ATLANTIK-BRÜCKE
CANADA

A Million Splinters of Light:
What is the Path to Ethical AI?

Dominic Martin

JUNE 2023

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS
LIST OF RECCOMMENDATIONS

EXECUTIVE SUMMARY

The main objective of this document is to provide an overview of research and initiatives in artificial intelligence (AI) ethics in the last decades, especially regarding the development of AI ethics frameworks, and make recommendations for future initiatives.

People respond to the impact of technology in various ways and there has been a strong response to the development of artificial intelligence (AI) in the last decade. One of these responses has been the development of ethical frameworks initiatives that aim to orient, oversee, or *frame* the development and usage of the technology.

A compilation of online directories, reports, and studies shows that the cumulative number of AI ethics frameworks increased globally from three frameworks developed during the years 2000 to 2014, inclusively, to 205 in 2020, under conservative estimates. The total number of frameworks is likely to be much larger.

After many calls for new regulations more suited to AI technologies, especially in the second half of the 2010s, the existing regime of AI regulation is likely to change with the roll-out, in the coming years, of new legislation in Canada, Germany, the EU, the United States, and other jurisdictions.

Private organizations, especially in the big tech industry, have been an important source of AI ethics frameworks, almost as important as governments. However, the private sector's response to AI regulation is complex and not homogeneous.

There is strong agreement that AI should be more ethical and that we should minimize the negative impacts of this technology on society. However, there is still debate about what constitutes 'ethical AI' and which requirements, standards, practices, and laws are needed for its realization. A series of considerations and recommendations need to be kept in mind regarding the development of AI ethics frameworks:

- We must ensure that there is enough convergence, proper groundings (in more fundamental moral notions) and coherence in our frameworks' initiatives.
- Expertise in AI ethics is important and we must avoid the instrumentalization of an ethical framework initiative.
- Listing axiomatic principles is the most popular approach for developing ethical frameworks, however, we must pay more attention to our views about social justice, the universal nature of a framework and possible conflicts in the application of multiple principles.
- AI ethics frameworks need to stay relevant. They also need to capture moral truths. It is difficult to achieve these goals without multiple iterations for the development of a framework.

RÉSUMÉ

Cette note a pour objectif de faire la synthèse de la recherche en matière d'intelligence artificielle (IA) au cours des dernières décennies et de donner un aperçu des initiatives les plus récentes, en se concentrant particulièrement sur le développement des cadres éthiques qui s'appliquent à l'IA, dans le but de faire des recommandations pour l'avenir.

Les individus répondent aux nouvelles technologies de diverses façons et ce fut le cas dans les derniers dix ans en ce qui concerne la forte réaction suscitée par l'IA. L'une de ces réponses a notamment consisté à imaginer des cadres éthiques visant à orienter, à contrôler ou à encadrer le développement et l'utilisation de cette technologie.

Une évaluation sommaire et conservatrice des guides, des rapports et des études disponibles en ligne montre qu'au niveau mondial, le nombre de cadres éthiques applicables à l'IA est passé de trois propositions pour les années 2000 à 2014 à au moins 205 pour la seule année 2020. Il est probable qu'il y en ait encore davantage.

Les appels nombreux pour de nouvelles réglementations applicables aux technologies d'IA, surtout depuis le milieu des années 2000, rendent inévitables un changement dans le régime juridique qui prévaut aujourd'hui, à la suite de la mise en œuvre de nouvelles législations au Canada, en Europe, aux États-Unis et dans d'autres pays.

Le secteur privé, notamment les entreprises de haute technologie, ont été une source importante de propositions sur l'encadrement éthique de l'IA, à un niveau presque équivalent à celui des gouvernements. En revanche, la réaction du secteur privé n'est pas nécessairement simple ou homogène.

Il existe un accord général que l'IA devrait avoir un caractère plus éthique et que nous devons minimiser ses impacts négatifs sur la société. Mais on débat encore de ce qui se qualifie comme « IA éthique » et des normes, standards, pratiques ou lois qui sont nécessaires à son avènement. Il existe néanmoins une série de considérations et de recommandations qui doivent être prises en compte dans le développement de tout cadre éthique applicable à l'IA :

- Il faut s'assurer que la proposition s'appuie sur un consensus large et cohérent, solidement ancré dans des normes morales fondamentales;
- L'expertise éthique en IA est importante et ne se réduit pas à des considérations instrumentales;
- L'approche la plus populaire quand vient le temps de développer un cadre éthique consiste à faire une liste de principes ou d'axiomes, mais si on est sérieux en matière de justice sociale ou d'équité, il faut privilégier des cadres à portée universelle et résoudre les contradictions qui naissent de l'application de multiples critères;
- Les cadres éthiques doivent rester en phase avec la technologie. Ils doivent aussi refléter un point de vue moral. Il est difficile de concilier ces exigences sans prévoir des itérations fréquentes entre un nouveau cadre et la réalité

INTRODUCTION

People respond to the impact of technology in various ways and there has been a strong response to the development of artificial intelligence (AI) in the last decade. This includes a fair amount of hype, concerns, and proposals to make AI more ethical. However, there is still debate about what constitutes 'ethical AI' and which requirements, standards, practices, and laws are needed for its realization.

One response to the development of AI has been the development of ethical frameworks initiatives that aim to orient, oversee, or *frame* the development and usage of the technology. Actors within governments, the private sector, civil society, non-governmental organizations, research and teaching institutions, intergovernmental organizations, and other groups have developed hundreds of frameworks, under conservative estimates. These frameworks put forward multiple principles, values, guidelines, checklists, and other forms of guidance to make AI more ethical.

If a metaphor is to be permitted, one could say that these frameworks and their content shine like a million splinters of light. But is this lighting a clear path to ethical AI? The main objective of this document is to provide an overview of research and initiatives in AI ethics in the last decades, especially regarding the development of AI ethics frameworks, and make recommendations for future initiatives.

The document is divided in two parts: an overview and a discussion with recommendations. Part one is the overview. The first section of part one goes over the principled response to AI. Among the various ethical framework initiatives launched in the second half of the 2010s, principles of transparency, justice, fairness, non-maleficence, responsibility, and privacy are the most popular. Actors from both Canada and Germany have engaged in multiple initiatives. The second and third sections also give an overview of current regulatory proposals for AI and the response of the private sector, respectively.

Part two is the discussion on the principled response to the development of AI leading to a series of recommendations. First, there is a lack of convergence, proper groundings (in more fundamental moral notions) and coherence in many frameworks' initiatives. This leads to the first three recommendations. In the second section, recommendations 4 to 6, inclusively, emphasize the importance of expertise in AI ethics and avoiding the instrumentalization of an ethical framework initiative.

The third section introduces a typology of six approaches to developing ethical frameworks. Among these approaches, the list of principles is the most popular approach by a large margin. However, recommendations 7 to 10, inclusively, are to pay more attention to our views about social justice, the universal nature of a framework, and possible conflicts in the application of multiple principles. The last recommendation in the fourth section of part two is to have multiple iterations for the development of a framework. AI ethics frameworks need to stay relevant. They also need to capture moral truths. This is difficult to achieve with one iteration of a framework. Our views about what is good or bad can develop over long periods.

I. OVERVIEW: THE LAST DECADE IN AI ETHICS

The years 2022 and 2023 have been marked by the deployment of new large-scale generative AI models such as ChatGPT, Stable Diffusion, Whisper, and DALL-E 2. These systems are part of a type of AI capable of generating text, images, or other media in response to questions or prompts from their user (Manyika et al. 2023). Once again, recent breakthroughs in AI generated both hype and concerns, and it is easy to lose sight of everything that happened in the field during the last ten years.

AI as a field is devoted to building systems that reproduce some functions of human or animal intelligence (Bringsjord and Govindarajulu 2018). An important branch of AI is machine learning (ML), an approach that allows creating system that learn automatically with less direct intervention from humans. Systems developed with this approach often use networks of artificial neurons whose weight is adjusted during a learning phase. Deep learning is part of the machine learning approach and refers to neural networks with multiple or 'deeper' layers (Mitchell 1997; Jordan and Mitchell 2015; Marcus and Davis 2019). Machine learning methods can find patterns in data automatically and this allows automating task, such as facial recognition, that humans cannot describe with a set of rules that are both formal and finite.

Before 2010, ML systems were fairly limited in their ability (Goodfellow, Bengio, and Courville 2016), and confined to university labs and theoretical research. The access to more data and computing power allowed us to train better models and the technology became one of the most important building blocks for new applications.

AI is now an integral part of modern technology, touching many aspects of people's daily lives, from personal assistant devices to health care applications and targeted advertising. During this period, there has been an evolution in our understanding of the social impact and the ethical issues raised by AI. We have developed more nuanced views of the existential risk and the impact of AI on work. Also, it is increasingly being asked why transparency is such an important consideration. But there has been increasing awareness of the issues of accountability, explainability, discrimination and disinformation, see the *APPENDIX – ETHICAL ISSUES AND SOCIAL IMPACT OF AI*. People responded to these issues in different ways.

I-1. PRINCIPLED RESPONSE

Several social actors sought to provide normative guidance regarding the technology. One of the first and strongest reflexes was to develop various ethical frameworks and we witness the proliferation of these initiatives during the second half of the 2010s (Jobin, Ienca, and Vayena 2019; Fjeld et al. 2020; Hagendorff 2020; Tidjon and Khomh 2022).

Broadly speaking, an AI ethics framework can point to any document, recommendation, policy, analysis, position statement or another type of initiative that expresses a moral preference for a defined course of action. These initiatives aim to orient, oversee, or *frame* the development and usage of AI. A framework can include a charter, guidelines, checklists, surveys, training, governance mechanisms, regulation, or other tools or mechanisms that will be useful to actors subscribing to the initiative (Jobin, Ienca, and Vayena 2019; Tidjon and Khomh 2022).

A compilation of the *AI Principles Map* (https://aiethicslab.com/big-picture/), the *AI Ethics Guidelines Global Inventory* (https://inventory.algorithmwatch.org/), Fjeld et al. (2020) report and Jobin, Ienca & Vayena (2019) study (hereafter the compilation) shows a very sharp increase in the number of AI ethics framework develop during the years 2015 to 2019. The cumulative number increased from three frameworks developed during the years 2000 to 2014, inclusively, to 205 in 2020, with a yearly increase of almost 76 frameworks in 2018. The total number of frameworks developed since 2000 amounts to 227. This is based on the first year that a final version (*i.e.,* not a draft) of a framework was published, subsequent versions, if existing, were not counted.

Furthermore, this compilation isn't exhaustive, and the total number of frameworks is likely to be much larger. The scientific reports and studies only include a limited subset of frameworks that have been specifically selected and coded. The online repositories are based on self-registration or self-declaration to various extents, and some organizations may not have registered their ethical framework, especially if they are not public organizations or if the framework was intended for internal use. Many private enterprises may have developed these internal ethical frameworks. Finally, the large number of frameworks already publicized by 2020 may have led some organizations to lose interest in publicizing their initiative. This might be one of the factors explaining why the pace of development plateaued as of 2020 and why fewer new frameworks were registered in the online directories in the following years, see Figure 1.

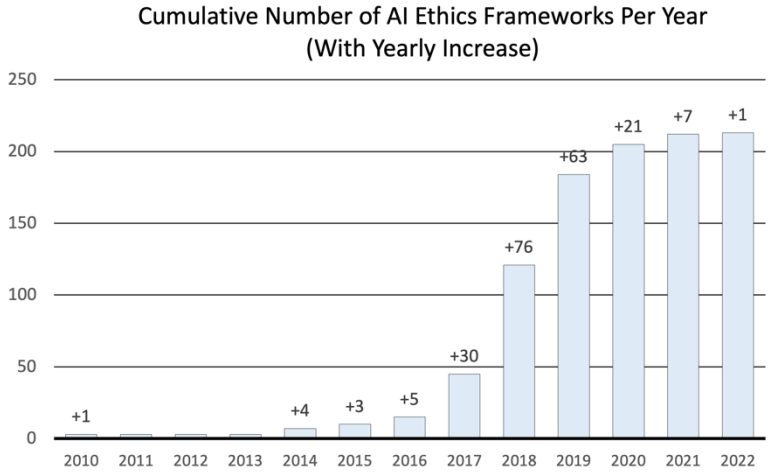

Figure 1: Frameworks in AI Ethics per Year

Source: Compilation of online directories, reports, and studies.

Still, this helps grasp of the magnitude of the phenomenon. Frameworks have been developed by many types of organizations, including governments, intergovernmental organizations, civil society, non-governmental organizations (NGO), research and teaching institutions, professional associations, the private sector, multistakeholder groups, and even political parties and religious organizations. These actors originate from almost forty countries, in addition to the European Union and international organizations.

Anna Jobin, Marcello Ienca and Effy Vayena (2019) conducted a review of 84 frameworks from the gray literature with academic and legal sources excluded. They identify eleven overarching ethical values and principles that reappear in multiple documents with various degrees of popularity, see Table 1. Their results reveal an emerging convergence around the first five principles:

1. Transparency;
2. Justice and fairness;
3. Non-maleficence;
4. Responsibility; and
5. Privacy.

In a recent study, Lionel Tidjon and Foutse Khomh (2022) provide a contextual analysis of 100 frameworks from 29 countries. Figure 2 shows the occurrence of ethical principles per country based on a random sample of these frameworks. The results are consistent with Jobin, Ienca & Vayena's study. In Canada, the most frequent principles are transparency, responsibility, privacy, sustainability, autonomy, and well-being. In Germany and other European countries, transparency is also the most frequent principle, then fairness, security, responsibility, and accountability.

Frameworks (*N*/84)	Ethical principle (bold) and variations
73	**Transparency** Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
68	**Justice and fairness** Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-) discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access, and distribution
60	**Non-maleficence** Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
60	**Responsibility** Responsibility, accountability, liability, acting with integrity
41	**Beneficence** Beneficence, well-being, peace, social good, common good
34	**Freedom and autonomy**

	Freedom and autonomy, consent, choice, self-determination, liberty, empowerment
28	**Trust**
14	**Sustainability** Sustainability, environment (nature), energy, resources (energy)
13	**Dignity**
6	**Solidarity** Solidarity, social security, cohesion

Table 1: Most Popular Principles in a Series of Ethical Frameworks

The number (*N*) represents the number of frameworks with a principle over a total of 84 frameworks.
Source: Jobin, Ienca & Vayena (2019, tbl. 3)

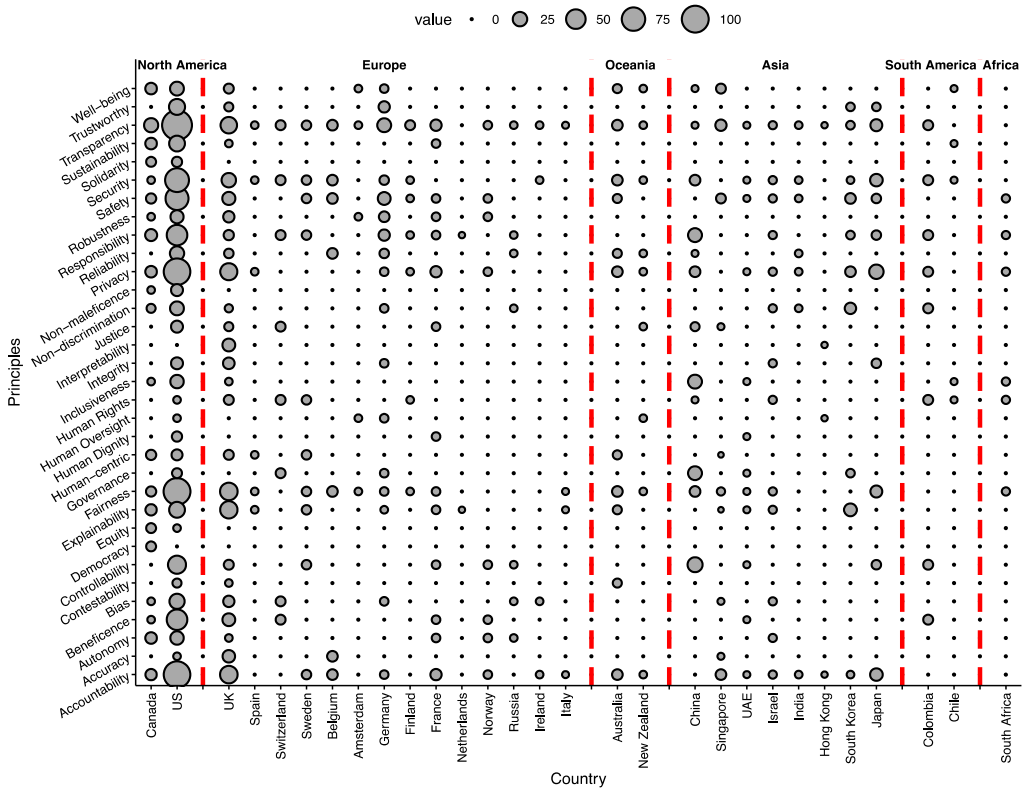

Figure 2: Occurrence of AI Ethics Principles per Country

Source: Tidjon & Khomh (2022, fig. 1).

Many ethics frameworks have been developed in Canada and Germany, especially if we take into account the participation in international initiatives and initiatives originating from the European Union (for Germany). Based on the compilation data, the United States is the most popular country with 49 frameworks (22%), then 41 frameworks originate from international initiatives (18%), 25 from the United Kingdom, 23 from Germany (10%), 10 from the European Union (4%) and 8 from Canada (4%), see Figure 3.

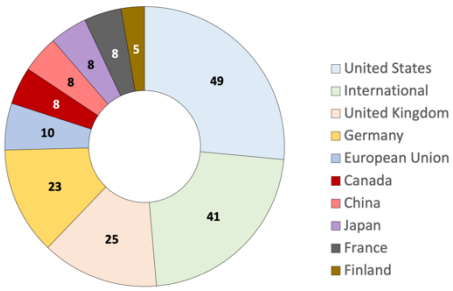

Figure 3: Countries and Geographical Origins with the Most Frameworks

Source: Compilation of online directories, reports, and studies.

I-1.1. AI ETHICS FRAMEWORKS IN CANADA

Most of the frameworks developed in Canada come from the government, and then the civil society or multistakeholder initiatives, with a strong presence of research institutions. The government of Canada officially expressed an ethical stance on AI through at least four different initiatives. First, the government committed to using AI in a manner that is compatible with core principles of administrative law through its *Directive on Automated Decision-Making*. The directive will continue to evolve to remain relevant in a context where technology is changing rapidly. Second, the Treasury Board developed the *Algorithmic Impact Assessment (AIA)* tool, a mandatory risk assessment tool to support the directive. The tool was developed after an extensive consultation campaign with experts and other stakeholders. The tool is itself an algorithm or automated questionnaire that produces an assessment score based on 51 risk and 34 mitigation questions, considering the design and decision type of an algorithm, among other factors. Third, the government produced a *Guideline on Service and Digital* to support the implementation of the Treasury Board Policy and the Directive on Service and Digital. These different initiatives are also grouped under the *Responsible use of artificial intelligence (AI)* framework that features its guiding principles and timeline.

Other influential frameworks in Canada include the *Montreal Declaration for a responsible development of artificial intelligence* that was initiated by the Université of Montréal and co-constructed with stakeholders from various sectors or industries. The declaration identifies 10 principles and values that should be applied to the digital and AI fields: (1) Well-being, (2) Respect for autonomy, (3) Protection of privacy and intimacy, (4) Solidarity, (5) Democratic participation, (6) Equity, (7) Diversity inclusion,

(8), Prudence, (9) Responsibility and (10) Sustainable development (Abrassart et al. 2018). Recommendations are made based on each of these principles to establish guidelines for the digital transition.

The Centre for International Governance Innovation (CIGI) participated in the G20 AI principles initiatives with, among others, the working paper *Toward a G20 Framework for Artificial Intelligence in the Workplace*. Other frameworks include the *Toronto Declaration* led by Amnesty International and Access Now, which expressed views from the global human rights community, see Table 2 for references and a complete list.

2019	***Directive on Automated Decision-Making*** Canada Government (Government) https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592
2019	***Algorithmic Impact Assessment (AIA)*** Canada Government, Office of the Chief Information Officer (OCIO) & Treasury Board of Canada Secretariat (TBS) (Government) https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
	Guideline on Service and Digital Canada Government (Government) https://www.canada.ca/en/government/system/digital-government/guideline-service-digital.html#ToC4_5
2019	***Responsible use of artificial intelligence (AI)*** Canada Government (Government) https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai.html
2018	***Toward a G20 Framework for AI in the Workplace*** Centre for International Governance Innovation (CIGI)(Civil Society and NGO) https://www.cigionline.org/publications/toward-g20-framework-artificial-intelligence-workplace
2018	***Montreal Declaration for a responsible development of artificial intelligence*** Université de Montréal (Multistakeholder) https://www.montrealdeclaration-responsibleai.com/
2018	***Toronto Declaration: Protecting the right to equality in machine learning*** Amnesty International; Access Now (Civil Society and NGO) https://www.torontodeclaration.org/
2019	***Human Ethics in Artificial Intelligence and Big Data Research*** National Research Council Canada (Government)

https://nrc.canada.ca/en/corporate/values-ethics/research-involving-human-participants/advisory-statement-human-ethics-artificial-intelligence-big-data-research-2017

Table 2: Overview of AI Ethics Frameworks in Canada

Source: Compilation of online directories, reports, and studies.

I-1.2. AI ETHICS FRAMEWORKS IN GERMANY

Almost three times more frameworks were developed in Germany in comparison to Canada, not counting participation in international and European initiatives. German governmental organizations developed a similar number of frameworks, but more initiatives originated from the private sector.

One of the first reports comes from the Ethics commission appointed by the Federal Minister of Transport and Digital Infrastructure. The commission published its report in 2017: *Automated and connected automated driving (*Automatisiertes und Vernetztes Fahren*).* The expert group from the Data ethics commission (Daten ethik kommmission) published its *Report of the Data Ethics Commission of the German Federal Government (*Gutachten der Datenethikkommission der Bundesregierung*)* in 2019, developing ethical standards, guidelines, and recommendations for the information age. There are other frameworks developed by the Conference of the independent data protection supervisory authorities in Germany (Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder).

Researchers at the Universities of Bonn and Cologne are developing standards for the inspection and certification of AI applications in a project led by the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) with the participation Federal Office for Information Security (BSI). The team published a white paper presenting the philosophical, ethical, legal, and technological issues that should serve as the basis for the certification: *Trustworthy Use of Artificial Intelligence (*Vertrauenswürdiger Einsatz Von Künstlicher Intelligenz*)*. The Handelsblatt Research Institute and Hochschule der Medien also developed their frameworks. The AI Ethics Impact Group, an interdisciplinary consortium led by VDE Association for Electrical, Electronic & Information Technologies and Bertelsmann Stiftung, also developed a framework entitled *From Principles to Practice: An interdisciplinary framework to operationalise AI ethics*. The framework aims to bring ethical principles into actionable practice as much as possible.

The German industry association Bitkom publicized at least two different frameworks. First, its *Guidelines for the use of Big Data (*Leitlinien für Big Data Einsatz*)* aimed at decision-makers, data protection authorities, private consumers, the general public and the media to help develop public policies and guide practices toward the ethical use of Big Data. Second, a series of *Recommendations for the responsible use of AI and automated decision-making (*Empfehlungen für den verantwortlichen Einsatz von KI und automatisierten Entscheidungen*)* with a broader focus on AI and digital automation. These private sector organizations also developed ethical frameworks: BMW, Bosch, Bundesverband KI, Deutsche Telekom, Ethikbeirat HR Tech, SAP, Verivox, and the Working group "Vernetzte Anwendungen und Plattformen für die digitale Gesellschaft."

Finally, the professional association Gesellschaft für Informatik developed a series of *Ethical Guideliunes*, see Table 3 for references and a more detailed list.

2017	***Automated and connected automated driving* (Automatisiertes und Vernetztes Fahren)** Federal Ministry of Transport and Digital Infrastructure, Ethics Commission (Ethkikkommission BuMi Verkehr und digitale infrastruktur)(Government) https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile
2019	***Report of the Data Ethics Commission of the German Federal Government* (Gutachten der Datenethikkommission der Bundesregierung)** Data ethics commission (Daten ethik kommmission) (Government) https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=4
2019	***Hambach Declaration on Artificial Intelligence – Seven requirements for data protection* (Hambacher Erklärung zur Künstlichen Intelligenz – Sieben datenschutzrechtliche Anforderungen)** Conference of the independent data protection supervisory authorities in Germany (Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder)(Government) https://www.datenschutz.rlp.de/fileadmin/lfdi/Konferenzdokumente/Datenschutz/DSK/Entschliessungen/097_Hambacher_Erklaerung.pdf
2019	***Trustworthy Use of Artificial Intelligence* (Vertrauenswürdiger Einsatz Von Künstlicher Intelligenz)** Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)(Multistakeholder) https://www.iais.fraunhofer.de/en/press/press-release-190702.html
	***Data protection and Big Data* (Datenschutz und Big Data)** Handelsblatt Research Institute (Research and Education Institution) https://www.umweltdialog.de/de-wAssets/docs/2014-Dokumente-zu-Artikeln/leitfaden_unternehmen.pdf
2017	***10 ethical guidelines for the digitalisation of companies* (10 ethische Leitlinien für die Digitalisierung von Unternehmen)** Hochschule der Medien (Research and Education Institution) https://www.hdm-stuttgart.de/digitale-ethik/digitalkompetenz/ethische_unternehmensleitlinien
2020	***From Principles to Practice: An interdisciplinary framework to operationalize AI ethics***

	AI Ethics Impact Group (AIEIG)(Multistakeholder) https://www.ai-ethics-impact.org/en
2015	***Guidelines for the use of Big Data (*Leitlinien für Big Data Einsatz*)*** Bitkom (Private Sector) https://www.bitkom.org/sites/default/files/file/import/150901-Bitkom-Positionspapier-Big-Data-Leitlinien.pdf
2018	***Recommendations for the responsible use of AI and automated decision-making (*Empfehlungen für den verantwortlichen Einsatz von KI und automatisierten Entscheidungen*)*** Bitkom (Private Sector) https://www.bitkom.org/Bitkom/Publikationen/Empfehlungen-fuer-den-verantwortlichen-Einsatz-von-KI-und-automatisierten-Entscheidungen-Corporate-Digital-Responsibility-and-Decision-Making.html
2020	***BMW Group Code of Ethics for AI*** BMW (Private Sector) https://www.press.bmwgroup.com/global/article/detail/T0318411EN/seven-principles-for-ai:-bmw-group-sets-out-code-of-ethics-for-the-use-of-artificial-intelligence?language=en
2020	***Code of Ethics for AI*** Bosch (Private Sector) https://www.bosch.com/stories/ethical-guidelines-for-artificial-intelligence/
2019	***KIBV Quality seal (KIBV Gütesiegel)*** Bundesverband KI (Private Sector) https://ki-verband.de/ki-guetesiegel-ai-made-in-germany
2018	***Guidelines for Artificial Intelligence*** Deutsche Telekom (Private Sector) https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366
2019	***Guidelines for the responsible use of artificial intelligence and other digital technologies in human resources (*Richtlinien für den verantwortungsvollen Einsatz von Künstlicher Intelligenz und weiteren digitalen Technologien in der Personalarbeit*)*** Ethikbeirat HR Tech (Ethics council HR Tech)(Private Sector) https://www.ethikbeirat-hrtech.de/wp-content/uploads/2019/09/Ethikbeirat_und_Richtlinien_Konsultationsfassung_final.pdf
2018	***SAP's guiding principles for Artificial Intelligence*** SAP (Private Sector) https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/
201	***Verivox/Pro7 Commitment (*Selbstverpflichtung*)***

9	Verivox (Private Sector) https://www.verivox.de/company/selbstverpflichtung/
2014	***Charter of digital networking*** Working group "Vernetzte Anwendungen und Plattformen für die digitale Gesellschaft" (Private Sector) https://charta-digitale-vernetzung.de/app/uploads/2016/11/Charter-of-Digital-Networking.pdf
2018	***Ethical Guidelines* (Ethische Leitlinien)** Gesellschaft für Informatik (German Society of Informatics) (Professional Association) https://gi.de/ueber-uns/organisation/unsere-ethischen-leitlinien/

Table 3: Overview of AI Ethics Frameworks in Germany

Source: Compilation of online directories, reports, and studies.

I-2. REGULATORY AND GOVERNANCE-FOCUSED RESPONSES

After many calls for new regulations more suited to AI technologies, especially in the second half of the 2010s, the existing regime of AI regulation is likely to change with the roll-out, in the coming years, of new legislation in Canada, Germany, the EU, the United States and other jurisdictions (Choudhry, Wall, and Reynolds 2023). Furthermore, it is expected that the recent generative AI boom will concentrate the power of the big tech industry even further, which create additional pressures for regulators to act rapidly (Kak and Myers West 2023; Heikkilä 2023).

In Canada, the most significant reform is the upcoming *Artificial Intelligence and Data Act* (AIDA)(Government of Canada 2023). The proposed federal Bill C-27, if passed, would create Canada-wide obligations and prohibitions about the design, development, and use of artificial intelligence systems in the course of international or interprovincial trade and commerce. This would apply to any "technological system that, autonomously or partly autonomously, processes data related to human activities through the use of a genetic algorithm, a neural network, machine learning or another technique in order to generate content or make decisions, recommendations or predictions" (Choudhry, Wall, and Reynolds 2023). Before the introduction of the AIDA, the Office of the Privacy Commissioner (OPC) of Canada also issued recommendations on how to amend the *Personal Information Protection and Electronic Documents Act* (PIPEDA) which would also impact how AI systems are used and developed (https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/). Another initiative is the proposed Bill C-18, the *Act respecting online communications platforms that make news content available to persons in Canada*, that would constrain media platform companies to negotiate deals to pay Canadian media companies for the content they link on their websites (https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c18_1.html).

In Germany, the Study Commission on Artificial Intelligence - Social Responsibility and Economic, Social and Ecological Potential of the 19th German Bundestag presented its final report In October 2020 (see the AI Watch on Germany: https://ai-

watch.ec.europa.eu/countries/germany/germany-ai-strategy-report_en). The Federal Government updated its national AI strategy the following month (Goverment of Germany 2020). The review sets out concrete measures to be implemented in the following fields of action: research, knowledge and expertise, transfer and application, regulatory framework, and society. These include the launch of a *Commission on Competition Law 4.0*, a review of the legislation concerning the use of non-personal data as well as copyright, and the implementation of a cyber security directive.

These initiatives are taking place as the European Union is considering far-reaching legislation on AI with the *Artificial Intelligence (AI) Act*. When effective, the new law will considerably shape the regulatory landscape in Europe and abroad. AI regulation and policies in the United States may also have an impact on Canada and Germany. Finally, initiatives in China should not be overlooked. Officials recently close a consultation on a second round of generative AI regulation which can lead to standards and regulations that will be influential globally (Heath 2023).

I-3. PRIVATE SECTOR RESPONSE

The private sector, and especially the big tech industry, has been a key stakeholder in the development of AI in the last decade. Until 2014, the most significant machine learning models were released by academia. Since then, the industry has taken over. In 2022, there were "32 significant industry-produced machine learning models compared to just three produced by academia" (Manyika et al. 2023, 23).

Private organizations have been an important source of AI ethics frameworks, almost as important as governments. Sixty-six (66) or almost 30 percent of ethical frameworks in the compilation come from private sector organizations, just after the 68 frameworks produced by governments. It is common to see the same organization put forward multiple different frameworks, this is the case with Google, Microsoft, IBM, and the professional services network Deloitte. Civil society organizations and NGOs launched 17 percent of the initiatives with 39 frameworks, then research and education institutions with 11 percent of the initiatives and 25 frameworks. The rest is divided between intergovernmental organizations, multistakeholder initiatives and other types of actors like professional associations, see Figure 4.

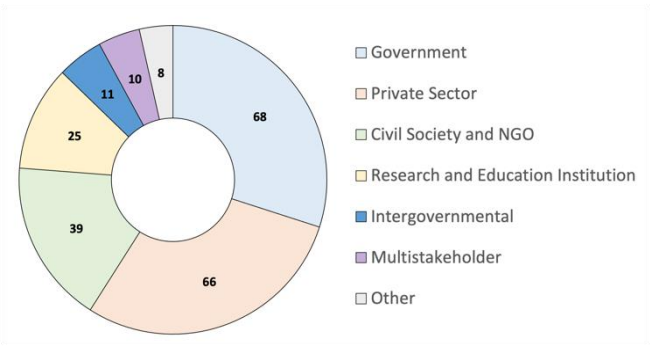

Figure 4: Frameworks Developed by Each Type of Author

Source: Compilation of online directories, reports, and studies.

The private sector's role in, and response to, AI regulation are complex and not homogeneous. The big tech industry often criticizes and opposes regulatory proposals on the basis that this would harm its consumers and would not serve public interests (Canadian Press 2023). However, this regulation is often detrimental to the commercial interests of the industry. At the same time, many people working in the industry seem to understand the need for more normative guidance (Klein 2023). In a recent congressional hearing in the United States Senate, OpenAI's Sam Altman largely agreed with the members of a the hearing subcommittee on the need to regulate the increasingly powerful A.I. technology being created (Kang 2023). An even greater diversity of views is observed outside the tech industry.

I-4. CONCLUSION

Technology often has an impact on society, and people will respond to this impact in various ways. The strength of the response to recent developments in artificial intelligence was particularly strong both in terms of the ethical frameworks that were developed and publicized, and the calls for additional regulation and governance mechanisms.

There have been other disruptive technologies in the pass that generated strong social responses: genetic engineering, nanotechnologies or even the development of information technologies during the first decade of the 2000s. Using these technologies as a basis of comparison may help to see the specificities and the strongness of the response to AI technologies.

II. DISCUSSION AND RECOMMENDATIONS

There is strong agreement that AI should be more ethical and that we should minimize the negative impacts of this technology on society. However, there is still debate about what constitutes 'ethical AI' and which requirements, standards, practices, and laws are needed for its realization. A series of considerations and recommendations need to be kept in mind to ensure we achieve desired outcomes, especially in terms of the ethical frameworks that have been developed during the last decade.

There are two main challenges with framework initiatives. First, an ethical framework must capture and express moral truths (Schwartz 2002). Second, it must generate sufficient uptake, compliance, or adhesion. The second challenge is very different nature, but this is especially important because AI ethics frameworks typically lack mechanisms to enforce their own normative claims (Hagendorff 2020).

II-1. CONVERGENCE, NORMATIVE GROUNDING AND COHERENCE

Ethical frameworks can be divided in two parts: foundation and implementation. First, they build on one or multiple principles, values or idea that serve as the basis for the initiative. For instance, the *Ethics Guidelines for Trustworthy AI* of the High-Level Expert Group on AI (HLEG AI 2019, 11–12) lists four principles that must be respected to "ensure that AI systems are developed, deployed and used in a trustworthy manner": (*i*) respect for human autonomy, (*ii*) prevention of harm, (*iii*) fairness, and (*iv*) explicability. These principles form the foundation of the ethical framework. The

Asilomar AI Principles are a list of 27 principles touching upon notions of transparency, responsibility, and value alignment, among others (Future of Life Institute 2017). The foundation of an ethical framework can also build on a single idea, such as the claim that AI and autonomous systems should be aligned with human morality (IEEE 2017).

The implementation part of a framework includes all the elements by which the framework is put into practice: guidelines, regulations and laws, standards, recommendations on governance mechanisms, checklists, software, training, etc. The implementation of a framework is derived from the foundation. If, for instance, a framework will emphasize the importance of democratic decision-making regarding the usage of AI in society, it may also include recommendations on how to organize civil deliberation forums to discuss and make suggestions on how to use the technology, see Figure 5.

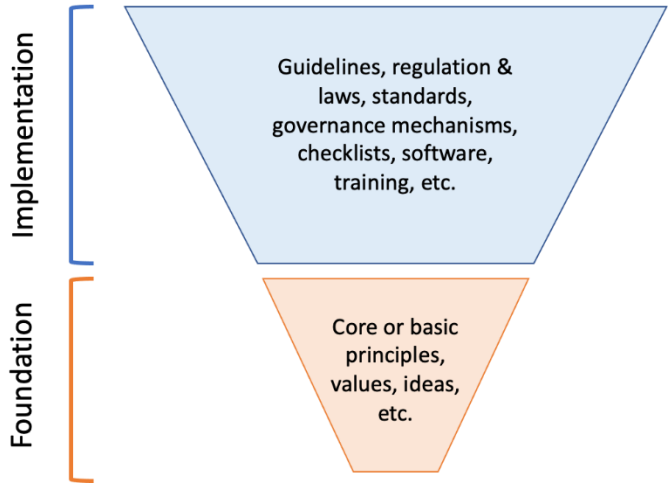

Figure 5: Foundation and Implementation of an Ethical Framework

The overview in the previous section and additional analyses show that a wide diversity of principles or values serve as the foundations for current ethical frameworks. One of the main conclusions of these analyses is the lack of convergence between different principles. Tidjon & Khomh (2022, 1) show the divergences from one country to the other and point out that operationalizing different AI ethics frameworks is difficult given "diversity and context-dependency," which results in gaps between AI ethics principles and their execution.

In their study, Jobin, Ienca & Vayena reach similar, if not more critical, conclusions. What is more, countries are not lumped together in their analysis, which allows us to see divergences between each framework within the same country. Even if 11 principles occur more often in a series of frameworks, their analysis reveals:

> substantive divergences among all 11 ethical principles in relation to four major factors: (1) how ethical principles are interpreted; (2) why they are

> deemed important; (3) what issue, domain or actors they pertain to; and (4) how they should be implemented. These conceptual and procedural divergences reveal uncertainty as to which ethical principles should be prioritized and how conflicts between ethical principles should be resolved, and it may undermine attempts to develop a global agenda for ethical AI. For example, the need for ever-larger, more diverse datasets to 'unbias' AI might conflict with the requirement to give individuals increased control over their data and its use in order to respect their privacy and autonomy. Similar contrasts emerge between avoiding harm at all costs and the perspective of accepting some degree of harm as long as risks and benefits are weighed against each other. Moreover, risk–benefit evaluations are likely to lead to different results depending on whose well-being will be optimized for and by which actors. Such divergences and tensions illustrate a gap at the cross-section of principle formulation and their implementation into practice. (Jobin, Ienca, and Vayena 2019, 8)

Important divergences are even observed within the same organization, especially private organizations. For instance, professional service network Deloitte put forward at least three different AI ethics frameworks in recent years. First, a report exploring the role of ethics in AI and the benefits of AI to governments and public sector entities (Hashmi 2019). Second, an article on the design principles for ethical AI that "can guide leaders when thinking about AI's ethical ramifications" (Guszcza et al. 2020). Third, Deloitte United States sells a *Trustworthy AI*™ framework as a service to its business clients, to help "bridge the ethics gap" between a lack of global AI regulation and business leaders' concerns while adopting the technology (https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html).

There is an overlap between the foundation of these three different frameworks because some of their core normative notions are similar: transparency, justice, fairness, robustness, trustworthiness, and so on. On the other hand, the three frameworks organize these notions in different ways, in different orders, and under different structures. Also, each framework uses other additional normative notions that diverge from one framework to the other, and the link between the different conceptions of ethical AI implied in these frameworks is not clearly stated.

A similar lack of convergence, although perhaps to a lesser extent, can be observed in the ethical framework initiatives of the Canadian government. At least four different frameworks were developed. Some effort has been made to regroup these initiatives under one appellation, the *Responsible use of artificial intelligence (AI)*, but it is not clear how these different documents — each quite rich and extensive in itself — will fit with each other or interact. What is more, the responsible AI overarching initiative also features its guiding principles and timeline.

> Recommendation 1: While developing an ethical framework, an organization or a government should ensure that the foundation of this initiative converges with other existing initiatives within the same organization or government.

Diverging principles is not necessarily an issue if different frameworks within a country or between countries are compared with each other. This could indicate a diversity of views on what constitutes ethical AI. But convergence within the same organization is important for ensuring the moral validity of a series of frameworks and uptake. If two or more initiatives diverge, it is unlikely that they are both able to capture moral truths. But also, people might disregard the normative guidance provided by an organization or a government if they are oriented in different directions.

A lack of convergence can also raise questions about the normative groundings of a framework (Franzke 2022; Stahl 2022; Coeckelbergh 2020). Many frameworks endorse a specific appellation such as trustworthy AI, responsible AI, AI for good, human-centred AI or ethically aligned AI. It is difficult to establish if these appellations, and the frameworks that use these appellations, are based on a systematic understanding of these notions.

For instance, there are at least nine frameworks in the compilation that endorse the appellation 'trustworthy AI,' including the HLEG initiative, one of Deloitte's frameworks, the *Principles for responsible stewardship of trustworthy AI (G20 AI Principles),* IBM's *AI Ethics Framework based on Trust and Transparency Principles*. We can observe important differences in the foundational principles, the structure, and the general content of these frameworks. Similar questions arise regarding other appellations.

> Recommendation 2: An AI ethics framework ought to be properly grounded at the fundamental level. This implies, among other things, systematic appellations derived from more fundamental moral notions.

A broader set of questions touches on the overall coherence of AI ethics frameworks at multiple levels, in addition to their grounding in substantive normative notions. For instance, the HLEG's *Ethics Guidelines for Trustworthy AI* are rooted in four principles. But the group also claims that AI systems should "improve individual and collective wellbeing" and that their principles are "rooted in fundamental rights" (11). They also propose a list of seven non-exhaustive requirements building on these principles: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination, and fairness; 6) societal and environmental wellbeing; and 7) accountability (14).

According to the guidelines, different groups of stakeholders — developers, deployers and end users — must ensure that the requirements are met, but all this bears the question: what exactly one ought to do to follow the guidelines? Is it ultimately a question of improving well-being, articulating human rights, complying with four principles, or following seven requirements? Each of these different elements in the HLEG framework comes with its implications and may lead to different course of action for the stakeholder that would want to put the framework into practice.

The *Asilomar AI principles* (Future of Life Institute 2017) is divided into three categories: research issues, ethics and values, and longer-term issues. But it is not clear why one of these categories is the 'ethics and values' principles when each of the principles is an ethical principle reflecting different values. This also raises questions about the structure and the presentation of the framework. Consider finally the

Declaration of Montreal, one of the most popular frameworks globally with the HLEG guidelines and Asilomar principles. The framework endorses the notion of responsible AI as its main appellation and approach, yet there is also a principle of responsibility among the 11 principles identified in the declaration. Why is the notion of responsibility both the overarching moral notion for the framework and one of its parts with ten other principles?

> Recommendation 3: Coherence at all levels is an essential feature of any ethical framework, especially in terms of the overall structure of an initiative, its foundation, its implementation, and the relation between these different parts.

II-2. ETHICS WASHING AND ETHICS AVOIDANCE

There is also a possibility that some ethical frameworks contribute to various forms of ethics washing (Wagner 2018). Private or public organizations often use marketing or communication techniques to promote a positive perception of their practices regarding sensitive ethical issues. Sometimes, these perceptions are not accurate. For instance, an enterprise may promote an AI system as being fair or unbiased, but in reality, it may still contain hidden biases or perpetuate unfairness (Gambs et al. 2021, sec. 9).

The idea of ethics washing is a generalization over notions such as greenwashing (Laufer 2003) or fairness washing (McMurtry 2009). This is a problem because this leads to a false sense of security or trust in the AI system. But also, this can be used as a strategy to prevent regulation (Wagner and Delacroix 2019) or at least be counterproductive in developing regulation.

Voices were raised in recent years to warn the public of the harmful effects of ethics washing and the possibility that some ethics frameworks may be developed to promote the false impression of respecting ethical values in AI (Yeung, Howes, and Pogrebna 2020; Floridi 2019). Other work also reflects upon the limits of ethical frameworks and what can be achieved with principled approaches (Mittelstadt 2019).

> Recommendation 4: Private and public organizations that develop ethics frameworks must also develop expertise in AI ethics. They must ensure that ethics is not instrumentalized to achieve other commercial or political aims.
>
> Recommendation 5: Avoiding ethics is not a solution to the complexities of ethical analysis and action, or the risk of ethics washing.

The work on the limits of ethical frameworks is important and relevant. The conclusion, however, is not that ethical initiatives in AI are useless, or that they should be disregarded, quite the opposite. This shows the importance of the continued development of expertise in AI ethics, both inside and outside organizations.

There has been a tendency in big tech companies to cut on existing resources, at least internal human resources, in ethics. In April 2019, Google fired its Ethics Board less

than two weeks after launch after Google employees signed a petition calling for the removal of one member (Jee 2019). Google also fired, at the beginning of 2021, AI ethics researchers Timnit Gebru and Margaret Mitchell, exposing company divisions on academic freedom, diversity and AI ethics (Dave and Dastin 2021). Microsoft laid off its entire ethics and society team this year as part of layoffs that affected 10,000 employees across the company (Schiffer and Newton 2023). We may wonder if these decisions contribute to the continued development of expertise in ethics in these organizations.

> Recommendation 6: It is not always necessary to develop a new ethical framework when there are more general ethical frameworks, policies, and laws providing a rich normative context.

Finally, it is not always necessary to develop an ethical framework. There has been a particular phenomenon in AI ethics and many actors launched these initiatives, but there are other ways to deal with ethical issues. One of these ways is to look at the existing normative context, the existing frameworks, documents, policies, or theories that can already provide guidance. Interpreting this context is already a work in itself that can foster more ethical practices.

II-3. FUNDAMENTAL APPROACHES

We can identify at least six different approaches for developing an ethical framework. To begin with, an ethical framework can be based on a conception of the good or a conception of social justice. Theories of the good and justice aim to specify what is morally right such as Immanuel Kant's (1785) moral philosophy or John Rawl's (1999) theory of justice as equity. Theories of justice are more limited in the sense that they apply to social institutions, but not to people's choices in their personal lives (Rawls 1985; 1988). Following these approaches, one could adopt a welfarist or utilitarian conception (Kymlicka 2002, chap. 2) to AI ethics and examine how the technology should be developed to maximize social welfare.

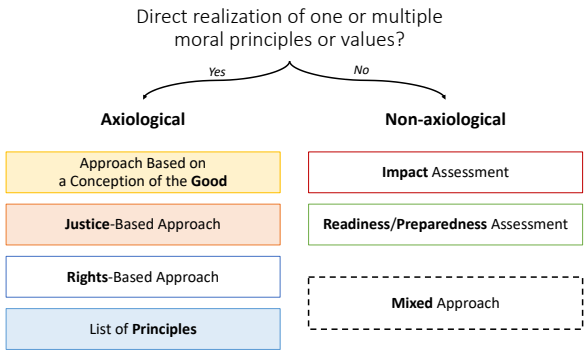

Figure 6: AI Ethics Framework Fundamental Approaches

These two approaches raise a challenge because it might be difficult to determine the implications of high-level moral theories. What is more, governments, let alone private organizations, rarely take a clear stance on these fundamental moral questions. Also, approaches based on a conception of the good may be perfectionist in the sense that they are not neutral among rival understandings of a good life. Many contemporary writers on politics would reject a moral conception for this reason (Wall 2021, sec. 3). However, the importance of these approaches should not be disregarded, especially the latter, because some questions of AI ethics will raise fundamental questions that cannot be answered without clarifying one's view on social justice.

> Recommendation 7: The importance of views about social justice should not be disregarded for any initiative in AI ethics. Although it may be overly challenging and impractical to derive an AI ethics framework from a conception of justice, some questions of AI ethics are likely to raise fundamental questions that cannot be answered without clarifying one's view in this regard.

Another approach would be to look at human or fundamental rights, or another charter of rights, as the foundation of an AI ethics framework. This approach is more common and it is easy to show examples of existing frameworks that claim to be rooted in fundamental rights, starting with the HLEG (2019, 11) guidelines. Many studies and reports also advocate for a rights-based approach (Fjeld et al. 2020; Yeung, Howes, and Pogrebna 2020; Access Now 2018). The more universal nature of some charters of rights and their stability is a strong aspect of the frameworks developed according to this approach. But a series of rights can also have conflicting outcomes and complex implications. Furthermore, charters of rights are sometimes understood as bottom or hard moral constraints — that should not be violated under any circumstances (Nickel 2021) — but not as the specification of the ultimate ideal that we should aim for with the development of the technology.

This may explain why the frameworks that claim to be derived from rights often end up proposing a series of principles without a clear demonstration of their derivation from these rights. It is unclear if these frameworks are part of a rights-based approach or a fourth different type of approach: the list of principles. This fourth approach is the most popular approach by a large margin. Most of the approaches mentioned so far feature series of principles or values that must be followed to ensure the ethical development and usage of AI.

The fourth type of approach has at least two strengths. First, list of principles make it easier to adapt a framework to current issues and views about social acceptability. Second, these frameworks may be easier to put into practice if the principles are defined in a way that is more practical and concrete. Finally, this type of framework is more popular. The last point is not a strength in itself, but the popularity of this approach may indicate other strengths that are not captured in the two previous points.

Lists of principles also raise important challenges. First, as mentioned above, there are significant divergences in the principles within existing ethical frameworks. Second, this approach may lead to frameworks that are not sufficiently grounded because the principles are treated as axioms. It is often asked that the people adopting the

framework accept the principles 'as is' without a normative argument to demonstrate the relevance of the principle or their derivation from more fundamental moral notions. Third, these lists are prone to incoherence and they are more likely to produce contradictions or conflicts, because different principles are likely to have different implications. This point was also raised above. Finally, there are particularities with current lists that are hard to explain. For instance, transparency is among the most popular principles (see also the discussion in the appendix) and there may be blind spots for other values. Few ethical frameworks include principles on the importance of economic development or growth, while it is likely to be an important consideration for most governmental and private sector organizations, as well as the public.

Recommendation 8: Ethical frameworks should build on universal foundations as much as possible. This is difficult to achieve with a list of axiomatic principles: statements that are considered established without a clear derivation from other more fundamental moral notions. Other approaches grounded in rights or a conception of justice may have a more universal dimension.

Recommendation 9: Frameworks with unitary, hierarchal, or minimalist foundations are less likely to produce conflicts or incoherence. Again, this issue is more likely to arise with frameworks founded on a list of axiomatic principles.

Recommendation 10: Notwithstanding the type of approach and the number of parts in the foundation of a framework, it is important to always specify how to make trade-offs or how to resolve the main conflicts that will arise with the application of the framework.

The four approaches presented so far imply the direct realization of one or multiple moral principles or values. In that sense, we may say that these approaches are axiological. They propose an axiology or a value theory that is "concerned with theoretical questions about value and goodness of all varieties" (Schroeder 2021). Fundamental approaches can also fall into a second category that is non-axiological. This category includes impact and readiness, or preparedness, assessments.

For instance, an impact assessment approach generally provides tools to determine the possible level of impact of a technological solution. The response to the technology should be proportional to the level of impact, or even abandoned altogether if the impact is too important. Two examples include the *Algorithmic Impact Assessment (AIA)* tool in Canada (https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html) and the United States (https://www.cio.gov/aia-eia-js/).

To be clear, non-axiological approaches are not morally neutral. What is considered an impactful algorithm, or an appropriate level of readiness, implies some evaluative notions about what is good or bad. But values are less directly involved in these approaches. This is both a strength and a weakness. The fact that a framework doesn't take a position as directly on what constitutes ethical AI makes it easier to adapt the

frameworks to various political agendas. But that may also lead to an insufficient determination of the permissible ways to use the technology. Furthermore, implicit values are, by definition, not explicit, and more work may be necessary to clarify these values to assess the relevance of a non-axiological framework.

Finally, the six different approaches are not mutually exclusive. A framework can build on multiple approaches, even if that might create additional risks of incoherence, see Figure 6 for an overview.

II-4. DEVELOPMENT CYCLE

A final consideration concerns the cycle of development of ethical frameworks. Most of the frameworks contained in the compilation have been developed and then publicized, with no indication of additional phases of development.

> Recommendation 11: Ethical framework initiatives should undergo a cycle of multiple iterations with successive development phases followed by the publication, adoption, and implantation of each version of the framework.

There are examples of influential ethical frameworks in the field of medical and research ethics that have undergone multiple iterations. First, the *Declaration of Helsinki* developed by the World Medical Association (WMA 2013) is a set of ethical principles regarding human experimentation. The declaration was originally adopted in June 1964 in Helsinki, Finland, and has since undergone seven revisions. In Canada, a new version of the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* was released in 2022 (https://ethics.gc.ca/eng/tcps2-eptc2_2022_introducing-presentation.html) after multiple revisions as well.

Within the field of AI ethics, we can point to two initiatives that have undergone, or plan to undergo, more than one iteration. First, the IEEE adopted two versions of its *Ethically Aligned Design* framework with intermediary requests for inputs. But even in that case, a more extensive calendar of development and revision could not be envisioned. In Canada, the authors of the *Directive on Automated Decision-Making* (https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592) explain that the technology is changing rapidly and the "directive will continue to evolve to ensure that it remains relevant." But apart from these two examples, few current initiatives seem to include plans for revisions.

Ethical frameworks need to stay relevant. They also need to capture moral truths. This is difficult to achieve with one iteration of a framework. For instance, the first versions of the *Declaration of Helsinki* raised concerning issues in terms of the definition of the patients that should be treated ethically (the declaration focussed on therapeutic research which allowed to experiment on an unfit subject with non-therapeutic research, see Doucet 1996, chap. 3) and these issues were addressed in subsequent versions. To use another example, *Just War Theory* is a series of principles or criteria that define what is morally justifiable before, during and after war (Orend 2008, sec. 2). The theory has enjoyed a long and distinguished pedigree in Western thought tradition, going back as far as Ancient Greece and Rome thinkers Augustine of Hippo, Cicero and Aristotle. There have been multiple revisions of the theory in the last two

thousand years until we were able to develop contemporary versions.

Moral knowledge tends to be gathered through long time periods with multiple back and forths. This reality should be acknowledged and integrated into our initiatives to make AI more ethical.

APPENDIX – ETHICAL ISSUES AND SOCIAL IMPACT OF AI

A fair amount of time and resources have been invested in the last decade to develop a deeper understanding of the ethical issues and impact of AI on society. Views regarding these issues have evolved at many levels.

Existential risk — Early concerns about AI tended to focus on the possibility that the technology could lead to catastrophic consequences or even the extinction of the human species. This could occur if AI were to become super-intelligent (Chalmers 2010) and beyond human control (Orseau and Armstrong 2016), or if AI was used deliberately for malicious purposes (Brundage et al. 2018). Raising awareness of these problems led prominent figures such as Elon Musk and Stephen Hawking to call for greater regulation and oversight of AI development and research (Economist 2016). Others have called for the development of friendly, trustable or human-compatible AI that is programmed to serve human interests and ensure its alignment with human values (Russell 2019). There has also been a growing movement of researchers and organizations dedicated to studying and addressing AI-related risks, these include the Future of Life Institute (https://futureoflife.org/), the Machine Intelligence Research Institute (https://intelligence.org/) and the Future of Humanity Institute at the University of Oxford (https://www.fhi.ox.ac.uk/).

It is often claimed that the impact of technology is overestimated in the short term and underestimated in the long term, an adage attributed to American scientist and futurist Roy Amara (Coates and Jarratt 1989). The response to the existential risk seems to conform to this tendency in multiple ways. First, and despite all the attention given to the subject, the development of AI was unlikely to produce catastrophic consequences in the short term. The year 2024 will mark the tenth anniversary of the publication of Nick Bostrom's (2014) book, *Superintelligence: Paths, Dangers, Strategies*, which is still one of the most important references on the topic. There is less debate and research on this issue today than there was 10 years ago because the attention shifted to more pressing issues (Crawford and Calo 2016).

That is not to say, however, that the existential risk no longer exists. The development of AI may still have an existential impact on the medium- to long-term. As a case in point, Geoffrey Hinton, one of the pioneers of deep learning, recently stepped down from his role as an AI researcher at Google, explaining that he wants to concentrate on the existential threat of AI (Douglas Heaven 2023).

Impact on work and technological unemployment — A report published in April 2023 by the Pew Research Center suggests that 62% of Americans believe AI will have a major impact on jobholders overall in the next 20 years (Rainie, Anderson, and Nolan 2023, 3). These perceptions are in line with trends observed for several years. In one of the center's first reports on artificial intelligence and robotics, in 2014, half of the surveyed experts believed that robots and digital agents would displace a "significant number of both blue- and white-collar workers" by 2025 (A. Smith and Anderson 2014, 5). These views were also reflected in scientific work on the economic impact of AI (Ford 2015; Miller 2017; Schlogl and Sumner 2018).

There is still a strong public perception that AI presents a risk for job replacement and income inequality, but expert views have shifted in recent years. Half of the experts surveyed in the Pew Research Center report of 2014 believed that AI would lead to a vast increase in income inequality, technological unemployment, and a general shrinkage of the labor market. Today's views are more nuanced. Fewer experts will be willing to claim there is a clear causal link between the development of AI and fewer jobs, let alone less economic growth (Aghion et al. 2019). A more common view is that AI can have both positive and negative effects, and we need more time, more data and more research to understand the economic impact of the technology (CEST 2021).

Transparency — The issue of transparency gained increasing attention as people became aware of the potential risks and negative consequences associated with opaque AI systems (Pasquale 2015; Campolo et al. 2017). The lack of transparency can make it difficult to understand how AI systems work and make decisions, which can lead to negative consequences such as biased or discriminatory outcomes.

While transparency is an important consideration in the development and deployment of any system, the question is increasingly being asked whether we should aim for this ideal directly (Ananny and Crawford 2018). At the very least, there is a tendency to focus more directly on underlying issues of accountability, explainability and discrimination that come with a lack of transparency.

Accountability — There is accountability when: *i*) an individual or a group of individuals; *ii*) provide an account (a justification or an explanation); *iii*) about a political decision, a policy, the functioning of a new product or service, and so on; i*v*) to another individual or group. The entity that receives the explanation must have some sanctioning power in the sense that it must be able to impose a punishment or corrective actions if the account is unsatisfactory (Binns 2018; Bovens 2010; Mulgan 2000). The issue of accountability in AI became more important as AI systems became more pervasive in society and take on important decision-making tasks. For instance, an AI system can make a bad decision or a recommendation that will harm someone, but it may be difficult to attribute this outcome to a particular individual or organization. The notion of accountability helps determine what is wrong in this scenario.

Accountability in AI is still an important topic of debate and research for multiple reasons: first, we may wonder about the type of ethical standards or frameworks that could prevent accountability gaps, the extent to which various actors have a responsibility to provide an account for their actions and the type of sanctions that can be imposed (Manyika et al. 2023). The idea is also intertwined with other notions that capture desirable aspects of our social arrangements: transparency, responsibility, answerability, attributability, and the proper auditing and sanctioning of algorithmic decision-makers (Shoemaker 2011; A. M. Smith 2012; Eshleman 2016).

Explainability — AI may raise an issue because the people that develop or use the technology are not sufficiently accountable. This lack of accountability can be attributed to a lack of proper mechanisms and standards, but there are also features of AI technologies that make it difficult to provide an account of the decision made by an AI algorithm. On the one hand, these algorithms may function as black boxes when the organizations that develop or use these systems provide limited information on their

inner functioning (Burrell 2016). On the other hand, some algorithms are very complex to the point where the best human experts cannot fully understand how they function. This is a particular challenge in ML where powerful models contain large series of parameters. Humans are not able to interpret the rules embedded in these parameters with symbols we can understand. In that case, the lack of explainability does not come from difficulty accessing information about an algorithm, but an inherent feature of the system and the way it is developed (Knight 2017a).

Explainability is important because inaccurate or biased models can lead to serious consequences, such as discrimination, unfairness, or even harm. But also because there are many instances where an organization ought to be able to justify the decisions made by its AI systems (Rudin 2019). Research and debates in the field of explainability are very lively today, as we still struggle to fully explain how many algorithms operate, especially ML algorithms (Molnar 2019; Biecek, Kozak, and Zawada 2022).

Discrimination — this issue has been a growing concern for several years as we developed a better understanding of the biases produced by AI technologies (Knight 2017b). A bias refers to the idea of a systematic error made by a system. Discrimination is the evaluative concept and implies there is an unjust distinction between people based on the group to which they belong (Altman 2020). Not all biases are morally wrong, for instance an algorithm used in employment may be favorable to the member of an ethnic group and this may compensate for other injustices. If an algorithm commits a systematic error that leads to treat the members of a group in a way that is considered unjust (based on their membership in this group), then it might be discriminatory.

Biases and discrimination can occur when AI systems are trained on data that reflects historical biases, leading to unequal outcomes for different groups and perpetuating existing biases and discrimination (Buolamwini and Gebru 2018). Technology can also amplify existing biases, often in ways that are not immediately apparent. For example, a candidate selection system trained on resumes from predominantly male job applicants may inadvertently learn to favor male candidates over equally qualified female candidates. Work from Julia Angwin and other collaborators at ProPublica (Angwin et al. 2016), Cathy O'Neil (2016) and Kate Crawford (2016) have contributed to bringing these issues to the fore.

There is widespread agreement today that addressing biases and discrimination in AI is a complex issue that requires ongoing attention and action from all actors at all levels: researchers, industry practitioners, policymakers, and civil society groups (West, Whittaker, and Crawford 2019; Park 2023). One key strategy involves developing AI systems that are designed to be transparent and explainable, allowing people to understand how decisions are being made and identify any potential biases, but these are also current challenges we face with the technology (Rudin 2019).

Fake news and disinformation — awareness about these issues crystallized with the apparition of the first generative system and the Facebook-Cambridge Analytica data scandal (Lapowsky 2018). The development of information technology, social media and now more advanced AI technology allows for the creation and dissemination of targeted and false information. The publication of the first deepfake videos in 2017

suddenly raised awareness about the potential of AI to create fake contents (Hao 2017). The expression deepfake is a portmanteau for *fake* media content generated with *deep learning* networks. While these fake contents were initially long and complicated to make, with mitigated results, the new wave of generative models makes it very easy to produce fake text, images and even videos (to a lesser extent) that are difficult to distinguish from reality.

REFERENCES

Abrassart, Christophe, Yoshua Bengio, Nathalie de Marcellis-Warin, Marc-Antoine Dilhac, Sébastien Gambs, Vincent Gautrais, Martin Gibert, et al. 2018. "Montreal Declaration For A Responsible Development Of Artificial Intelligence." Montréal. https://www.declarationmontreal-iaresponsable.com/.

Access Now. 2018. "Human Rights in the Age of Artificial Intelligence." Access Now. https://www.accessnow.org/wp-content/uploads/2018/11/AI-and-Human-Rights.pdf.

Aghion, Philippe, Céline Antonin, Simon Bunel, Diane Coyle, Zia Qureshi, Mary O'Mahony, Michael J. Böhm, et al. 2019. *Work in the Age of Data*. BBVA. https://www.bbvaopenmind.com/en/books/work-in-the-age-of-data/.

Altman, Andrew. 2020. "Discrimination." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/discrimination/.

Ananny, Mike, and Kate Crawford. 2018. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20 (3): 973–89. https://doi.org/10.1177/1461444816676645.

Angwin, Julia, Jeff Larson, Surya Mattu, Lauren Kirchner, and ProPublica. 2016. "Machine Bias." *ProPublica*, May. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Biecek, Przemysław, Anna Kozak, and Aleksander Zawada. 2022. *The Hitchhiker's Guide to Responsible Machine Learning: Interpretable and eXplainable Artificial Intelligence with examples in R*. Warschau: Scientific Foundation SmarterPoland.pl.

Binns, Reuben. 2018. "Algorithmic Accountability and Public Reason." *Philosophy & Technology* 31 (4): 543–56. https://doi.org/10.1007/s13347-017-0263-5.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bovens, Mark. 2010. "Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism." *West European Politics* 33 (5): 946–67. https://doi.org/10.1080/01402382.2010.486119.

Bringsjord, Selmer, and Naveen Sundar Govindarajulu. 2018. "Artificial Intelligence." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. 2018. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." Future of Humanity Institute, University of Oxford. https://maliciousaireport.com/.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. New York, NY, USA: PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 2053951715622512. https://doi.org/10.1177/2053951715622512.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. "AI Now 2017 Report." New York: AI Now.

Canadian Press. 2023. "Google Is Blocking Some Canadians from Seeing Online News | CBC News." *CBC*, February 23, 2023. https://www.cbc.ca/news/business/google-blocking-news-1.6757500.

CEST. 2021. "Les effets de l'intelligence artificielle sur le monde du travail et la justice sociale : automatisation, précarité et inégalités." Québec, Québec: Commission de l'éthique en science et en technologie. https://www.ethique.gouv.qc.ca/media/viipye0b/ia_travail_web.pdf.

Chalmers, David. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65.

Choudhry, Mavra, Nic Wall, and Molly Reynolds. 2023. "Guide to Artificial Intelligence Regulation in Canada." Torys LLP.

Coates, Joseph F., and Jennifer Jarratt. 1989. *What Futurists Believe*. Bethesda, Maryland: Lomond Pubns.

Coeckelbergh, Mark. 2020. *AI Ethics*. CogNet. MIT Press Direct. https://doi.org/10.7551/mitpress/12549.001.0001.

Crawford, Kate. 2016. "Artificial Intelligence's White Guy Problem." *The New York Times*, June 25, 2016. http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

Crawford, Kate, and Ryan Calo. 2016. "There Is a Blind Spot in AI Research." *Nature* 538 (7625): 311–13. https://doi.org/10.1038/538311a.

Dave, Paresh, and Jeffrey Dastin. 2021. "Google Fires Second AI Ethics Leader as Dispute over Research, Diversity Grows." *Reuters*, February 19, 2021, sec. U.S.

Legal News. https://www.reuters.com/article/us-alphabet-google-research-idUSKBN2AJ2JA.

Doucet, Hubert. 1996. *Au pays de la bioéthique : l'éthique biomédicale aux États-Unis*. Genève: Labor et Fides.

Douglas Heaven, Will. 2023. "Geoffrey Hinton Tells Us Why He's Now Scared of the Tech He Helped Build." *MIT Technology Review*, May 2, 2023. https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/.

Economist. 2016. "Frankenstein's Paperclips." *The Economist*, June 23, 2016. https://www.economist.com/special-report/2016/06/23/frankensteins-paperclips.

Eshleman, Andrew. 2016. "Moral Responsibility." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." 2020–1. Berkman Klein Center. https://doi.org/10.2139/ssrn.3518482.

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.

Ford, Martin. 2015. *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York: Basic Books.

Franzke, Aline Shakti. 2022. "An Exploratory Qualitative Analysis of AI Ethics Guidelines." *Journal of Information, Communication and Ethics in Society* 20 (4): 401–23. https://doi.org/10.1108/JICES-12-2020-0125.

Future of Life Institute. 2017. "Asilomar AI Principles." Future of Life Institute. January 2017. https://futureoflife.org/ai-principles/.

Gambs, Sébastien, Ulrich Aïvodji, Céline Castets-Renard, Ignacio Cofone, Aude-Marie Marcoux, and Dominic Martin. 2021. "Privacy and AI Ethics: Understanding the Convergences and Tensions for the Responsible Development of Machine Learning." Office of the Privacy Commissioner of Canada's Contributions Program. https://www.priv.gc.ca/en/opc-actions-and-decisions/research/funding-for-privacy-research-and-knowledge-translation/completed-contributions-program-projects/2020-2021/p_2020-21_09/.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge: The MIT Press.

Goverment of Germany. 2020. "Artificial Intelligence Strategy of the German Federal Government, 2020 Update." Federal Government of Germany. https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf.

Government of Canada. 2023. "The Artificial Intelligence and Data Act (AIDA) – Companion Document." Innovation, Science and Economic Development Canada. March 13, 2023. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.

Guszcza, James, Michelle A. Lee, Beena Ammanath, and Dave Kuder. 2020. "Human Values in the Loop: Design Principles for Ethical AI." Deloitte. https://www2.deloitte.com/content/dam/Deloitte/xe/Documents/About-Deloitte/WGS%20report%20l%20AI%20Ethics.pdf.

Hagendorff, Thilo. 2020. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines* 30 (1): 99–120. https://doi.org/10.1007/s11023-020-09517-8.

Hao, Karen. 2017. "Researchers Have Figured out How to Fake News Video with AI." *Quartz*, July 19, 2017. https://qz.com/1031624/researchers-have-figured-out-how-to-fake-news-video-with-ai.

Hashmi, Ali. 2019. "AI Ethics: The Next Big Thing in Government." Deloitte. https://www2.deloitte.com/content/dam/Deloitte/xe/Documents/About-Deloitte/WGS%20report%20l%20AI%20Ethics.pdf.

Heath, Ryan. 2023. "China Races Ahead of U.S. on AI Regulation." *Axios*, May 8, 2023. https://www.axios.com/2023/05/08/china-ai-regulation-race.

Heikkilä, Melissa. 2023. "Generative AI Risks Concentrating Big Tech's Power. Here's How to Stop It." MIT Technology Review. April 17, 2023. https://mailchi.mp/technologyreview.com/generative-ai-concentration-of-big-tech-power?e=db7b51eb1b.

HLEG AI. 2019. "Ethics Guidelines for Trustworthy AI." Brussels: High-Level Expert Group on Artificial Intelligence, European Commission.

IEEE. 2017. "Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems, Version 2." Institute of Electrical and Electronics Engineers (IEEE). http://standards. ieee.org/develop/indconn/ec/autonomous_ systems.html.

Jee, Charlotte. 2019. "Google Has Now Cancelled Its AI Ethics Board after a Backlash from Staff." *MIT Technology Review*, April 5, 2019. https://www.technologyreview.com/2019/04/05/136188/google-has-now-cancelled-its-ai-ethics-board-after-a-backlash-from-staff/.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science* 349 (6245): 255–60. https://doi.org/10.1126/science.aaa8415.

Kak, Amba, and Sarah Myers West. 2023. "AI Now 2023 Landscape: Confronting Tech Power." AI Now Institute. https://ainowinstitute.org/2023-landscape?mc_cid=5b274fd045.

Kang, Cecilia. 2023. "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing." *The New York Times*, May 16, 2023, sec. Technology. https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

Kant, Immanuel. 1785. *Groundwork of the Metaphysics of Morals*. Edited by Mary Gregor. Cambridge Texts in the History of Philosophy. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809590.

Klein, Ezra. 2023. "Opinion | The Surprising Thing A.I. Engineers Will Tell You If You Let Them." *The New York Times*, April 16, 2023, sec. Opinion. https://www.nytimes.com/2023/04/16/opinion/this-is-too-important-to-leave-to-microsoft-google-and-facebook.html.

Knight, Will. 2017a. "The Dark Secret at the Heart of AI." *MIT Technology Review*, April 11, 2017. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

———. 2017b. "Forget Killer Robots—Bias Is the Real AI Danger." MIT Technology Review. October 3, 2017. https://www.technologyreview.com/2017/10/03/241956/forget-killer-robotsbias-is-the-real-ai-danger/.

Kymlicka, Will. 2002. *Contemporary Political Philosophy: An Introduction*. Second ed. Oxford: Oxford University Press.

Lapowsky, Issie. 2018. "Mark Zuckerberg Answers to Congress For Facebook's Troubles." *Wired*, April 10, 2018. https://www.wired.com/story/mark-zuckerberg-congress-facebook-troubles/.

Laufer, William S. 2003. "Social Accountability and Corporate Greenwashing." *Journal of Business Ethics* 43 (3): 253–61. https://doi.org/10.1023/A:1022962719299.

Manyika, James, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. 2023. "The AI Index 2023 Annual Report." Stanford, CA: Institute for Human-Centered AI Institute (HAI), Stanford University. https://aiindex.stanford.edu/ai-index-report-2021/.

Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon.

McMurtry, J. J. 2009. "Ethical Value-Added: Fair Trade and the Case of Café Femenino." *Journal of Business Ethics* 86 (1): 27–49. https://doi.org/10.1007/s10551-008-9760-x.

Miller, Claire Cain. 2017. "Evidence That Robots Are Winning the Race for American Jobs." *The New York Times*, March 28, 2017, sec. The Upshot. https://www.nytimes.com/2017/03/28/upshot/evidence-that-robots-are-winning-the-race-for-american-jobs.html.

Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill Series in Computer Science. New York: McGraw-Hill. http://catdir.loc.gov/catdir/toc/mh022/97007692.html.

Mittelstadt, Brent. 2019. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1 (11): 501–7. https://doi.org/10.1038/s42256-019-0114-4.

Molnar, Christoph. 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Morisville, North Carolina: Lulu.

Mulgan, Richard. 2000. "'Accountability': An Ever-Expanding Concept?" *Public Administration* 78 (3): 555–73. https://doi.org/10.1111/1467-9299.00218.

Nickel, James. 2021. "Human Rights." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/rights-human/.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Orend, Brian. 2008. "War." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2008. http://plato.stanford.edu/archives/fall2008/entries/war/.

Orseau, Laurent, and Stuart Armstrong. 2016. "Safely Interruptible Agents." In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 557–66. UAI'16. Arlington, Virginia, USA: AUAI Press.

Park, Yong Jin. 2023. "How We Can Create the Global Agreement on Generative AI Bias: Lessons from Climate Justice." *AI & SOCIETY*, April. https://doi.org/10.1007/s00146-023-01679-0.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Reprint edition. Cambridge, Massachusetts: Harvard University Press.

Rainie, Lee, Monica Anderson, and Haley Nolan. 2023. "AI in Hiring and Evaluating Workers: What Americans Think." Pew Research Institute. https://www.pewresearch.org/internet/2023/04/20/ai-in-hiring-and-evaluating-workers-what-americans-think/.

Rawls, John. 1985. "Justice as Fairness: Political Not Metaphysical." *Philosophy & Public Affairs* 14: 223–51.

———. 1988. "The Priority of Right and Ideas of the Good." *Philosophy & Public Affairs* 17 (4): 251–76.

———. 1999. *A Theory of Justice*. Rev. ed. Cambridge: Belknap Press of Harvard Univeristy Press.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Schiffer, Zoë, and Casey Newton. 2023. "Microsoft Just Laid off One of Its Responsible AI Teams." *Platformer* (blog). March 13, 2023. https://www.platformer.news/p/microsoft-just-laid-off-one-of-its.

Schlogl, Lukas, and Andy Sumner. 2018. "The Rise of the Robot Reserve Army: Automation and the Future of Economic Development, Work, and Wages in Developing Countries." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3208816.

Schroeder, Mark. 2021. "Value Theory." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/value-theory/.

Schwartz, Mark S. 2002. "A Code of Ethics for Corporate Code of Ethics." *Journal of Business Ethics* 41 (1): 27–43. https://doi.org/10.1023/A:1021393904930.

Shoemaker, David. 2011. "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121 (3): 602–32.

Smith, Aaron, and Jana Anderson. 2014. "AI, Robotics, and the Future of Jobs." Pew Research Institute. http://www.pewinternet.org/2014/08/06/future-of-jobs/.

Smith, Angela M. 2012. "Attributability, Answerability, and Accountability: In Defense of a Unified Account." *Ethics* 122 (3): 575–89. https://doi.org/10.1086/664752.

Stahl, Bernd Carsten. 2022. "From Computer Ethics and the Ethics of AI towards an Ethics of Digital Ecosystems." *AI and Ethics* 2 (1): 65–77. https://doi.org/10.1007/s43681-021-00080-1.

Tidjon, Lionel Nganyewou, and Foutse Khomh. 2022. "The Different Faces of AI Ethics Across the World: A Principle-To-Practice Gap Analysis." *IEEE Transactions on Artificial Intelligence*, 1–20. https://doi.org/10.1109/TAI.2022.3225132.

Wagner, Ben. 2018. "Ethics As An Escape From Regulation. From 'Ethics-Washing' To Ethics-Shopping?" In *Being Profiled: Cogitas Ergo Sum*, edited by Emre Bayamiloglu, Irina Baraliuc, Liisa Janssens, and Mireille Hildebrandt. Amsterdam University Press. http://oapen.org/search?identifier=1004973;keyword=9789463722124.

Wagner, Ben, and Sylvie Delacroix. 2019. "Constructing a Mutually Supportive Interface between Ethics and Regulation." SSRN Scholarly Paper ID 3404179. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3404179.

Wall, Steven. 2021. "Perfectionism in Moral and Political Philosophy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/perfectionism-moral/.

West, Sarah Myers, Meredith Whittaker, and Kate Crawford. 2019. "Discriminating Systems: Gender, Race and Power in AI." New York: AI Now Institute.

WMA. 2013. "World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects." *JAMA* 310 (20): 2191–94. https://doi.org/10.1001/jama.2013.281053.

Yeung, Karen, Andrew Howes, and Ganna Pogrebna. 2020. "AI Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 0. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.5.