technology and Ethics in a Pandemic (STEP). 2020. « Society, Technology and Ethics in a Pandemic. Expert Advisory Group Report ». CIFAR. https://cifar.ca/cifarnews/2020/05/07/expert-advisory-group-on-society-technology-ethics-in-a-pandemic-(step/.

Field, Hayden. 2020. « An A.I. Training Tool Has Been Passing Its Bias to Algorithms for Almost Two Decades ». *Medium*, 20 août 2020. https://onezero.medium.com/the-troubling-legacy-of-a-biased-data-set-2967ffdd1035.

Fineman, Howard. 2008. *The Thirteen American Arguments: Enduring Debates That Inspire and Define Our Country*. Random House Publishing Group. https://books.google.ca/books?id=XJFBLEhFgXcC&hl=fr.

Fjeld, Jessica, Hannah Hilligoss, Nele Achten, Maia Levy Daniel, Joshua Feldman, et Sally Kagay. 2019. « Principled Artificial Intelligence. A Map of Ethical and Rights-Based Approaches ». Berkman Klein Center for Internet & Society at Harvard University. 4 juillet 2019. https://ai-hr.cyber.harvard.edu/primp-viz.html.

Floridi, Luciano. 2014. *The Ethics of Information*. Oxford Scholarship Online. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199641321.001.0001/acprof-9780199641321-chapter-13.

Floridi, Luciano. 2019. « Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical ». *Philosophy & Technology* 32 (2) : 185-93. https://doi.org/10.1007/s13347-019-00354-x.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Lütge, et al. 2018a. « AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations ». *Minds and Machines* 28 (novembre) : 689-707. https://doi.org/10.1007/s11023-018-9482-5.

Floridi, Luciano, Josh Cowls, Thomas C. King, et Mariarosaria Taddeo. 2020. « How to Design AI for Social Good: Seven Essential Factors ». *Science and Engineering Ethics*, n° 26 (avril) : 1771-96. https://doi.org/10.1007/s11948-020-00213-5.

Floridi, Luciano, Mariarosaria Taddeo, Brent Mittelstadt, Sandra Wachter, et Corinne Cath. 2018b. « Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach ». *Science and Engineering Ethics* 24 (2) : 505–528. https://doi.org/10.1007/s11948-017-9901-7.

Flyvbjerg, Bent. 2001. *Making Social Science Matter: Why Social Inquiry Fails and How it Can Succeed Again*. Cambridge, UK : Cambridge University Press.

Flyvbjerg, Bent, Todd Landman, et Sanford Schram. 2012. *Real Social Science : Applied Phronesis*. Cambridge ; New York : Cambridge University Press. https://doi.org/10.1017/CBO9780511719912.

Foot, Philippa. 1978. *Virtues and Vices and Other Essays in Moral Philosophy*. Berkeley and Los Angeles : University of California Press.

Foucault, Michel. 1993. *Surveiller et punir : naissance de la prison*. Collection Tel ; 225. [Paris] : Gallimard. http://catalogue.bnf.fr/ark:/12148/cb355675932.

Franssen, Maarten, Gert-Jan Lokhorst, et Ibo van de Poel. 2018. « Philosophy of Technology ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entriesechnology/.

franzke, aline shakti, Anja Bechmann, Michael Zimmer, Charles Melvin Ess, et The Association of Internet Researchers. 2020. « Internet Research: Ethical Guidelines 3.0. » https://aoir.org/reports/ethics3.pdf.

Frischmann, Brett M., et Evan Selinger. 2018. *Re-Engineering Humanity*. Cambridge, United Kingdom ; New York, NY : Cambridge University Press. https://doi.org/10.1017/9781316544846.

Future of Life Institute. 2015. « AI Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence ». Future of Life Institute. 2015. https://futureoflife.org/ai-open-letter/.

Future of Life Institute. s. d.a « AI Policy ». Future of Life Institute. Consulté le 2 octobre 2019a. https://futureoflife.org/ai-policy/.

Future of Life Institute. s. d.b « Lethal Autonomous Weapons Pledge ». Future of Life Institute. Consulté le 5 août 2020b. https://futureoflife.org/lethal-autonomous-weapons-pledge/.

Future of Life Institute (FLI). 2017. « Asilomar AI Principles ». Future of Life Institute (FLI). 2017. https://futureoflife.org/ai-principles/.

G20 Countries. 2019. « G20 AI Principles ». https://www.g20.org/pdf/documents/en/annex_08.pdf.

Gadamer, Hans Georg. 1982. *L'Art de comprendre. Écrits II. Herméneutique et Champ de l'expérience humaine*. Bibliothèque philosophique. Aubier.

Gadamer, Hans Georg. 1990. *Vérité et méthode. Les grandes lignes d'une herméneutique philosophique*. Paris : Éditions du Seuil.

Gagné, Jean-François. 2019. « Putting AI Guidelines to Work ». *Element AI* (blog). 2019. https://www.elementai.com/news/2019/putting-ai-ethics-guidelines-to-work.

Gallie, W. B. 1956. « Essentially Contested Concepts ». *Proceedings of the Aristotelian Society* 56 : 167— 98. https://www.jstor.org/stable/4544562.

Garner, James W. 1907. « Political Science and Ethics ». *International Journal of Ethics* 17 (2) : 194-204. https://www.jstor.org/stable/2375844.

Gerlat, Pierre-Yves. 2020. « Le Parlement européen ouvre la voie à une première série de règles sur l'intelligence artificielle ». *Actu IA*, 26 octobre 2020. https://www.actuia.com/actualite/le-parlement-europeen-ouvre-la-voie-a-une-premiere-serie-de-regles-sur-lintelligence-artificielle/.

German Government, Federal Ministry of Transport and Digital Infrastructure. 2017. « Ethics Commission [Report]: Automated and Connected Driving ». https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile.

Gerrish, Sean. 2018. *How Smart Machines Think*. Cambridge, MA : The MIT Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1910393.

Gibert, Martin. 2019. « « Ethique artificielle », version Grand Public ». Dans, édité par M. Kristanek. http://encyclo-philo.fr/etique-artificielle-gp/.

Gibert, Martin. 2020. *Faire la morale aux robots : Une introduction à l'éthique des algorithmes*. Atelier 10. https://www.amazon.ca/-/fr/Martin-Gibert-ebook/dp/B088GLD7D1.

Gilligan, Carol. 2008. *Une voix différente : pour une éthique du care*. Champs [Flammarion [Firme]]. Essais 844; Paris : Flammarion.

Gingras, Yves, et Maxime Colleret. 2020. « Element AI, un « fleuron » ? » *La Presse*+, 5 décembre 2020, sect. DÉBATS. https://plus.lapresse.ca/screens/a14fee43-0f24-4a82-a528-187c02f0d42c__7C___0.html.

Good, Irving John. 1966. « Speculations Concerning the First Ultraintelligent Machine**Based on talks given in a Conference on the Conceptual Aspects of Biocommunications, Neuropsychiatric Institute, University of California, Los Angeles, October 1962; and in the Artificial Intelligence Sessions of the Winter General Meetings of the IEEE, January 1963 [1, 46].The first draft of this monograph was completed in April 1963, and the present slightly amended version in May 1964.I am much indebted to Mrs. Euthie Anthony of IDA for the arduous task of typing. » Dans *Advances in Computers*, édité par Franz L. Alt et Morris Rubinoff, 6:31 — 88. Elsevier. https://doi.org/10.1016/S0065-2458(08)60418-0.

Google. 2018. « Artificial Intelligence at Google: Our Principles ». Google AI. 2018. https://ai.google/principles/.

Google. 2019a. « Perspectives on Issues in AI Governance ». https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

Google. 2019b. « Responsible Development of AI ». https://ai.google/static/documents/responsible-development-of-ai.pdf.

Gordon, David, et Mises Institute. 2007. « John Stuart Mill, Lysander Spooner and Herbert Spencer ». The History of Political Philosophy: From Plato to Rothbard. https://podcasts.apple.com/ca/podcast/the-history-of-political-philosophy-from-plato-to-rothbard/id380678853.

Gouvernement du Canada, Conseil du Trésor du Canada. 2019. « Directive sur la prise de décision automatisée ». https://www.tbs-sct.gc.ca/pol/doc-fra.aspx?id=32592.

Government of the United States, National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee. 2016. « The National Artificial Intelligence Research and Development Strategic Plan ». https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf.

Greene, Gretchen. 2018. « Potholes, Rats and Criminals: How to Think about the Ethics of AI in Government ». *Government Technology*. 20 avril 2018. https://www.govtech.com/data/Potholes-Rats-and-Criminals-How-to-Think-about-the-Ethics-of-AI-in-Government.html.

Grondin, Jean. 2010. « L'herméneutique ». *Que sais-je ?* 2e éd. (3758). http://www.cairn.info/revue-que-sais-je-2008-3758-p-3.htm.

Groupe d'experts ad hoc (GEAH) pour l'élaboration d'un projet de recommandation sur l'éthique de l'intelligence artificielle. 2020. « Document final : première version du projet de recommandation sur l'éthique de l'intelligence artificielle — UNESCO Bibliothèque Numérique ». UNESDOC — Bibliothèque numérique. https://unesdoc.unesco.org/ark:/48223/pf0000373434_fre.

Groupe d'experts de haut niveau sur l'IA (GEHN IA). 2019. « Lignes directrices en matière d'éthique pour une IA digne de confiance ». https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Guiton, Amaelle. 2020. « Lois sur le renseignement : quel avenir pour les «boîtes noires» ? » *Libération.fr*, 30 septembre 2020. https://www.liberation.fr/france/2020/09/30/lois-sur-le-renseignement-quel-avenir-pour-les-boites-noires_1800891.

Hagendorff, Thilo. 2019. « The Ethics of AI Ethics. An Evaluation of Guidelines ». https://arxiv.org/ftp/arxiv/papers/1903/1903.03425.pdf.

Hall, Edward. 2020. *Value, Conflict, and Order*. Chicago : University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/V/bo52781681.html.

Hampshire, Stuart. 1976. « A Delicate Balance ». *Index on Censorship* 5 (3) : 88-89. https://doi.org/10.1080/03064227608532563.

Hampshire, Stuart. 1989. *Innocence and Experience*. Cambridge, Mass. : Harvard University Press.

Hampshire, Stuart. 2002. « Justice Is Strife ». *Philosophy & Social Criticism* 28 (6) : 635-45. https://doi.org/10.1177/019145370202800603.

Hampshire, Stuart. 2011. *La justice est conflit*. Inférences. Genève : Markus Haller.

Harari, Yuval Noah. 2018. « Why Technology Favors Tyranny ». *The Atlantic*, octobre 2018. https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/.

Hart, Vi. 2019. « Changing My Mind about AI, Universal Basic Income, and the Value of Data – The Art of Research ». *The Art of Research* (blog). 2019. https://theartofresearch.org/ai-ubi-and-data/.

Haugeland, John. 1989. *Artificial Intelligence: The Very Idea*. MIT Press.

Hébert-Dolbec, Anne-Frédérique. s. d. « Nos robots peuvent-ils être vertueux? » *Le Devoir*. Consulté le 25 mai 2020. https://www.ledevoir.com/lire/579355/essai-nos-robots-peuvent-ils-etre-vertueux.

Heritage, J. 2001. « Conversation Analysis: Sociological ». Dans *International Encyclopedia of the Social & Behavioral Sciences*, édité par Neil J. Smelser et Paul B. Baltes, 2741— 44. Oxford : Pergamon. https://doi.org/10.1016/B0-08-043076-7/02001-5.

Höffe, Otfried. 1993. *Petit dictionnaire d'éthique*. Paris; Fribourg : Cerf; Éditions universitaires Fribourg. https://books.google.ca/books/about/Petit_dictionnaire_d_%C3%A9thique.html?id=VZiPCvb01yoC&redir_esc=y.

Horvitz, Eric, et Mustafa Suleyman. 2016. « Introduction from the Founding Co-Chairs ». The Partnership on AI. 28 septembre 2016. https://www.partnershiponai.org/introduction-from-the-founding-co-chairs/.

House of Commons of the United Kingdom, Science and Technology Committee. 2016. « Robotics and Artificial Intelligence ». Fifth Report of Session 2016–17 HC 145. https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf.

Hsieh, Nien-hê. 2016. « Incommensurable Values ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Spring 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2016/entries/value-incommensurable/.

Hursthouse, Rosalind, et Stephen Darwall. 2003. « Normative Virtue Ethics ». Dans *Virtue Ethics*, 184-2002. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Hursthouse, Rosalind, et Glen Pettigrove. 2018. « Virtue Ethics ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Winter 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/.

Hutcheson, Francis. 2003. « From An Inquiry into the Origine of Our Idea of Virtue ». Dans *Virtue Ethics*, édité par Stephen Darwall, 51— 62. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Institute of Electrical and Electronics Engineers, Incorporated (IEEE). 2016. « Ethically Aligned Design, Version I. A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems ». The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf.

Institute of Electrical and Electronics Engineers, Incorporated (IEEE). 2017. « Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2 ». https://standards.ieee.org/industry-connections/ec/ead-v1.html.

Institute of Electrical and Electronics Engineers, Incorporated (IEEE). 2019. « Ethically Aligned Design, First Edition ». https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined.

Institute of Electrical and Electronics Engineers, Incorporated (IEEE). 2020. « Ethics in Action ». IEEE. 2020. https://ethicsinaction.ieee.org/.

International Business Machines Corporation (IBM). 2019. « Everyday Ethics for Artificial Intelligence ». https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

ITU & the XPRIZE Foundation. 2017. « AI for Good Global Summit #AI for Good. Artificial Intelligence can help solve humanity's greatest challenges. Report ». Genève, Suisse. https://www.itu.int/en/ITU-T/AI/Documents/Report/AI_for_Good_Global_Summit_Report_2017.pdf.

ITU & the XPRIZE Foundation. 2018. « AI for Good Global Summit 2018 Report. Accelerating progress towards the SDGs ». https://2ja3zj1n4vsz2sq9zh82y3wi-wpengine.netdna-ssl.com/wp-content/uploads/2018/12/SDGs-Report.pdf.

Jobin, Anna, Marcello Ienca, et Effy Vayena. 2019. « Artificial Intelligence: The Global Landscape of Ethics Guidelines ». *Nature Machine Intelligence* 1 (9) : 389-99. https://doi.org/10.1038/s42256-019-0088-2.

Johnson, Gabbrielle M. s.d. « Algorithmic Bia: On the Implicit Biases of Social Technology ». *Synthese* 1 (21). Consulté le 17 septembre 2020. https://doi.org/10.1007/s11229-020-02696-y.

Johnson, Khari. 2020. « Microsoft Researchers Create AI Ethics Checklist With ML Practitioners From a Dozen Tech Companies ». *VentureBeat*, 10 mars 2020. https://venturebeat.com/2020/03/10/microsoft-researchers-create-ai-ethics-checklist-with-ml-practitioners-from-a-dozen-tech-companies/.

Johnson, Robert, et Adam Cureton. 2019. « Kant's Moral Philosophy ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Spring 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/kant-moral/.

Kaminski, Margot E., et Andrew D. Selbst. 2019. « The Legislation That Targets the Racist Impacts of Tech ». *The New York Times*, 7 mai 2019, sect. Opinion. https://www.nytimes.com/2019/05/07/opinion/tech-racism-algorithms.html.

Kant, Emmanuel. 1848a. « Fondements de la métaphysique des mœurs ». Dans *Critique de la raison pratique : précédée des Fondements de la métaphysique des mœurs*, par Emmanuel Kant, traduit par Jules-Romain Barni. Libr. Philosophique de Ladrange. https://books.google.ca/books/about/Critique_de_la_raison_pratique.html?id=NO8TAAAAQAAJ&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false.

Kant, Emmanuel. 1848b. « Critique de la raison pratique ». Dans *Critique de la raison pratique : précédée des Fondements de la métaphysique des mœurs*, par Emmanuel Kant, traduit par Jules-Romain Barni. Libr. Philosophique de Ladrange. https://books.google.ca/books/about/Critique_de_la_raison_pratique.html?id=NO8TAAAAQAAJ&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false.

Kolodny, Niko, et John Brunero. 2020. « Instrumental Rationality ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Spring 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/rationality-instrumental/.

Kurzweil, Ray. 2006. *The Singularity Is Near: When Humans Transcend Biology*. New York : Penguin Books.

Lamb, Creig, et Sarah Doyle. 2017. « Future-proof: Preparing young Canadians for the future of work ». Brookfield Institute for Innovation + Entrepreneurship. http://brookfieldinstitute.ca/wp-content/uploads/2017/03/FINAL-FP-report-Onlinev3.pdf.

Lampert, Jay. 2018. « Kierkegaard : Decisionism in Religion. Infinite Futures ». Dans *The Many Futures of a Decision*, 105— 24. London : Bloomsbury Academic. http://www.bloomsburycollections.com/book/the-many-futures-of-a-decision.

Larson, Christina. 2018. « Who Needs Democracy When You Have Data? » *MIT Technology Review*, 20 août 2018. https://www.technologyreview.com/s/611815/who-needs-democracy-when-you-have-data/.

Lazari-Radek, Katarzyna de, et Peter Singer. 2017. *Utilitarianism. A Very Short Introduction*. Oxford, UK : Oxford University Press.

Leben, Derek. 2018. *Ethics for Robots: How to Design a Moral Algorithm*. First edition.. Boca Raton, FL : Routledge, an imprint of Taylor and Francis. https://www.taylorfrancis.com/books/9781351769068.

Leikas, Jaana, Nadhezda Gotcheva, et Raija Koivisto. 2019. « Ethical Framework for Designing Autonomous Intelligent Systems ». *Journal of Open Innovation: Technology, Market, and Complexity* 5 (18) : 1-12. https://doi.org/10.3390/joitmc5010018.

« Letter to Google C.E.O. » (2018) https://static01.nyt.com/files/2018/technology/googleletter.pdf.

Levy, Neil L, et Robert M Ross. 2020. « The Cognitive Science of Fake News ». Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/3nuzj.

Liberty. s. d. « Liberty Wins Ground-Breaking Victory Against Facial Recognition Tech ». *Liberty*. Consulté le 14 août 2020. https://www.libertyhumanrights.org.uk/issue/liberty-wins-ground-breaking-victory-against-facial-recognition-tech/.

Lindbergh, Ben. 2019. « Carry That Weight: Let's Debate the Morality of 'Yesterday' ». *The Ringer*, 2 juillet 2019. https://www.theringer.com/movies/2019/7/2/20678700/beatles-yesterday-ethics-debate.

Lister, Andrew. 2015. « Reasonable Pluralism ». Dans *The Cambridge Rawls Lexicon*, édité par Jon Mandle et David A. Reidy, 700— 702. 180. Cambridge : Cambridge University Press. https://doi.org/10.1017/CBO9781139026741.181.

López-Molina, Naiara Bellio. s. d. « Spain's Largest Bus Terminal Deployed Live Face Recognition Four Years Ago, but Few Noticed ». *AlgorithmWatch*. Consulté le 14 août 2020. https://algorithmwatch.org/en/story/spain-mendez-alvaro-face-recognition/.

Luño, Angel Rodríguez. s.d. « La ética de las instituciones políticas ». Dans *Introducción a la Ética política*, par Angel Rodríguez Luño. Consulté le 18 novembre 2019. http://www.eticaepolitica.net/eticapolitica/IntrEtPol1.pdf.

MacIntyre, Alasdair. 1984. *After Virtue*. 2e éd. Notre Dame, Indiana : University of Notre Dame Press.

Maclure, Jocelyn, et Marie-Noëlle Saint-Pierre. 2018. « Le Nouvel Âge de l'intelligence Artificielle : Une Synthèse Des Enjeux Éthiques ». *Les Cahiers de Propriété Intellectuelle* 30 (3) : 741— 65. https://www.academia.edu/37920007/Le_nouvel_%C3%A2ge_de_lintelligence_artificielle_une_synth%C3%A8se_des_enjeux_%C3%A9thiques.

Maclure, Jocelyn, et Charles Taylor. 2010. *Laïcité et liberté de conscience*. Montréal : Boréal.

Madaio, Michael A., Jennifer Wortman Vaughan, Luke Stark, et Hanna Wallach. 2020. « Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI ». Dans, 1— 14. Honolulu, HI, USA. http://www.jennwv.com/papers/checklists.pdf.

Malli, Nisa, Melinda Jacobs, et Sarah Villeneuve. 2018. « Intro to AI for Policymakers: Understanding the Shift ». Brookfield Institute. http://brookfieldinstitute.ca/research-analysis/intro-to-ai-for-policymakers/.

Manheim, Karl M., et Lyric Kaplan. 2019. « Artificial Intelligence: Risks to Privacy and Democracy ». SSRN Scholarly Paper ID 3273016. Rochester, NY : Social Science Research Network. https://papers.ssrn.com/abstract=3273016.

Mantelero, Alessandro. 2018. « Report on Artificial Intelligence. Artificial Intelligence and data protection: Challenges and envisaged remedies ». Consultative Committee of Convention for the protection of individuals with regard to automatic processing of personal data (Convention 108). https://rm.coe.int/report-on-artificial-intelligence-artificial-intelligence-and-data-pro/16808b2e39.

Manuguerra-Gagné. 2018. « Les embouteillages dans la ligne de mire de l'intelligence artificielle ». *Radio-Canada*, 2018. https://ici.radio-canada.ca/nouvelle/1137639/embouteillages-intelligence-artificielle-bouchon-circulation-voiture-autonome-auto.

Martin, Dominic. 2017. « Who Should Decide How Machines Make Morally Laden Decisions? » *Science and Engineering Ethics* 23 (4) : 951-67. https://doi.org/10.1007/s11948-016-9833-7.

Martin, Dominic. 2018. « Shedding light on confusion around AI and work ». *Policy Options*, 26 février 2018. http://policyoptions.irpp.org/magazines/february-2018/shedding-light-on-confusion-around-ai-and-work/.

McAfee, Andrew, et Erik Brynjolfsson. 2017. « Human Work in the Robotic Future », 26 janvier 2017. https://www.foreignaffairs.com/articles/2016-06-13/human-work-robotic-future.

McAleer, Michael. 2018. « Philosophers Key as Artificial Intelligence and Biotech Advance ». The Irish Times. 2018. https://www.irishtimes.com/business/innovation/philosophers-key-as-artificial-intelligence-and-biotech-advance-1.3629588.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, et Claude E. Shannon. 2006. « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 ». *AI Magazine* 27 (4) : 12-12. https://doi.org/10.1609/aimag.v27i4.1904.

McDonald, Henry. 2020. « Home Office to Scrap "racist Algorithm" for UK Visa Applicants ». *The Guardian*, 4 août 2020. http://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants.

McDowell, John. 2003. « Virtue and Reason ». Dans *Virtue Ethics*, édité par Stephen Darwall, 121—39. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

McNamara, Andrew, Justin Smith, et Emerson Murphy-Hill. 2018. « Does ACM's Code of Ethics Change Ethical Decision Making in Software Development? » Dans *Proceedings of the 26th ACM Joint European Software Engineering Conference and Symposium on the Foundations of*

Software Engineering (ESEC/FSE '18), November 4–9, 2018, Lake Buena Vista, FL, USA, 1-5. New York : ACM. https://people.engr.ncsu.edu/ermurph3/papers/fse18nier.pdf.

Medhora, Rohinton P. 2018. « Three Paths Towards Global Governance of Artificial Intelligence ». *Centre for International Governance Innovation* (blog). 29 octobre 2018. https://www.cigionline.org/articles/three-paths-towards-global-governance-artificial-intelligence.

Merleau-Ponty, Maurice. 2016. *Phénoménologie de la perception*. Collection Tel ; 4. Paris : Gallimard.

Metz, Cade. 2019. « Is Ethical A.I. Even Possible? » *The New York Times*, 5 mars 2019, sect. Business. https://www.nytimes.com/2019/03/01/business/ethics-artificial-intelligence.html.

Mhlambi, Sabelo. 2020. « From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance ». *Carr Center Discussion Paper Series, Harvard Kennedy School*, nº 2020-009. https://carrcenter.hks.harvard.edu/publications/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial.

Microsoft. 2018a. « Microsoft AI Principles ». Microsoft. 2018. https://www.microsoft.com/en-us/ai/our-approach-to-ai.

Microsoft. 2018b. *The Future Computed. Artificial Intelligence and Its Role in Society*. Redmond, WA : Microsoft Corporation. https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/2018/02/The-Future-Computed_2.8.18.pdf.

Mill, John Stuart. 2003. « Utilitarianism ». Dans *Consequentialism*, édité par Stephen Darwall. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Millar, Jason. 2016. « An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars ». *Applied Artificial Intelligence* 30 (8) : 787–809. https://doi.org/10.1080/08839514.2016.1229919.

Millar, Jason, Nick Novelli, Anne Boily, Carlos Ignacio Gutierrez, Courtney Doagoo, Kathryn Bouskill, Elizabeth Wright, et al. 2019. « Mob.Ly App Makes Driving Safer by Changing How Drivers Navigate | AI Pulse ». *AI Pulse : Accessible, Cutting-Edge AI Policy Research* (blog). 2019. https://aipulse.org/mob-ly-app-makes-driving-safer-by-changing-how-drivers-navigate/.

Mittelstadt, Brent. 2019. « AI Ethics – Too Principled to Fail? » *SSRN Electronic Journal*, janvier. https://doi.org/10.2139/ssrn.3391293.

Moore, G.E. 2003. « Principia Ethica ». Dans *Consequentialism*, édité par Stephen Darwall, 89— 92. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Morley, Jessica, Luciano Floridi, Libby Kinsey, et Anat Elhalal. 2019. « 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]*, septembre. http://arxiv.org/abs/1905.06876.

Müller, Vincent C. 2019. « Policy Documents & Institutions - ethical, legal and socio-economic issues of robotics and artificial intelligence ». Philosophy & Theory of Artificial Intelligence. 2019. http://www.pt-ai.org/TG-ELS/policy/.

Müller, Vincent C. 2020. « Ethics of Artificial Intelligence and Robotics ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/ethics-ai/.

National Institute of Standards and Technology, U.S. Department of Commerce. 2019. « U.S. Leadership in AI : A Plan for Federal Engagement in Developing Technical Standards and Related Tools. » https://www.nist.gov/sites/default/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.

Newman, Jessica Cussins. s. d. « Decision Points in AI Governance: Three Case Studies Explore Efforts to Operationalize AI Principles ». UC Berkeley Center for Long-Term Cybersecurity White Paper Series. Consulté le 4 août 2020. https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf.

Nguyen, Dang Khoa. 2018. « L'intelligence artificielle pour un meilleur contrôle de l'épilepsie réfractaire ». *UdeM nouvelles*, 1 mai 2018. https://nouvelles.umontreal.ca/article/2018/11/27/l-intelligence-artificielle-pour-un-meilleur-controle-de-l-epilepsie-refractaire/.

Nielsen, Karen M. 2007. « Dirtying Aristotle's Hands? Aristotle's Analysis of 'Mixed Acts' in the Nicomachean Ethics III, 1 » Phronesis (52) : 270-300. https://www.jstor.org/stable/40387934.

Nilsson, Nils J. s. d. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Web Version Print version published by Cambridge University Press http://www.cambridge.org/us/0521122937. Consulté le 26 novembre 2020. https://ai.stanford.edu/~nilsson/QAI/qai.pdf.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York : New York University Press.

Normandin, Pierre-André. 2019. « Services : Montréal misera sur l'intelligence artificielle ». *La Presse*, 19 mars 2019. https://www.lapresse.ca/actualites/grand-montreal/201903/18/01-5218722-services-montreal-misera-sur-lintelligence-artificielle.php.

Nørskov, Marco, et Raffaele Rodogno. Forthcoming. « The Automation of Ethics: The Case of Self-Driving Cars ». Dans *Designing Robots — Designing Humans*, édité par D.M. Søndergaard et C. Hasse. Routledge. https://pure.au.dk/portal/files/143302936/The_automation_of_ethics_Revised_FInal.pdf.

Nørskov, Marco, et Søren Schack Andersen, éd. 2016. *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016-TRANSOR 2016*. Frontiers in Artificial Intelligence and Applications, Volume 290. Amsterdam, Netherlands: IOS Press.

Nurock, Vanessa. 2019. « L'intelligence artificielle a-t-elle un genre ? » *Cites* N° 80 (4) : 61— 74. https://www.cairn.info/revue-cites-2019-4-page-61.htm.

Nussbaum, Martha C. 2000. « Why Practice Needs Ethical Theory: Particularism, Principle, and Bad Behavior ». Dans *Moral Particularism*, édité par Brad Hooker et Margaret Olivia Little, 227–55. Oxford University Press.

Nussbaum, Martha C. 1990. « Aristotelian Social Democracy ». Dans *Liberalism and the Good*, édité par R. Bruce Douglass, Gerald M. Mara, Richardson, Henry S., et Georgetown University, 203— 52. New York, N.Y. : Routledge.

Nussbaum, Martha C. 1995. « Aristotle on Human Nature and the Foundations of Ethics ». Dans *World, Mind, and Ethics. Essays on the Ethical Philosophy of Bernard Williams.*, édité par J.E.J. Altham. Cambridge : Cambridge University Press. https://doi.org/10.1017/CBO9780511621086.

Nussbaum, Martha C. 2009. « Bernard Williams : Tragedies, Hope, Justice ». Dans *Reading Bernard Williams*, édité par Daniel Callcut, 213— 38. London ; New York : Routledge. https://books.google.ca/books?id=gqZ5AgAAQBAJ&printsec=frontcover&hl=fr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

Nussbaum, Martha C. 2016. *La fragilité du bien : fortune et éthique dans la tragédie et la philosophie grecques*. Polemos. Paris : Éditions de l'éclat.

ÓhÉigeartaigh, Seán S., Jess Whittlestone, Yang Liu, Yi Zeng, et Zhe Liu. 2020. « Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance ». *Philosophy & Technology*, mai. https://doi.org/10.1007/s13347-020-00402-x.

Omohundro, Steve. 2014. « Autonomous Technology and the Greater Human Good ». *Journal of Experimental & Theoretical Artificial Intelligence* 26 (3) : 303-15. https://doi.org/10.1080/0952813X.2014.895111.

OpenAI. 2018. « OpenAI Charter ». OpenAI. 2018. https://blog.openai.com/openai-charter/.

Organisation for Economic Co-operation and Development (OECD). 2019. « OECD Principles on Artificial Intelligence ». OECD. Better Policies for Better Lives. mai 2019. https://www.oecd.org/going-digital/ai/principles/.

Organisation for Economic Co-operation and Development (OECD). 2020a. « Parliamentarians: OECD Global Parliamentary Network ». OECD. Better Policies for Better Lives. 2020. https://www.oecd.org/parliamentarians/.

Organisation for Economic Co-operation and Development (OECD). 2020b. « The OECD Artificial Intelligence Policy Observatory ». OECD.AI : Policy Observatory. 2020. https://oecd.ai/.

Panic, Branka. 2020. « AI Explained: Non-Technical Guide for Policymakers ». Technology. https://www.slideshare.net/BrankaAcademic/ai-explained-nontechnical-guide-for-policymakers.

Partnership on AI. s.d.a. « Frequently Asked Questions, dans "About Us" ». Partnership on AI. s.d. Consulté le 16 avril 2020. https://www.partnershiponai.org/about/.

Partnership on AI. s.d.b. « Publication Norms for Responsible AI ». The Partnership on AI. s.d. Consulté le 20 mai 2020. https://www.partnershiponai.org/case-study/publication-norms/.

Pasquale, Frank. 2015. *Black Box Society : Les Algorithmes Secrets qui Contrôlent L'économie et L'information*. Collection « Présence ». Limoges : FYP éditions.

Pettit, Philip. 1991. « Consequentialism ». Dans *A Companion to Ethics*, édité par Peter Singer. Oxford : Blackwell.

Pilon-Larose, Hugo. 2020. « Application de traçage : l'opposition craint un «faux sentiment de sécurité» ». *La Presse*, 12 août 2020. https://www.lapresse.ca/covid-19/2020-08-12/application-de-tracage-l-opposition-craint-un-faux-sentiment-de-securite.php.

Piper, Kelsey. 2018. « The case for taking AI seriously as a threat to humanity ». *Vox*, 21 décembre 2018. https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment.

Platon. s. d. *Timée*. Philosophie, Volume 8 : version 1.01. La Bibliothèque électronique du Québec (Édition de référence : Classiques Garnier). Consulté le 25 août 2020. https://beq.ebooksgratuits.com/Philosophie/Platon-Timee.pdf.

Postman, Neil. 1986. *Se distraire à en mourir*. Paris : Flammarion.

Postman, Neil. 1993. *Technopoly: The Surrender of Culture to Technology*. 1st ed.. New York : Knopf.

Postman, Neil. 1998. « Five Things We Need to Know About Technological Change ». Dans. Denver, Colorado. https://www.student.cs.uwaterloo.ca/~cs492/papers/neil-postman--five-things.html.

Price, Richard. 2003. « A Review of the Principal Questions in Morals ». Dans *Deontology*, édité par Stephen Darwall, 35— 54. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Pringle, Ramona. 2019. « The Writing of This AI Is so Human That Its Creators Are Scared to Release It ». *CBC News*, 25 février 2019. https://www.cbc.ca/news/technology/ai-writer-disinformation-1.5030305.

Prunkl, Carina, et Jess Whittlestone. 2020. « Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society ». *arXiv:2001.04335 [cs]*, janvier. http://arxiv.org/abs/2001.04335.

Queffelec, Derwell. 2019. « Le panoptique à l'origine de la société de surveillance ». France Culture. 4 décembre 2019. https://www.franceculture.fr/societe/le-panoptique-a-lorigine-de-la-societe-de-surveillance.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, et Parker Barnes. 2020. « Closing the AI Accountability Gap: Defining an End-To-End Framework for Internal Algorithmic Auditing ». Dans *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. FAT* '20. Barcelona, Spain: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873.

Rawls, John. 2005. *Political Liberalism*. Columbia University Press. https://books.google.ca/books/about/Political_Liberalism.html?id=vXGZRYCkaNsC&redir_esc=y.

Rawls, John. 2009. *Théorie de la justice*. Points. Essais ; 354. Paris : Éditions Points.

Reisman, Dillon, Jason Schultz, Kate Crawford, et Meredith Whittaker. 2018. « Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability ». AI Now Institute. https://ainowinstitute.org/aiareport2018.pdf.

Ricœur, Paul. 1996. *Soi-même comme un autre*. Points; Essais. Paris : Seuil.

Robertson, Kate, Cynthia Khoo, et Yolanda Song. 2020. « To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada ». Citizen Lab (Munk School of Global Affairs & Public Policy, University of Toronto) and the International Human Rights Program (Faculty of Law, University of Toronto). https://citizenlab.ca/wp-content/uploads/2020/09/To-Surveil-and-Predict.pdf.

RodriguezRamos, Jaime. 2019. « A Jesuit Priest, a Singularity Prophet and Elon Musk ». *Medium*, 16 septembre 2019. https://medium.com/@jaimerrf/a-jesuit-priest-a-singularity-prophet-and-elon-musk-5e7c18f3d940.

Roochnik, David. 2009. « What is Theoria? Nicomachean Ethics Book 10.7–8 ». *Classical Philology* 104 (1) : 69-82. https://doi.org/10.1086/603572.

Ropert, Pierre. 2014. « La société de surveillance de Foucault ». France Culture. 13 juin 2014. https://www.franceculture.fr/philosophie/la-societe-de-surveillance-de-foucault.

Ross, W.D. 2003. « The Right and the Good ». Dans *Deontology*, édité par Stephen Darwall, 55— 80. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Russell, Stuart, Sabine Hauert, Russ Altman, et Manuela Velosco. 2015. « Ethics of Artificial Intelligence ». *Nature* 521 (7553) : 415-18. https://www.nature.com/news/robotics-ethics-of-artificial-intelligence-1.17611.

Russell, Stuart J., et Peter Norvig. 2010. *Artificial Intelligence. A Modern Approach*. 3e éd. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, N.J. : Pearson Education, Inc.

Sample, Ian. 2018. « Thousands of Leading AI Researchers Sign Pledge Against Killer Robots ». *The Guardian*, 18 juillet 2018, sect. Science. https://www.theguardian.com/science/2018/jul/18/thousands-of-scientists-pledge-not-to-help-build-killer-ai-robots.

Sandel, Michael J. 2016. *Justice*. Albin Michel.

Sayre-McCord, Geoff. 2014. « Metaethics ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Summer 2014. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2014/entries/metaethics/.

Schiff, Daniel, Jason Borenstein, Justin Biddle, et Kelly Laas. 2020. « What's Next for AI Ethics, Policy, and Governance? A Global Overview ». Dans *2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES'20)*, 153-58. https://dl.acm.org/doi/10.1145/3375627.3375804.

Schwab, Klaus. 2017. *La quatrième révolution industrielle*. Traduit par Jean-Louis. Clauzier et Laurence traductrice). Coutrot. http://catalogue.bnf.fr/ark:/12148/cb451944875.

Scott, Zach. 2018. « Data Science's Reproducibility Crisis ». *Medium*, mai. https://towardsdatascience.com/data-sciences-reproducibility-crisis-b87792d88513.

Searle, John. 1980. « Minds, Brains and Programs ». *Behavioral and Brain Sciences* 3 (3) : 417-57. http://cogprints.org/7150/1/10.1.1.83.5248.pdf.

Selinger, Evan. 2019. « The Efficiency Delusion ». *OneZero*, avril. https://onezero.medium.com/the-efficiency-delusion-f6a97241e1e1.

Sen, Amartya. 1980. « Plural Utility ». *Proceedings of the Aristotelian Society* 81 : 193-215. https://www.jstor.org/stable/4544973.

Sen, Amartya. 2009. *The Idea of Justice*. Harvard University Press. https://books.google.ca/books?id=enqMd_ze6RMC.

Serebrin, Jacob. 2019. « E is for Ethics in AI — and Montreal's Playing a Leading Role ». *Montreal Gazette*, 30 mars 2019. https://montrealgazette.com/news/local-news/can-montreal-become-a-centre-not-just-for-artificial-intelligence-but-ethical-ai?fbclid=IwAR0gkm4iy8fQtEZGGxOJ7HWUsi2kJYG4k8iM0UNN6SQ1LyxmzULV-eNcBKg.

Shane, Scott, et Daisuke Wakabayashi. 2018. « 'The Business of War': Google Employees Protest Work for the Pentagon ». *The New York Times*, 4 avril 2018, sect. Technology. https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html.

Shapiro, Joel, et Reid Blackman. s. d. « Four Steps for Drafting an Ethical Data Practices Blueprint ». *TechCrunch* (blog). Consulté le 4 août 2020. https://social.techcrunch.com/2020/07/24/four-steps-for-an-ethical-data-practices-blueprint/.

Shestakofsky, Benjamin. 2017. « Working Algorithms: Software Automation and the Future of Work ». *Work and Occupations* 44 (4) : 376-423. https://doi.org/10.1177/0730888417726119.

Shipman, Matt. 2018. « Code of Ethics Doesn't Influence Decisions of Software Developers ». NC State News. 2018. https://news.ncsu.edu/2018/10/software-developer-ethics/.

Shorey, Samantha, et Philip N. Howard. 2016. « Automation, Big Data, and Politics: A Research Review ». *International Journal of Communication*, nº 10 : 5032–5055. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2016/10/shoreyhoward.pdf.

Sidgwick, Henry. 2011. *The Methods of Ethics*. Cambridge Library Collection - Philosophy. Cambridge : University Press. http://dx.doi.org/10.1017/CBO9781139136617.

Simon, Robert L. 2002. *The Blackwell Guide to Social and Political Philosophy*. Édité par Blackwell Reference Online. Blackwell Philosophy Guides. Malden, Mass.: Blackwell. http://onlinelibrary.wiley.com/book/10.1002/9780470756621.

Simonite, Tom. 2020. « Google Offers to Help Others With the Tricky Ethics of AI ». *Wired*, 28 août 2020. https://www.wired.com/story/google-help-others-tricky-ethics-ai/.

Sinnott-Armstrong, Walter. 2019. « Consequentialism ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/consequentialism/.

Slote, Michael. 2003. « Agent-Based Virtue Ethics ». Dans *Virtue Ethics*, édité par Stephen Darwall, 203— 26. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Smart, J.J.C. 1997. « Esquisse d'un système de l'éthique utilitariste ». Dans *Utilitarisme. Le pour et le contre*, édité par J.J.C. Smart et Bernard Williams, 9— 69. Le champ éthique 30. Genève : Labor et Fides.

Smellie, Sarah. 2019. « Ethics and Artificial Intelligence: These Researchers Say Tech Has to Have a Moral Backbone ». *CBC*, 20 juin 2019. https://www.cbc.ca/news/canada/newfoundland-labrador/nl-scientist-speaking-up-about-china-1.5179788.

Snyder Caron, Mirka. 2020. « Réponse à la Commission d'accès à l'information du Québec portant sur les amendements potentiels à la Loi sur la protection des renseignements personnels dans le secteur privé particuliers à l'intelligence artificielle. » Montreal AI Ethics Institute. L'Institut d'éthique en intelligence artificielle de Montréal. https://montrealethics.ai/wp-content/uploads/2020/04/FINAL-RAPPORT-MAIEI-AU-CAIQ.pdf.

Snyder Caron, Mirka, et Abhishek Gupta. 2019. « The Social Contract for AI ». https://aiforgood2019.github.io/papers/IJCAI19-AI4SG_paper_36.pdf.

Spaemann, Robert. 1997. *Bonheur et bienveillance. Essai sur l'éthique*. Paris : Presses Universitaires de France.

Spaemann, Robert. 1999. *Notions fondamentales de morale*. Champ essais. Paris : Flammarion.

Sproule, Tessa. 2018. « It's Time to Talk About Ethics in Artificial Intelligence ». *Medium*, 2018. https://medium.com/@TessaSproule123/its-time-to-talk-about-ethics-in-artificial-intelligence-e886133f6f60.

Stark, Philip B. 2018. « Before Reproducibility Must Come Preproducibility ». *Nature* 557 (613). https://doi.org/10.1038/d41586-018-05256-0.

Susskind, Daniel, et Richard E. Susskind. 2015. *The future of the Professions : How Technology Will transform the Work of Human Experts*. Oxford, UK : Oxford University Press. https://ebookcentral.proquest.com/lib/umontreal-ebooks/detail.action?docID=2186874.

Susskind, Jamie. 2018. *Future Politics: Living Together in a World Transformed by Tech*. First edition.. Oxford : Oxford University Press.

Tarantino, Giancarlo. 2017. « Being Wise Before Wisdom: The Historical Development of Phronēsis from Homer to Aristotle, and Its Consequences for Hans-Georg Gadamer's Hermeneutic Ethics ». ProQuest Dissertations Publishing. http://search.proquest.com/docview/1960769510/?pq-origsite=primo.

Taylor, Charles. 1982. « The Diversity of Goods ». Dans *Utilitarianism and Beyond*, édité par Bernard Williams et Amartya Sen, 129— 44. Cambridge University Press. https://doi.org/10.1017/CBO9780511611964.008.

Taylor, Charles. 1985. « Cognitive Psychology ». Dans *Human Agency and Language. Philosophical Papers 1*, par Charles Taylor, 187— 212. Cambridge University Press. 10.1017/CBO9781139173483.

Taylor, Charles. 1992. *Grandeur et misère de la modernité*. L'Essentiel. Québec : Bellarmin.

Taylor, Charles. 1993. « The Motivation Behind a Procedural Ethics ». Dans *Kant and Political Philosophy: The Contemporary Legacy*, édité par Ronald Beiner et William James Booth, 337-60. Yale University Press. https://books.google.ca/books/about/Kant_and_Political_Philosophy.html?id=aPxBoCq454UC&redir_esc=y.

Taylor, Charles. 1994a. « Human Rights, Human Difference ». *Compass*, août, 18— 19.

Taylor, Charles. 1994b. « Justice After Virtue ». Dans *After MacIntyre: critical perspectives on the work of Alasdair MacIntyre*, édité par John Horton et Susan Mendus, 16-43. Notre Dame, Ind. : University of Notre Dame Press.

Taylor, Charles. 1994c. « Reply and Re-Articulation ». Dans *Philosophy in an Age of Pluralism: The Philosophy of Charles Taylor in Question*, édité par Charles Taylor, Daniel M. Weinstock, et James Tully, 213-57. Cambridge : Cambridge University Press.

Taylor, Charles. 1995. « A Most Peculiar Institution ». Dans *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*, édité par Ross Harrison et J.E.J. Altham, 132-55. Cambridge : Cambridge University Press.

Taylor, Charles. 1997a. « Irreducibly Social Goods ». Dans *Philosophical Arguments*, 127— 45. Cambridge, London : Harvard University Press.

Taylor, Charles. 1997b. « La conduite d'une vie et le moment du bien ». Édité par Philippe de Lara. *Esprit (1940—)*, n° 230/231 (3/4) : 151— 73. https://www.jstor.org/stable/24276816.

Taylor, Charles. 1997c. « To Follow A Rule ». Dans, 165— 80. Cambridge, MA : Harvard University Press.

Taylor, Charles. 2002. « Gadamer on the Human Sciences ». Dans *The Cambridge Companion to Gadamer*, édité par Robert J. Dostal, 126— 42. Cambridge companions to philosophy. Cambridge ; New York, Cambridge : Cambridge University Press. http://www.myilibrary.com?id=41822.

Taylor, Charles. 2003. *Les Sources du moi. La formation de l'identité moderne*. Essai. Montréal : Boréal Compact.

Tegmark, Max. 2017. *Life 3.0 : Being Human in the Age of Artificial Intelligence*. New York : Alfred AKnopf.

The Canadian Press. 2019. « Feds set rules on use of AI in government services amid wider testing ». *The Province*, 4 mars 2019. https://theprovince.com/pmn/news-pmn/canada-news-pmn/feds-set-rules-on-use-of-ai-in-government-services-amid-wider-testing/wcm/12c8275c-4009-4a1c-8757-e91f16b5a90c.

The Editors of Encyclopaedia Britannica. 2016. « Teleology ». Dans *Britannica*. Encyclopædia Britannica. https://www.britannica.com/topic/teleology.

The FLI Team. 2017. « A Principled AI Discussion in Asilomar ». Future of Life Institute. 2017. https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/.

Thomson, Judith Jarvis. 2003. « The Trolley Problem ». Dans *Deontology*, par Stephen Darwall, 139— 61. Blackwell Readings in Philosophy. Malden, MA: Blackwell Publishing.

Thornton, T. 2006. « Judgement and the role of the metaphysics of values in medical ethics ». *Journal of Medical Ethics* 32 (6) : 365-70. https://doi.org/10.1136/jme.2005.012518.

Totschnig, Wolfhart. 2019. « The Problem of Superintelligence: Political, Not Technological ». *AI and Society* 34 (4) : 907–920. https://doi.org/10.1007/s00146-017-0753-0.

Tronto, Joan C. 2011. « Who is Authorized to Do Applied Ethics? Inherently Political Dimensions of Applied Ethics ». *Ethical Theory and Moral Practice* 14 (4) : 407-17. https://www.jstor.org/stable/41472608.

Turek, Matt. s.d. « Explainable Artificial Intelligence (XAI) ». Defense Advanced Research Projects Agency (DARPA). s.d. Consulté le 10 octobre 2019. https://www.darpa.mil/program/explainable-artificial-intelligence.

Turing, A. M. 1950. « Computing Machinery and Intelligence ». *Mind*, nº 49 : 433-60. https://www.csee.umbc.edu/courses/471/papers/turing.pdf.

United Nations Development Group (UNDG). 2017. « Data Privacy, Ethics and Protection. Guidance Note on Big Data for Achievement of the 2030 Agenda ». https://undg.org/wp-content/uploads/2017/11/UNDG_BigData_final_web.pdf.

Université Stanford. 2016. « One Hundred Year Study on AI: Artificial Intelligence and Life in 2030 ». https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai_100_report_0901fnlc_single.pdf.

Utrecht Data School, Utrecht University. s.d.b. « App », Data Ethics Decision Aid (DEDA) ». Utrecht Data School. s.d.a. Consulté le 14 juin 2021. https://dataschool.nl/en/deda/app/.

Utrecht Data School, Utrecht University. s.d.a. « Data Ethics Decision Aid (DEDA) ». Utrecht Data School. s.d.b. Consulté le 14 juin 2021. https://dataschool.nl/deda/deda-for-research/.

Utrecht Data School, Utrecht University. s.d.c. « «Poster», sur Data Ethics Decision Aid (DEDA) ». Utrecht Data School. s.d.c. Consulté le 14 juin 2021. https://dataschool.nl/deda/poster/.

Vallor, Shannon. 2016. *Technology and the Virtues : a Philosophical Guide to a Future Worth Wanting*. New York, NY : Oxford University Press. https://atrium.umontreal.ca/primo-explore/fulldisplay?docid=UM-ALEPH002447135&context=L&vid=UM&lang=fr_FR&search_scope=Tout_sauf_articles&adaptor=Local%20Search%20Engine&tab=default_tab&query=creator,contains,shannon%20vallor,AND&sortby=rank&mode=advanced&offset=0.

Vallor, Shannon. 2018. « An Ethical Toolkit for Engineering/Design Practice ». Markkula Center for Applied Ethics at Santa Clara University. 2018. https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/.

Velasquez, Manuel, Claire Andre, et Thomas Shanks, S.J. 2014. « The Common Good ». *Markkula Centre for Applied Ethics, Santa Clara University*, août. https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/the-common-good/.

Wachter, Sandra, Brent Mittelstadt, et Chris Russell. 2020. « Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI ». SSRN Scholarly Paper ID 3547922. Rochester, NY : Social Science Research Network. https://doi.org/10.2139/ssrn.3547922.

Wadell, Kaveh. 2019. « A Tug-of-War Over Biased AI ». *Axios*, 14 décembre 2019. https://www.axios.com/ai-bias-c7bf3397-a870-4152-9395-83b6bf1e6a67.html.

Warnke, Georgia. 2002. « Hermeneutics, Ethics, and Politics ». Dans *The Cambridge Companion to Gadamer*, édité par Robert J. Dostal, 79— 100. Cambridge companions to philosophy. Cambridge ; New York, Cambridge : Cambridge University Press. http://www.myilibrary.com?id=41822.

Weber, Max. 1949. « 'Objectivity' in Social Science and Social Policy ». Dans *The Methodology of the Social Sciences*, édité par Edward A Shils et Henry A. Finch. New York : Macmillan Publishing Co.

Weinberger, David, et Sabelo Mhlambi. 2020. « Q&A : Sabelo Mhlambi on What AI Can Learn from Ubuntu Ethics ». *People + AI Research — Medium*, 6 mai 2020. https://medium.com/people-ai-research/q-a-sabelo-mhlambi-on-what-ai-can-learn-from-ubuntu-ethics-4012a53ec2a6.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. New York : W.H. Freeman and Company.

Werner, Freya. 2019. « It's the Structures, Not the Tech: Os Keyes ». *Exberliner. Berlin in English since 2002*, 13 juin 2019. http://www.exberliner.com/whats-on/disruption-network-lab-os-keyes-interview/.

Wiener, Norbert 1894-1964. 1969. *The Human Use of Human Beings: Cybernetics and Society*. New York : Discus Books, published by Avon Books.

Williams, Bernard. 1981. *Moral Luck: Philosophical Papers, 1973-1980*. Cambridge Cambridgeshire ; New York : Cambridge University Press. https://doi.org/10.1017/CBO9781139165860.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. Cambridge, Mass. : Harvard University Press.

Williams, Bernard. 1993. *Morality. An Introduction to Ethics*. Cambridge University Press. https://doi.org/10.1017/CBO9781107325869.003.

Williams, Bernard. 1994. *La fortune morale : moralité et autres essais*. 1re éd.. Philosophie morale. Paris : Presses universitaires de France.

Williams, Bernard. 1997. « Une critique de l'utilitarisme ». Dans, édité par J.J.C. Smart et Bernard Williams, 73— 141. Le champ éthique 30. Genève : Labor.

Williams, Bernard. 2011. *Ethics and the Limits of Philosophy*. London : Routledge. https://doi.org/10.4324/9780203828281.

Wiltshire, Travis J. 2015. « A Prospective Framework for the Design of Ideal Artificial Moral Agents: Insights from the Science of Heroism in Humans ». *Minds and Machines* 25 (1) : 57-71. https://doi.org/10.1007/s11023-015-9361-2.

Winfield, Alan. 2019. « An Updated Round Up of Ethical Principles of Robotics and AI ». *Alan Winfield's Web Log* (blog). 18 avril 2019. https://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html.

Winner, Langdon. 1980. « Do Artifacts Have Politics? » *Daedalus*, Modern Technology: Problem or Opportunity?, 109 (1) : 121-36. http://www.jstor.org/stable/20024652?origin=JSTOR-pdf.

Wolf, Susan. 2007. « Moral Psychology and the Unity of the Virtues ». *Ratio* 20 (2) : 145-67. https://doi.org/10.1111/j.1467-9329.2007.00354.x.

Woollard, Fiona, et Frances Howard-Snyder. 2016. « Doing vs. Allowing Harm ». Dans *The Stanford Encyclopedia of Philosophy*, édité par Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/doing-allowing/.

World Economic Forum (WEF). s.d.a. « Centre for the Fourth Industrial Revolution ». World Economic Forum. s.d. Consulté le 7 octobre 2019. https://www.weforum.org/centre-for-the-fourth-industrial-revolution/.

World Economic Forum (WEF). s.d.b. « Project Empowering AI Leadership ». World Economic Forum. s.d. (Consulté le 8 octobre 2019) https://www.weforum.org/projects/ai-board-leadership-toolkit.

World Economic Forum (WEF). s.d.c. « Project on Artificial Intelligence and Machine Learning ». World Economic Forum. s.d. Consulté le 8 octobre 2019. https://www.weforum.org/platforms/shaping-the-future-of-technology-governance-artificial-intelligence-and-machine-learning.

World Economic Forum (WEF). s.d.d. « Project TAIE (Teaching AI Ethics) ». World Economic Forum. s.d. Consulté le 8 octobre 2019. https://www.weforum.org/projects/teaching-ai-ethics.

World Economic Forum (WEF). s.d.e « The Fourth Industrial Revolution, by Klaus Schwab ». *World Economic Forum* (blog). Consulté le 27 juillet 2020. https://www.weforum.org/about/the-fourth-industrial-revolution-by-klaus-schwab/.

World Economic Forum (WEF). 2019a. « White Paper: AI Governance. A Holistic Approach ». https://weforum.my.salesforce.com/sfc/p/#b0000000GycE/a/0X000000cPl1/i.8ZWL2HIR_kAnvckyqVA.nVVgrWIS4LCM1ueGy.gBc.

World Economic Forum (WEF). 2019b. « White Paper: Guidelines for AI Procurement ». http://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf.

Wright, Nicholas. 2019. « How Artificial Intelligence Will Reshape the Global Order », 19 février 2019. https://www.foreignaffairs.com/articles/world/2018-07-10/how-artificial-intelligence-will-reshape-global-order.

Wright, Nicholas D. 2020. « Sharp Power and Democratic Resilience Series | Artificial Intelligence and Democratic Norms ». https://www.ned.org/sharp-power-and-democratic-resilience-series-artificial-intelligence-and-democratic-norms/.

Wynsberghe, Aimee van, Denise Soesilo, Kristen Thomasen, et Noel Sharkey. 2018. « Drones in the Service of Society ». Responsible Robotics. https://responsible-robotics-myxf6pn3xr.netdna-ssl.com/wp-content/uploads/2018/08/Drones-in-the-Service-of-SocietyFINAL.pdf.

Yampolskiy, Roman V. 2019. « Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach ». *Medium*, 22 avril 2019. https://medium.com/@romanyam/artificial-intelligence-safety-engineering-why-machine-ethics-is-a-wrong-approach-5fcfa2ca5e75.

Zeng, Yi, Enmeng Lu, et Cunqing Huangfu. 2018. « Linking Artificial Intelligence Principles ». *In the Proceedings of the AAAI Workshop on Artificial Intelligence Safety (AAAI-Safe AI 2018)*. https://arxiv.org/ftp/arxiv/papers/1812/1812.04814.pdf.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. First edition.. New York : PublicAffairs.

Annexe

Tableau 1. – Directives éthiques analysées (Annexe 1.)

#	Origine/ Secteur	Auteur	Titre	Type d'institution	Année	Type de document	Destinataires	Catégorie
1	Privé	Institute of Electrical and Electronics Engineers, Incorporated (IEEE)	Ethically Aligned Design. Version I	Association professionnelle	2016	Étude de standards techniques	Les « technologistes » : « […] toute personne impliquée dans la recherche, la conception, la fabrication ou la diffusion de messages sur l'IA/SA, y compris les universités, les organisations et les entreprises qui font de ces technologies une réalité pour la société. » (4)	Société civile et organisations à multiples partenaires
2	Privé	IEEE	Ethically Aligned Design. Version II	Association professionnelle	2017	Étude de standards techniques	Les « technologistes » (*idem* IEEE 2016)	Société civile et organisations à multiples partenaires
3	Privé	DeepMind	DeepMind Ethics and Society	Entreprise spécialisée en IA	2017	Énoncé d'engagement éthique	Concepteurs et développeurs en IA	Entreprise privée
4	Privé	Open AI	Open AI Charter	Laboratoire de recherche en IA	2018	Charte de principes éthiques	Concepteurs et développeurs en IA	Entreprise privée
5	Privé	Microsoft	Microsoft AI Principles	Multinationale informatique	2018	Charte de principes éthiques	Grand public et clients	Entreprise privée

#	Origine/ Secteur	Auteur	Titre	Type d'institution	Année	Type de document	Destinataires	Catégorie
6	Privé	Microsoft	*The Future Computed*	Multinationale informatique	2018	Livre	« Chefs d'entreprises, décideurs politiques, chercheurs, universitaires et représentants d'ONG » (57)	Entreprise privée
7	Privé	Google	Our Principles	Entreprise de technologie	2018	Charte de principes éthiques	Grand public et clients	Entreprise privée
8	Privé	Google	Perspectives on Issues in AI Governance	Entreprise de technologie	2019	Livre blanc	Gouvernements et société civile	Entreprise privée
9	Privé	Gagné, pour ElementAI	Putting AI Guidelines to Work	Entreprise spécialisée en IA	2019	Énoncé public d'alignement sur les lignes directrices européennes concernant l'IA	Grand public et clients	Entreprise privée
10	Privé	IBM	Everyday *Ethics* for Artificial Intelligence	Multinationale informatique	2019	Rapport/étude éthique	Concepteurs et développeurs en IA	Entreprise privée
11	Privé	Partnership on AI (PAI)	Closing Gaps in Responsible AI	Coalition à but non lucratif	2020	Consultation publique visant un programme	« […] les acteurs du changement, les militants et les décideurs politiques […]. » (s.p.)	Société civile et organisations à multiples partenaires

#	Origine/ Secteur	Auteur	Titre	Type d'institution	Année	Type de document	Destinataires	Catégorie
12	Public	Future of Life Institute (FLI)	Asilomar AI Principles	Organisme de recherche	2017	Charte de principes éthiques	Concepteurs et développeurs en IA et grand public	Société civile et organisations à multiples partenaires
13	Public	Commission mondiale d'éthique des connaissances scientifiques et des technologies UNESCO	Report of COMEST on Robotic Ethics	Agence spécialisée d'une organisation internationale	2017	Rapport éthique	Concepteurs et développeurs en IA et grand public	Organisations gouvernementales, intergouvernementales ou internationales
14	Public	Amnesty International & Access Now	The Toronto Declaration	Organisme à but non lucratif	2018	Déclaration et charte de principes éthiques	Gouvernements et acteurs du secteur privé	Société civile et organisations à multiples partenaires
15	Public	Comité d'élaboration de la Déclaration de Montréal	La Déclaration de Montréal pour un développement responsable de l'IA	Comité interuniversitaire, Université de Montréal	2018	Déclaration et charte de principes éthiques	« […] toute personne, toute organisation de la société civile et toute compagnie désireuses de participer au développement de l'intelligence artificielle de manière responsable […] [et] aux responsables politiques, élus ou nommés […] » (p.6)	Société civile et organisations à multiples partenaires

#	Origine/ Secteur	Auteur	Titre	Type d'institution	Année	Type de document	Destinataires	Catégorie
16	Public	European Commission (EC)	Statement on Artificial Intelligence, Robotics and «Autonomous» Systems	Organisation internationale	2018	Déclaration éthique	Gouvernements des pays européens	Organisations gouvernementales, intergouvernementales ou internationales
17	Public	High-Level Expert Group on AI (EC)	Ethics Guidelines for Trustworthy AI	Groupe d'experts nommé par une organisation internationale	2019	Charte et lignes directrices éthiques	« […] à l'ensemble des parties prenantes de l'IA qui conçoivent, mettent au point, déploient, mettent en œuvre, utilisent l'IA ou sont soumises à ses incidences, et […] aux entreprises, aux organisations, aux chercheurs, aux services publics, organismes gouvernementaux, institutions, organisations de la société civile, particuliers, travailleurs et consommateurs. » (7)	Organisations gouvernementales, intergouvernementales ou internationales
18	Public	G20 Countries	G20 AI Principles	Association de pays	2019	Charte de principes éthiques	Grand public	Organisations gouvernementales, intergouvernementales ou internationales

#	Origine/ Secteur	Auteur	Titre	Type d'institution	Année	Type de document	Destinataires	Catégorie
19	Public	Organisation for Economic Co-operation and Development (OECD)	OECD Principles on Artificial Intelligence	Organisation internationale	2019	Charte de principes éthiques	Gouvernements du monde entier et grand public	Organisations gouvernementales, intergouvernementales ou internationales
20	Public	World Economic Forum (WEF)	AI Governance. A Holistic Approach	Organisation internationale	2019	Livre blanc	Concepteurs et développeurs en IA ainsi qu'aux gouvernements	Organisations gouvernementales, intergouvernementales ou internationales