Chapter 7
Global AI Ethics Documents: What They Reveal About Motivations, Practices, and Policies

Daniel S. Schiff, Kelly Laas, Justin B. Biddle, and Jason Borenstein

Abstract In recent years, numerous organizations worldwide have produced normative documents identifying potential benefits, harms, and associated recommendations related to artificial intelligence (AI). This chapter examines why these AI ethics documents are being produced and what they can tell us about the motivations, practices, and policies that surround AI. While much of the literature to-date discusses whether consensus on ethical principles is emerging, critical unanswered questions remain around representation and power, the translation of principles to practices, and the complex set of reasons that underlie the creation of these documents. Our work brings attention to these underexplored issues through a comprehensive literature review, and by proposing a novel typology of motivations that helps to characterize the creation of AI ethics documents. Finally, drawing on the recent case of gene-editing ethics documents, we argue that AI ethics stakeholders can achieve more beneficial impacts for society by fostering more diverse and inclusive participatory processes.

Keywords Artificial intelligence · Ethics codes · Technology governance · Organizational motivations

D. S. Schiff (✉) · J. B. Biddle · J. Borenstein
School of Public Policy, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: schiff@gatech.edu; justin.biddle@pubpolicy.gatech.edu; borenstein@gatech.edu

K. Laas
Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, IL, USA
e-mail: laas@iit.edu

K. Laas et al. (eds.), *Codes of Ethics and Ethical Guidelines*, The International Library of Ethics, Law and Technology 23,
https://doi.org/10.1007/978-3-030-86201-5_7

7.1 Introduction

Since 2016, numerous organizations worldwide have produced documents identifying the potential benefits and harms of artificial intelligence (AI). Documents that focus on the ethical aspects of AI have taken many forms, including codes of ethics, normative guidelines, and policy strategies. Such documents differ from traditional scholarly publications in that they often represent official viewpoints of the authoring organizations. This development is largely in response to the profound impacts that AI technologies are expected to have on human life. As such, the AI ethics documents typically reflect on AI's benefits and potential harms, offer ethical principles to minimize risks, and in some cases, include recommendations that could be realized through internal change or external influence. These normative documents provide us with an opportunity to understand how influential and, in some cases, politically powerful entities and global thought leaders imagine AI's impacts and how they intend to shape them. For this reason, AI ethics documents are valuable sources of information and important objects of study.

In this chapter, we seek to examine why AI ethics documents are being produced and what they suggest about the motivations, practices, and policies that surround AI. While much of the current literature discusses whether consensus on ethical principles is emerging, critical unanswered questions remain around representation and power, the translation of principles to practices, and the complex set of reasons that underlie the creation of these documents. Our work seeks to contribute by bringing attention to these underexplored issues through a comprehensive literature review, and by proposing a novel typology of motivations that helps to characterize the creation of AI ethics documents. Our examination suggests that AI ethics documents are likely to play an important – if complex – role in shaping future practices, norms, and regulations surrounding AI.

After reviewing the recent history surrounding AI ethics documents in Section 2, we summarize the recent literature on AI ethics documents in Section 3 and briefly describe our own study in Section 4. Section 5 produces a typology to examine the multiple motivations of organizations that are producing AI documents. In light of these motivations, Section 6 considers characteristics that are likely to make AI documents more effective at reaching the goals that they are trying to achieve. In Section 7, we examine regulatory and other responses to the ethical and social risks surrounding gene editing as a way of providing insight into the possible future of AI ethics documents. Section 8 concludes the discussion.

7.2 The New AI Spring and the Codification of AI Ethics

Since the 1950s, there have been several waves of interest in AI; the 2010s have marked the beginning of a new 'AI Spring' – a period of increased funding, research, development, and public attention. Venture capital, publications, patents,

conference attendance, and employment in this field have grown substantially. Estimates indicate that AI's economic impact will be in the trillions (PricewaterhouseCoopers 2017), and many multinational 'Big Tech' companies have reorganized their operations away from 'mobile-first' and 'cloud-first' to 'AI-first.' Some authors even claim that the AI age will be heralded as the most important economic and social transformation in recent human history, a general-purpose technology with sweeping impacts across human society and the key to the "Fourth Industrial Revolution" (Villani et al. 2018; Schwab 2016).

The tremendous excitement for AI is largely the result of technical advances in computer science, specifically natural language understanding and generation, image recognition, and search optimization, among other domains. These advances are themselves the result of two key developments: 1) increased processing power that made feasible the application of algorithms for deep neural networks; and 2) massive increases in the availability of 'big data,' including from online shopping, search, and social media sources (Duan et al. 2019). The movement towards the digitization of information, including health records, has also been an important driving factor (Mai 2016).

At the same time, a suite of ethical, legal, policy, and social concerns have emerged in relation to AI, which are increasingly drawing the attention of scholars, practitioners, policymakers, and the public. Debates are reawakening, for example, about the role of automation in replacing human labor and which work sectors are most vulnerable to displacement (Frey and Osborne 2017). While supporters of AI admit that some jobs will be lost to 'creative destruction', they contend that net-positive benefits will emerge for job creation and economic growth (McKinsey Global Institute 2018). Even if their assertions are correct – and they might not be – questions must still be addressed about the distribution of these impacts across different populations (West 2018).

In addition to concerns about job displacement, the capacity of facial recognition and big data more generally to enable widespread surveillance, micro-targeting, and digital manipulation exacerbates traditional concerns about privacy and autonomy and raises new facets of these concerns (Bennett and Raab 2017). The capacity of algorithms to reflect and reproduce societal biases, such as when deciding who should be eligible for a bank loan or job opportunity, and to do so without sufficient transparency or public scrutiny, brings to light the risks of placing increasingly weighty decisions into the hands of machines. Concern is growing stronger that some decisions with profound legal implications might be handed over to AI (Scherer 2016). In many sectors of human life ranging from AI in healthcare (Char et al. 2018) to autonomous vehicles (Bagloee et al. 2016), the risks – some of which may not be entirely foreseen at this point – must be juxtaposed against the transformative potential of this powerful set of technologies.

Many scholars have sought to identify and parse ethical issues pertaining to AI. Acronyms such as FEAT, FAT, or FATE, referring to some combination of fairness, ethics, accountability, and transparency, have become mainstream in academia, including through ACM's Fairness, Accountability, and Transparency in Machine Learning (FAccT)* or AAAI's AI, Ethics, and Society conferences.

Normative AI documents represent one category of attempts by key organizational actors – governments, corporations, NGOs, and others – to grapple with this balancing act. Organizations have pursued multiple avenues outside of academic scholarship and the creation of ethics documents to reflect on AI's benefits and risks. They often make recommendations for change through regulatory strategies, best practices, or other means. Some companies and universities have created new in-person and online courses on AI ethics (Gartenberg 2018; Vincent 2019). Organizational and technical frameworks for engaging in responsible AI development have also begun to appear (Chatila and Havens 2019; The Institute for Ethical AI & ML 2020). Governmental task forces and cooperatives have emerged to consider evidence and possible regulatory action (Automated Decision Systems and Task Force 2019; OECD 2019) along with specialized business units and governance boards (Fast Company 2017; Todd 2019). In addition, new organizations and collaborations have recently been launched with the mission of addressing AI ethics, such as AI Now and the Wadhwani Institute. Finally, many organizations have crafted various ethics and policy documents, which is the focus of this chapter.

7.3 A Review of Research Studies on AI Ethics Documents

Scholars had begun to examine normative AI documents through a variety of approaches, including qualitative content and thematic analysis (Floridi and Cowls 2019) and quantitative text analysis (Zeng et al. 2018). Table 7.1 summarizes some of the existing meta-analytic research on normative AI documents:

The studies in Table 7.1 pay the most attention to the following topics or themes: 1) to what degree there is consensus on which ethics topics are the most important to examine; 2) how existing documents reflect issues of representation and power, and 3) what the goals of the documents are and whether they are effective at generating change. We examine each item below, noting the state of recent scholarly discussions, how our work potentially contributes, and what current knowledge gaps are.

7.3.1 Consensus on Ethical Topics

Nearly all of the aforementioned analyses engage in an inductive search for ethics categories. They ultimately identify a small number of core principles or themes (typically between five and ten), such as accountability, transparency, beneficence, justice, and explainability. The ethical principles are parsed and categorized in a variety of ways to simplify and compare across documents. For example, Floridi and Cowls (2019) map 47 AI ethical values onto five core principles (beneficence, non-maleficence, autonomy, justice, and explicability), while Zeng et al. (2018) articulate ten themes.

Table 7.1 Review of meta-analytical research on normative AI documents

Study	# of documents	Method of analysis	Key findings
Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': The US, EU, and UK approach.	3	Comparative analysis of 3 governmental strategies (US, UK, and European Parliament)	Finds that transparency, accountability, and positive impact are shared values of a 'good AI society,' as well as cooperation across sectors. There remains divergence in nature of shared responsibility, specific ethical values, and how a broad vision is to be implemented in a certain kind of society.
Daly, A., Hagendorff, T., Li, H., Mann, M., Marda, V., Wagner, B., Wang, W., & Witteborn, S. (2019). Artificial Intelligence, Governance, and Ethics: Global Perspectives.	16	Comparative overview of ethics initiatives (primarily documents) from the perspective of countries, with several intergovernmental bodies, NGOs, and corporations discussed	Finds transparency, accountability, and privacy as common principles, and identifies missing issues like hidden human and energy costs. Discusses the importance of competition and collaboration, public and international engagement, and considers challenges in implementing and enforcing the principles.
Dutton, T., Barron, B., & Boskovic, G. (2018). Building an AI World: Report on National and Regional AI Strategies. CIFAR.	18	Reviews 18 national AI strategies	Maps documents' discussion of research, AI talent, future work, industrial strategy, ethics, data, AI in government, and inclusion. Also reviews current status of each national strategy and funding as of 2018.

(continued)

Table 7.1 (continued)

Study	# of documents	Method of analysis	Key findings
Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (SSRN Scholarly Paper ID 3518482). Berkman Klein Center for Internet & Society.	36	Coding of 36 high-profile sets of principles and inductive thematic analysis and frequency mapping of eight themes and components of each	Identifies eight themes: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. Notes whether documents have reference to human rights, argues that there is a general and growing consensus, and suggests that principles are unlikely to be effective without a larger governance ecosystem.
Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society.	6	Comparative analysis of high-profile sets of principles from NGOs and government entities	Identifies 47 ethics principles and maps these to four core principles from bioethics (beneficence, non-maleficence, autonomy, injustice) plus an additional principle – explicability. Notes a lack of geographic, cultural, and social diversity in these documents.
Gibert, M., Mondin, C., & Chicoisne, G. (2018). Montréal Declaration of Responsible AI: 2018 Overview of International Recommendations for AI Ethics (pp. 78–97). University of Montréal.	7	Comparative analysis of ethical concepts and recommendations in seven reports, typology based on citizen recommendations and Montréal Declaration categories	Discusses seven ethical concepts: well-being, autonomy, justice, privacy, knowledge, democracy, and responsibility. Notes a divergence between public sector and private sector documents in terms of where solutions should be applied. Identifies an overall convergence in ethical concepts.

(continued)

Table 7.1 (continued)

Study	# of documents	Method of analysis	Key findings
Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning.	7	Frame analysis to identify second-order ethical grounding assumptions	Identified higher-order themes, including a focus on expert oversight, deterministic assumptions of AI progress, assignment of ethical responsibility to designers instead of others, and the use of public engagement to confer legitimacy.
Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines.	21	Qualitative mapping of topics and frequencies and comparative analysis	Identifies accountability, privacy, fairness as most prominent values, and existential threats, machine consciousness, and social cohesion as omissions. Reviews issues like business vs. ethics interests, US vs. China competition, lack of effectiveness of ethics codes, and recommends virtue ethics over deontology.
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines.	84	Qualitative content analysis and code mapping	Identifies transparency, justice and fairness, non-maleficence, responsibility, and privacy as converging topics, and sustainability and solidarity as underrepresented. Reviews issues like precedence of non-maleficence over beneficence, differences in how principles are interpreted, and lack of clarity around implementation.

(continued)

Table 7.1 (continued)

Study	# of documents	Method of analysis	Key findings
Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI.	NA	Reflects on other meta-analyses, noting that at least 84 ethics documents exist. Contrasts AI ethics with medical ethics	Argues that "AI development lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms." (p.1) Offers recommendations including bottom-up ethics, licensure, shifting to business ethics, and approaching ethics as a process, not solution.
Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices.	NA	Creates a typology, mapping publicly available AI ethics design tools/ methodologies to five core ethical principles (Floridi and Cowls 2019)	Identifies existing tools/ methodologies to apply AI ethics principles in AI development, design, training, deployment, etc. Highlights the emphasis in the available tools/methods on explicability, and inability to assess individual impacts, and the lack of guidance towards using the tools/methods.
Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking Artificial Intelligence Principles.	27	Quantitative content analysis to identify topic and keyword frequencies	Defines keywords for 10 manually-chosen themes: Humanity, Collaboration, Share, Fairness, Transparency, Privacy, Security, Safety, Accountability, AGI/ASI. Notes some sector differences, like a corporate focus on collaboration rather than security and privacy. Recommends a focus on safety, AGI, and societal transformation.

The studies investigate whether a global consensus is emerging on which AI ethics topics are (or at least appear) most worthy of attention. Most of the studies conclude that there is, though the specific framing and parsing of ethics concepts differ by document. This can lead to a messy proliferation of somewhat interrelated ideas. As a result, Morley et al. (2019) describe the consensus on ethics topics as "fragile."

Nevertheless, the degree to which similarities are appearing across sectors and regions of the world is noteworthy. While this may be a positive step towards global harmonization and governance, a legitimate concern persists that underrepresented groups, such as people in poorer regions of the world, might not have a true voice in this process (Fejerskov 2017).

The studies also highlight neglected topics and differences across sectors or regions. For example, Hagendorff (2020) argues that the erosion of social cohesion and potential existential threats related to AI are underemphasized in AI ethics documents. Daly et al. (2019) note the lack of attention to the hidden ethical costs of AI, including energy usage and human labor. Zeng et al. (2018) comment on differences in focus between the public and private sectors. Gibert et al. (2018) argue that different stakeholders are responsible for addressing particular AI ethics challenges.

However, the discussion of neglected topics in the AI ethics realm is relatively underdeveloped, as most attention is paid to identifying shared themes and consensus. More research should be done, for example, on whether the lack of attention to particular topics in normative AI documents leads to a failure to address those topics (e.g., by failing to seek regulatory solutions or to develop technical or governance strategies). Additional study should also examine whether differences across sectors or regions reflect power dynamics, as organizations attempt to frame problems and solutions in light of their own interests. Our research, discussed in this chapter and elsewhere (Schiff et al. 2021), is helping to fill this gap by focusing on topical omissions as well as variances across three organizational sectors (public, private, and NGO).

7.3.2 Representation and Power

Some studies address issues of representation by examining who writes or participates in the development of AI ethics documents. For example, Floridi and Cowls (2019) note a lack of geographic, cultural, and social diversity in the documents. Daly et al. (2019) question whether academic experts and civil society groups are sufficiently represented. Hagendorff (2020) notes the relatively low number of women amongst document authors and argues that AI ethics principles appear to be molded mostly by men.

Greene et al. (2019) challenge prevailing assumptions that they argue underlie discourse surrounding AI, such as the sense that AI innovation will proceed deterministically, with little room for humans to shape AI's trajectory. They suggest that the focus of private sector documents on expert oversight and their assignment of ethical responsibility to designers serves as a means to avoid scrutiny of higher-level business decision-makers and economic systems. Furthermore, they suggest that public engagement in creating AI ethics principles is aimed at sanitizing or rubber-stamping conceptual frames and strategies that favor experts, elites, and the systems they prefer.

Based on the state of the current literature, we offer several observations. First, the fact that nearly all public and private sector AI ethics documents and national policy strategies are coming from high-income and powerful countries and multinational corporations is a serious concern (Kak 2020; Schiff et al. 2020). Admittedly, since much of the scholarly literature focuses on English-language documents, some items may have been missed. Yet even if that is the case, low-income countries are still underrepresented. Ethics and policy documents that adopt a national or regional perspective (e.g., the European Union) run the risk of articulating ethical issues through a narrow lens. For example, a country that directs its attention toward alleviating inequality domestically may pursue investment and innovation strategies that exacerbate inequality in low-income countries (Osoba 2020). In sum, the lenses of the leading countries and organizations are likely to shape the financial, developmental, and regulatory aspects of AI. Because the Global North seems to dominate the current conversation, fields such as international development, postcolonial studies, and others can provide contrasting perspectives that can and should inform AI ethics and policy.

Along related lines, research that examines the role of the public in shaping AI ethics and policy needs to be reframed to become more inclusive so that it more fully takes into account a broader range of perspectives on gender, race, and socioeconomic factors. Moreover, research on participatory processes and the role of the public versus experts in shaping decision-making can scrutinize and inform the trajectory of AI ethics and governance.

7.3.3 Principles to Practice

Most of the aforementioned studies are also concerned with the goals of the documents, the motivations of the actors producing the documents, and whether and how normative documents can effectively lead to changes in practice. Some of the studies suggest that normative AI documents are not really intended to produce substantive change; rather, the intent is to improve the public image of organizations and/or protect them from public scrutiny. Hagendorff, for example, states that in the context of AI, "ethical considerations are mainly used for public relations purposes" (2020, 11). Hagendorff cites Boddington (2017, 56) in support of this claim, though to be fair, Bodington merely states that there is "a danger" that codes of ethics might only serve public relations purposes.

Others question whether broad ethical principles can effectively change practice, even if they are intended to do so (e.g., Cath et al. 2018; Hagendorff 2020). Hagendorff also cites McNamara et al. (2018) as evidence that ethics codes have virtually no effect on practice. In the study by McNamara and colleagues, 63 software engineering students and 105 professional software developers were given software-related scenarios and asked questions about ethical decision-making. Study participants who received a copy of the Association for Computing Machinery (ACM) code of ethics did not exhibit statistically significant differences in their

decision-making in 11 ethical scenarios compared to participants in the control group who did not receive a copy of the ACM code.

While the findings from the McNamara et al. study may seem troubling, they do not necessarily prove that ethics codes lack value. Yet, Fjeld et al. (2020) suggest that ethical principles are only "gently persuasive" *unless* they are enmeshed in governance structures. Similarly, Daly et al. (2019) note challenges in enforcing ethical principles for AI, while Jobin et al. (2019), Mittelstadt (2019), and Morley et al. (2019) argue that guidance and tools to address AI ethics issues are not sufficiently available and developed.

Furthermore, many question the goals and motivations of AI organizations and raise concerns about "ethics washing," especially with regard to corporations. On this perspective, organizations may lack even the fundamental motivations to carry out effective change. Yet, it is notoriously difficult to determine motivations, particularly if an entity has a complex organizational structure (Abebe et al. 2020; Bietti 2020). A variety of research methods (document review, interviews, observations, text analysis) and theoretical perspectives should be applied to better understand organizational motivations.

Moreover, even when it is clear that an organization's motivation is to translate ethical principles into practice, it is challenging to measure efficacy in achieving this goal. Morley et al. (2019) have done systematic work on mapping tools and methodologies to address ethical issues in the AI-development pipeline, but existing tools and methodologies, including governance structures and processes of evaluation, require further advancement. The question of how to translate principles to practice should be a top priority for all AI stakeholders (Schiff et al. 2020, 2021). Our work described in the next section may contribute to this priority area by mapping AI ethics documents' levels of engagement with law and regulation as one proxy for how seriously they may be thinking through the implementation of ethical considerations.

7.4 Building on the AI Ethics Literature

In an ongoing project, we are examining a collection of documents that, in their entirety or at least in part, seek to identify ethical issues and/or make recommendations about ethical practice related to AI. Our collection consists of more than 110 normative AI documents across 25 countries from the public, private, or NGO sectors. By "public sector," we mean that the authoring organization is connected to a governmental entity. "Private sector" refers to corporate entities such as Microsoft or Tencent. The "NGO sector" includes non-profits, professional organizations such as IEEE, and hybrid entities that involve collaborations across sectors such as the Partnership on AI. To fit within our inclusion criteria, documents must be publicly available, published between 2016 and July 2019, and have an English-language version. The collection of documents includes frameworks, policy strategies, and

reports with ethics sections. It does not include traditional academic publications or single-authored opinion articles.

We included documents addressing AI or similar terms, such as machine intelligence, machine learning, and automation, that have at least some ethics component. AI ethics intersects with robot ethics (e.g., concerns related to social robots and military robots), and robot ethics has a long and rich history. However, we excluded documents on robot ethics as belonging to a different realm of discourse unless they address AI to a significant degree. We also excluded documents that focus narrowly on a single category of technology, such as autonomous vehicles or military robots. Of 224 potential documents identified, our final sample of 112 consists of 54 from the public sector, 26 from the private sector, and 32 from the NGO sector. Our data search and inclusion process and the methodologies used to study the frequency of ethical topics across public, private, and NGO sectors are detailed in other work (Schiff et al. 2020, 2021).

In general, the documents we analyzed are not codes of ethics in its narrow/formal definition. In that sense of the term, a code of ethics is a set of moral principles or standards designed to guide the behavior of the members of a particular organization or profession. According to Davis (2013), a code of ethics applies "...to participants in a legitimate voluntary activity." Codes of ethics usually seek to achieve one or more of the following: provide authoritative rules or guidance for individuals new to the profession, remind experienced members of the ethical standards they are expected to uphold, and call attention to new areas of concern. They also can serve as a framework for resolving disputes among members about what constitutes ethical practice (Davis 2015). Moreover, they help individuals outside of the group (for example, the public) to calibrate their expectations about those who are within the group (for example, physicians in the case of the AMA) (Davis 2013).

In contrast to formal codes of ethics, many of the documents we reviewed do not have specific or directly defined target audiences. In addition, many documents identify clusters of ethical issues as being important, without necessarily articulating rules or standards for behavior. While the documents we reviewed are not limited to formal codes of ethics, they do share many of the characteristics identified by Davis (2015). They exist alongside other ethics-related efforts, sometimes serving as the motivation for new activities and sometimes as their consequence. These documents also can be classified as "practical" or "institutional" ethics documents, insofar as they apply to individuals or organizations engaged in developing, utilizing, or governing a specific activity, in this case, AI (Davis 2015). Furthermore, the majority of these documents articulate ethical standards that developers, practitioners, and users of AI should follow that go beyond what common morality demands, and they push their audiences to consider how these standards should play out in both the current and future uses of these new technologies. While the documents analyzed here do not encompass the totality of current AI ethics efforts, we believe they are a valuable distillation of organizational perspectives and activities from around the world.

Informed by our review of the AI ethics literature, we extrapolated a set of six main organizational motivations that may underlie the creation of AI ethics

documents (Schiff et al. 2020). In the next section, we introduce our typology of motivations; the typology challenges the notion that motives can be simply sorted in a binary of "sincere" or "disingenuous." Instead, it seeks to move towards a more nuanced understanding of the reasons why organizations are engaging in AI ethics initiatives.

7.5 A Typology of Motivations

On the basis of our review of global AI ethics documents and the existing literature on these documents, we concluded that it is important to develop a typology of motivations that could reveal insights about an organization's target audiences and goals, as well as illuminate the purposes of AI ethics documents and their prospects for success. As noted in Section 3, a number of commentators have argued that AI ethics documents are largely exercises in public relations and "ethics washing." While public relations is certainly a factor in the creation of some ethics documents, it is not the only one, and getter clearer about the range of possible motivations and goals can help us to understand better what these documents might help to achieve.

For any given AI ethics document – or ethics document more generally – ascertaining the goal of the document or the motivations of those who produce it is challenging. Stated motivations can differ from actual ones, and actual motivations are not always clear potentially even to those who develop a document. Goals and motivations can overlap and, in some instances, conflict with one another. In the case of complex organizations such as governments or Big Tech firms, different divisions might have conflicting goals. Throughout our research, we developed a typology of six different types of motivations that we believe to be operative in organizations that produce AI ethics documents. The six motivation-types can be clustered into three pairs that are conceptually distinct and potentially overlapping. Moreover, the two motivations within each pair are also not mutually exclusive. The first pair addresses end goals, the second pair addresses strategies for achieving those end goals, and the third focuses on perception and public relations. Thinking of these motivations as ideal types or constructs that are only partially instantiated in practice may be helpful in understanding behavior (Weber 1949) (Table 7.2).

7.5.1 Motivations One and Two: Goals

The first two motivation-types are *social responsibility* and *competitive advantage.* The former is the motivation to promote social benefits and reduce the risk of harm, while the latter is the motivation to gain or increase an advantage (e.g., economic or political) over others. Both of these connect to achieving particular goals, and the two types of motivations are in many cases consistent with one another, such that an organization could be motivated by either or both simultaneously. Furthermore,

Table 7.2 Typology of motivations

	Motivation types	
Goals	*Social Responsibility*—the motivation to promote social benefits and reduce the risk of harm	*Competitive Advantage*—the motivation to gain or increase an advantage (e.g., economic or political) over others
Strategies	*Strategic Planning*—the motivation to aid with internal strategic planning or organizational change	*Strategic Intervention*—the motivation to intervene in the surrounding (external) environment, including the legal and regulatory environment
Signals	*Signaling Social Responsibility*—the motivation to be perceived to be promoting social benefits and reducing risk of harm, whether or not one is actually doing so	*Signaling Leadership*—the motivation to be perceived to be a leader in the field of AI, or to be perceived to have a particular sort of competitive advantage

even though it can be difficult to determine in every individual case which of these motivations of operative, both types of motivations are helpful in understanding the production of normative AI documents.

Those who draw attention to ethics washing are concerned that the production of ethics documents may be solely or predominantly tied to achieving a competitive advantage (Johnson 2019), and these worries are perhaps justified with regard to some organizations or governmental entities. Yet, many ethics documents seem to be trying to promote societal good (Bietti 2020). For example, IEEE's *Ethically Aligned Design* (2019) states that its goal is to align AI to "values and ethical principles that prioritize human well-being in a given cultural context," and SAP's Guiding Principles for Artificial Intelligence (2018) describes SAP's motivation "to help the world run better and improve people's lives" It is difficult to explain the production of documents such as these, the *Montreal Declaration for Responsible AI* (2018), and the European Commission's *Ethical Guidelines for Trustworthy AI* (2019) without reference to some level of concern for social responsibility.

7.5.2 Motivations Three and Four: Strategies

Motivations three and four are *strategic planning* and *strategic intervention*. Regarding the former, an organization produces a normative AI document in order to aid with internal strategic planning or organizational change. For example, a corporation could develop an ethics document to serve as a foundation for best practice guidelines that influence the norms, policies, and procedures for its labs or the culture of its workplace, or a government could produce a normative AI document to serve as a blueprint for a national AI strategy. Many countries, including France, Mexico, and Qatar, have produced documents for this stated purpose (British Embassy in Mexico, Oxford Insights, and C Minds 2018; Villani et al. 2018; Qatar Center for Artificial Intelligence 2019). For example, Qatar's *Blueprint National Artificial Intelligence Strategy* expresses a motivation to "identify the key pillars to

build a great AI research and innovation ecosystem in Qatar and follow those with recommendations for action."

Regarding motivation four, an organization develops a document in order to intervene in the surrounding (external) environment, including the legal and regulatory environment. A firm could produce a document as a part of a strategy of intervening in the regulatory environment to gain an economic advantage, such as by blocking government regulation through the promise of voluntary self-regulation. For example, Microsoft's The Future Computed (2018, 9–10) suggests that while regulation is important, it will take "more than a couple of years… but almost certainly less than two decades." According to the Microsoft report, "AI technology needs to continue to develop and mature before rules can be crafted to govern it."

We refer to these two motivation-types as strategic because they are, to a significant extent, means for effecting broader ends – in particular, the broader ends of social responsibility and/or competitive advantage. As is the case with the first pair of motivation-types, motivations three and four are largely independent of one another, in that, an organization could pursue either or both simultaneously. Furthermore, this second pair is orthogonal to the first, in that either type in the second pair could be adopted in pursuit of either type in the first pair. A firm could be motivated to intervene in the regulatory environment in order to gain an economic advantage, but it could also do this out of a genuine concern for social responsibility.

7.5.3 Motivations Five and Six: Signaling

Motivations five and six are what we call *signaling social responsibility* and *signaling leadership*. The first of these is the motivation to be *perceived* to be promoting social benefits and mitigating risks – whether or not one actually is doing so. Some organizations might desire to signal social responsibility even if they are not genuinely motivated by social responsibility. Alternatively, some organizations might be motivated to both signal and promote social responsibility. Indeed, they might reasonably believe that signaling their own commitment to social responsibility will actually generate social benefits by encouraging others to act responsibly as well.

In signaling leadership, an organization is motivated to be perceived to be a leader in the field of AI – or, to put it differently, motivated to be perceived to have a particular sort of competitive advantage. An organization might signal leadership in order to expand markets, improve its reputation, or gain a seat at the planning table. This might be important for countries that are attempting to attract investment, firms that are trying to expand, or NGOs that are seeking influence in policymaking; it is particularly crucial for those that wish to have a leadership role but are not already perceived to have such a role. The AI ethics documents of the governments of Australia, India, Mexico, New Zealand, and Tunisia all reference the documents created by other governments, and they describe their own documents as entryways into this more established group. For example, Mexico's document

(2018) states that it is the "first nation in Latin America to join this elite club" and expresses pride in being "one of the first ten countries in the world to deliver a National Strategy for AI." Qatar (2019) expresses explicitly that its "vision is to have AI so pervasive in all aspects of life, business and governance in Qatar that everyone looks up to Qatar as a role model."

However, signaling could be an exercise in ethics washing. Roughly stated, the concept refers to the practice of trying to appear to be ethical but not performing behaviors that are consistent with ethical practice. Within the AI context, Johnson (2019) describes ethics washing as "...the practice of fabricating or exaggerating a company's interest in equitable AI systems that work for everyone." Google is one of a number of prominent companies that has been accused of this behavior, in part because its AI ethics board allegedly had no real veto power and was quickly abandoned after public criticism of its composition (Hao 2019).

7.6 Efficacy

In light of this set of motivations, this section looks at the key question of the efficacy of AI documents; in other words, do the documents generate the sorts of changes that their authors intend? One metric of the efficacy of such documents is whether they contribute to the modification of internal policies or practices within an organization. This could take the form, for example, of recommending, and then later implementing, a new ethics review process internal to a company that scrutinizes the AI systems the company develops. For example, SAP (2018) was among the first companies to propose an "AI Ethics Steering Committee" and "AI Ethics Advisory Board" and Microsoft has proposed designation of internal "Responsible AI Champions," a strategy that is now being emulated by the U.S. Department of Defense's Joint Artificial Intelligence Center (Barnett 2020; O'Brien et al. 2020).

Another metric is whether a document helps to generate change external to the authoring organization. Within a document, a company might try, for example, to convince a government to develop a new law or regulation or revise the mission of a government agency. Intel's *Artificial Intelligence: The Public Policy Opportunity* recommends governments remove "barriers to the access of data," "identify and mitigate discrimination caused by the use of AI," and "encourage investment in AI R&D" (Intel 2017). The Information Technology Industry Council (2017) recommends the creation of public-private partnerships and expanded efforts to improve STEM education and workforce training and adjustment programs.

A different potential target of the documents is research practices within academia or industry. The Institute of Business Ethics (2018), for instance, encourages organizations to "Establish a multi-disciplinary Ethics Research Unit to examine the implications of AI research and potential applications." Along related lines, many of the documents voice a call to redesign the computing curriculum or the educational system more generally. A report by the Future of Humanity Institute and other collaborators states that "Educational efforts might be beneficial in

highlighting the risks of malicious applications to AI researchers, and fostering preparedness to make decisions about when technologies should be open, and how they should be designed, in order to mitigate such risks" (Brundage et al. 2018). It is difficult to establish direct causal links between specific AI documents and the renewed push for ethics education in the computing curriculum, but changes are certainly happening. For instance, the Mozilla Foundation (2018) is sponsoring the Responsible Computing Science Challenge, which aims to "unearth and spark innovative coursework." An overarching hope is that the documents are contributing to a culture shift in terms of computing fields taking ethics more seriously. This, in part, would involve more fully integrating ethical considerations into computing research, design, and implementation.

A variety of factors may shape whether documents achieve these internal or external changes. We posit that documents are more likely to generate tangible impacts if: they engage with issues of regulation and policy; articulate their goals and strategies in detail rather than superficially; include participatory and public engagement in the document's creation; encourage mechanisms for monitoring and enforcement, and have plans for iteration and follow-up.

7.7 Lessons from CRISPR

What can be accomplished through the creation and distribution of AI ethics documents remains to be determined. At least some of the drafting authors and organizations likely hope for substantive changes to industry practices and meaningful regulations pertaining to AI: yet it is far from clear what will transpire. Here, recent history may be instructive, as many of the previous pushes to create ethics documents also occurred in response to emerging technologies, such as nuclear energy and recombinant DNA. What resulted from these initiatives could provide some guideposts in terms of what the legacy and impact of AI documents might be and can help direct the attention of AI stakeholders on what pitfalls to avoid. One such guidepost surrounds the need for diverse and public participation in how ethical consensus is shaped.

To illuminate this point, we will focus on a relatively recent example, the use of CRISPR, clusters of regularly interspaced short palindromic repeats, to edit human embryos. The case of CRISPR represents a clear instance in which a significant global consensus emerged on the need to ban the use of germline gene editing. This consensus resulted from a concerted effort to articulate key ethical principles relevant to this technology and to involve key stakeholders both in the articulation and use of these principles. Though there was a failure to prevent the Chinese scientist He Jiankui from using CRISPR to edit viable human embryos, the global scientific community collectively condemned He's actions and called for stronger international guidelines and oversight for this technology. This suggests that the gene editing community has progressed in the articulation of meaningful regulations and practices.

Guidelines governing the modification of the human genome, such as the Council of Europe's Convention on Human Rights and Biomedicine from 1997, have been in existence for more than 20 years. Concerns around CRISPR specifically started to intensify in the 2010's as researchers began to use the technique to edit both human somatic cells and human germline cells, as well as to potentially eliminate genetic diseases (Evitt et al. 2015; Ledford 2020). Many ethics documents were generated in response, including statements put out by the organizing committee of the International Summit on Human Gene Editing (2015, 2017), the Council of Europe's Committee on Bioethics "Statement on Genome Editing Technologies" (2015), the National Academies' "Human Genome Editing: Science, Ethics and Governance" (2017), and the Alliance for Regenerative Medicine's "Statement of Principles on Genome Editing" (2019).

Documents on AI and gene editing have many similarities. Like AI ethics documents (Floridi and Cowls 2019), the authors of these guidelines drew heavily on existing bioethical norms and standards, such as the Universal Declaration on the Human Genome and Human Rights, and ethical principles pertaining to human subjects research. Fundamental principles of human rights, safety, social justice, and an emphasis on the need for expanded social discussion and debate about the ethical issues involved in gene editing are standard throughout these guidelines and statements (Brokowski 2018). The motivations behind these statements are also similar to those found in the AI guidelines, including discussions of social responsibility and leadership, as well as efforts to signal social responsibility and change external perceptions.[1]

Given these similarities, it is instructive to consider whether ethics documents surrounding gene editing led to meaningful change. Of note, modification of the human germline is now banned in most countries. According to a 2014 survey, regulations had been enacted by 39 different countries, with 29 having an outright ban on gene modification, 9 with ambiguous regulation, and the remaining country, the USA, severely regulating its use in clinical trials (Araki and Ishii 2014). These regulations align with mainstream sentiments in ethics documents produced on this topic, according to work by Brokowski (2018). Brokowski examined 61 ethics statements by governments, non-governmental organizations, and private companies from 2015–2018: 65% of the statements indicated that the clinical use of germline editing should be impermissible at this current time, 30% expressed no clear stance on its permissibility, and only 5% of the reports expressed openness to further exploring the possible applications of gene editing technologies in this area.

What led to this relative clarity of purpose? Foremost, the CRISPR case seems to emphasize the importance of having a robust participatory process. For example, the guidelines developed by international conferences such as the U.S. National Academies' international conference in 2015 and the International Summit on Human Gene Editing in 2015 and 2018 represent international gatherings of experts

[1] For examples, see documents by Merck (2017) and the Organizing Committee of the Second International Summit on Human Genome Editing (2018).

that produced guidelines critical to the future ethical use of CRISPR. The 2018 International Summit, held only a few days after the announcement that Dr. He Jiankui had edited viable human embryos, is especially striking in its involvement of 500 individuals; it included not only researchers but ethicists and patient group representatives as well.

As Jasanoff, Hurlbut, and Saha (2015) stated in their article on lessons to be learned for deliberations around the ethics of CRISPR, "…studies of technical controversies have repeatedly shown that public opposition reflects not technical misunderstanding but different ideas from those of experts about how to live well with emerging technologies." To ensure the safe, equitable, and ethical use of emerging technologies, the voices of all stakeholders must play a role in setting standards for the future. Fostering such a process in the development of AI ethics documents may make it more likely that they will have a tangible and positive impact on practices and regulations in the international AI community and ultimately on how the public is affected by AI in the future. Considering this observation, the fact that so many AI ethics documents are produced by a limited set of leading countries and multinational organizations in the Global North is a lingering source of concern.

7.8 Conclusion

In this chapter, we sought to provide context for why AI ethics documents are being created, to review the current literature on these documents, and to move beyond the discussion on whether a presumed consensus on important ethical principles exists. We highlighted unanswered questions around representation and power, the translation of principles to practice, and the need to examine more deeply, something we endeavored to do here, the motivations underlying document creation. Finally, drawing from lessons in the recent case of CRISPR, we argued that for AI ethics documents to achieve beneficial impacts for society, it is essential to foster more diverse and inclusive participatory processes.

References

Abebe, Rediet, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260. FAT* '20. Barcelona, Spain: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372871.

Alliance for Regenerative Medicine. 2019. *The Alliance for Regenerative Medicine Releases Statement of Principles on Genome Editing*. Washington, DC: Alliance for Regenerative Medicine.

Araki, M., and T. Ishii. 2014. International regulatory landscape and integration of corrective genome editing into in vitro fertilization. *Reproductive Biology and Endocrinology* 12: 108. https://doi.org/10.1186/1477-7827-12-108.

Automated Decision Systems Task Force. 2019. *New York City Automated Decision Systems Task Force Report*. New York: Automated Decision Systems Task Force.

Bagloee, Saeed Asadi, Madjid Tavana, Mohsen Asadi, and Tracey Oliver. 2016. Autonomous Vehicles: Challenges, Opportunities, and Future Implications for Transportation Policies. *Journal of Modern Transportation* 24 (4): 284–303. https://doi.org/10.1007/s40534-016-0117-3.

Barnett, Jackson. 2020. JAIC Launches Pilot for Implementing New DOD AI Ethics Principles. *FedScoop* 2 (April): 2020.

Bennett, Colin J., and Charles D. Raab. 2017. *The Governance of Privacy: Policy Instruments in Global Perspective*. London: New York Routledge.

Bietti, Elettra. 2020. From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–219. FAT* '20. Barcelona, Spain: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372860.

Boddington, Paula. 2017. *Towards a Code of Ethics for Artificial Intelligence*, Artificial Intelligence: Foundations, Theory, and Algorithms. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-60648-4.

British Embassy in Mexico, Oxford Insights, and C Minds. 2018. *Towards an AI Strategy in Mexico: Harnessing the AI Revolution*. Mexico City: British Embassy in Mexico, Oxford Insights, and C Minds.

Brokowski, Carolyn. 2018. Do CRISPR Germline Ethics Statements Cut It? *The CRISPR Journal* 1 (2): 115–125. https://doi.org/10.1089/crispr.2017.0024.

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, Centre for the Study of Existential Risk, Center for a New American Security, Electronic Frontier Foundation, OpenAI. http://arxiv.org/abs/1802.07228.

Cath, Corinne, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. 2018. Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach. *Science and Engineering Ethics* 24 (2): 505–528. https://doi.org/10.1007/s11948-017-9901-7.

Char, Danton S., Nigam H. Shah, and David Magnus. 2018. Implementing Machine Learning in Health Care – Addressing Ethical Challenges. *The New England Journal of Medicine* 378 (11): 981–983. https://doi.org/10.1056/NEJMp1714229.

Chatila, Raja, and John C. Havens. 2019. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In *Robotics and Well-Being*, Intelligent Systems, Control and Automation: Science and Engineering, ed. Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Osman Tokhi, and Endre E. Kadar, 11–16. Cham: Springer. https://doi.org/10.1007/978-3-030-12524-0_2.

Committee on Bioethics, Council of Europe. 2015. *Statement on Genome Editing Technologies*. Strasbourg.

Daly, Angela, Thilo Hagendorff, Hui Li, Monique Mann, Vidushi Marda, Ben Wagner, Wei Wang, and Saskia Witteborn. 2019. *Artificial Intelligence, Governance and Ethics: Global Perspectives*, SSRN Scholarly Paper ID 3414805. Rochester: Social Science Research Network. https://papers.ssrn.com/abstract=3414805.

Davis, Michael. 2013. Codes of Ethics. In *International Encyclopedia of Ethics*, edited by Hugh Lafollette, wbiee018. Oxford: Blackwell Publishing Ltd. https://doi.org/10.1002/9781444367072.wbiee018.

———. 2015. Codes of Ethics. In *Ethics, Science, Technology, and Engineering: A Global Resource*, ed. J. Britt Holbrook and Carl Mitcham, 2. ed. Farmington Hills, Mich.: Gale, Cengage Learning/Macmillan Reference USA.

Duan, Yanqing, John S. Edwards, and Yogesh K. Dwivedi. 2019. Artificial Intelligence for Decision Making in the Era of Big Data – Evolution, Challenges, and Research Agenda. *International Journal of Information Management* 48 (October): 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021.

Dutton, Tim, Brent Barron, and Gaga Boskovic. 2018. *Building an AI World: Report on National and Regional AI Strategies*. Toronto: CIFAR.

European Commission, High-Level Expert Group on Artificial Intelligence (AI HLEG). 2019. *Ethical Guidelines for Trustworthy AI*. Brussels: European Commission, High-Level Expert Group on Artificial Intelligence (AI HLEG).

Evitt, Niklaus H., Shamik Mascharak, and Russ B. Altman. 2015. Human Germline CRISPR-Cas Modification: Toward a Regulatory Framework. *The American Journal of Bioethics* 15 (12): 25–29. https://doi.org/10.1080/15265161.2015.1104160.

Fast Company. 2017. How Apple, Facebook, Amazon, And Google Use AI To Best Each Other. *Fast Company* 11 (October): 2017.

Fejerskov, Adam Moe. 2017. The New Technopolitics of Development and the Global South as a Laboratory of Technological Experimentation. *Science, Technology, & Human Values* 42 (5): 947–968. https://doi.org/10.1177/0162243917709934.

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*, SSRN Scholarly Paper ID 3518482. Rochester: Berkman Klein Center for Internet & Society. https://papers.ssrn.com/abstract=3518482.

Floridi, Luciano, and Josh Cowls. 2019. "A Unified Framework of Five Principles for AI in Society." *Harvard Data Science Review*. June. https://doi.org/10.1162/99608f92.8cd550d1.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. The Future of Employment: How Susceptible Are Jobs to Computerisation? *Technological Forecasting and Social Change* 114 (January): 254–280. https://doi.org/10.1016/j.techfore.2016.08.019.

Gartenberg, Chaim. 2018. Google Wants to Teach More People AI and Machine Learning with a Free Online Course. *The Verge* 28 (February): 2018.

Gibert, Martin, Christophe Mondin, and Guillaume Chicoisne. 2018. *Montréal Declaration of Responsible AI: 2018 Overview of International Recommendations for AI Ethics*. University of Montréal.

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *Critical and Ethical Studies of Digital and Social Media* 10.

Hagendorff, Thilo. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds & Machines*. https://doi.org/10.1007/s11023-020-09517-8.

Hao, Karen. 2019. In 2020, Let's Stop AI Ethics-Washing and Actually Do Something. *MIT Technology Review* 27 (December): 2019.

Information Technology Industry Council. 2017. *ITI AI Policy Principles*. Washington, DC: Information Technology Industry Council (ITI).

Institute for Business Ethics. 2018. *Business Ethics and Artificial Intelligence*. 58. Institute for Business Ethics.

Intel. 2017. *Artificial Intelligence: The Public Policy Opportunity*. Santa Clara: Intel.

Jasanoff, Sheila, Krishanu Saha, and J.Benjamin Hurlbut. 2015. CRISPR democracy: Gene editing and the need for inclusive deliberation. *Issues in Science and Technology* 32: 12.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1 (9): 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Johnson, Khari. 2019. How AI Companies Can Avoid Ethics Washing. *VentureBeat*, July 17, 2019, sec. AI.

Kak, Amba. 2020. 'The Global South Is Everywhere, but Also Always Somewhere': National Policy Narratives and AI Justice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 307–312. AIES '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375859.

Ledford, Heidi. 2020. Quest to Use CRISPR against Disease Gains Ground. *Nature* 577 (7789): 156–156. https://doi.org/10.1038/d41586-019-03919-0.

Mai, Jens-Erik. 2016. Big Data Privacy: The Datafication of Personal Information. *The Information Society* 32 (3): 192–199. https://doi.org/10.1080/01972243.2016.1153010.

McKinsey Global Institute. 2018. *Notes From the AI Frontier: Modeling the Impact of AI on the World Economy*. McKinsey Global Institute.

McNamara, Andrew, Justin Smith, and Emerson Murphy-Hill. 2018. Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, 729–733. (Association for Computing Machinery. https://doi.org/10.1145/3236024.3264833.

Merck. 2017. Genome Editing Technology – Principle. Merck.

Mittelstadt, Brent. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1: 501–507. https://doi.org/10.1038/s42256-019-0114-4.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 3 (2): 205395171667967. https://doi.org/10.1177/2053951716679679.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2019. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices." *Science and Engineering Ethics*, December. https://doi.org/10.1007/s11948-019-00165-5.

Mozilla Foundation. 2018. Announcing a Competition for Ethics in Computer Science, , with up to $3.5 Million in Prizes. *The Mozilla Blog*. October 10, 2018.

National Academies of Sciences, Engineering, and Medicine. 2015. *International Summit on Human Gene Editing: A Global Discussion*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21913.

———. 2017. *Human Genome Editing: Science, Ethics, and Governance*. Washington, DC: National Academies Press. https://doi.org/10.17226/24623.

O'Brien, Tim, Steve Sweetman, Natasha Crampton, and Venky Veeraraghavan. 2020. A Model for Ethical Artificial Intelligence. *World Economic Forum* 14 (January): 2020.

OECD. 2019. *OECD: Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449. OECD Legal Instruments. Paris: OECD.

Organizing Committee of the Second International Summit on Human Genome Editing. 2018. *Statement by the Organizing Committee of the Second International Summit on Human Genome Editing*. Washington, DC: National Academies of Sciences, Engineering, and Medicine.

Osoba, Osonde A. 2020. Technocultural Pluralism: A 'Clash of Civilizations' in Technology?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 132–137. AIES '20. New York: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375834.

PricewaterhouseCoopers. 2017. *Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise?* London: PricewaterhouseCoopers.

Qatar Center for Artificial Intelligence. 2019. *Blueprint: National Artificial Intelligence Strategy for Qatar*. Ar-Rayyan: Qatar Center for Artificial Intelligence (QCAI), Qatar Computing Research Institute (QCRI), Hamad Bin Khalifa University.

SAP. 2018. *SAP's Guiding Principles for Artificial Intelligence*. Waldorf: SAP.

Scherer, Matthew U. 2016. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology* 29 (2): 353–400.

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–158. AIES '20. New York: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375804.

Schiff, Daniel, Jason Borenstein, Kelly Laas and Justin Biddle. 2021. AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection. *IEEE Transactions on Technology and Society* 2 (1): 31–42. https://doi.org/10.1109/TTS.2021.3052127.

Schwab, Klaus. 2016. *The Fourth Industrial Revolution*. First U.S. edition. New York: Crown Business.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems, First Edition*. Piscataway: IEEE.

The Institute for Ethical AI & ML. 2020. *The AI-RFX Procurement Framework*.

Todd, Deborah. 2019. Microsoft Reconsidering AI Ethics Review Plan. *Forbes*, June 24, 2019, sec. Innovation.

Villani, Cédric, Marc Schoenauer, Yann Bonnet, Charly Berthet, Anne-Charlotte Cornut, François Levin, and Bertrand Rondepierre. 2018. *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*. French Parliament. https://frenchamerican.org/young-leader/cedric-villani/

Vincent, James. 2019. Finland Is Making Its Online AI Crash Course Free to the World. *The Verge* 18 (December): 2019.

Weber, Max. 1949. 'Objectivity' in Social Science and Social Policy. In *The Methodology of the Social Sciences, by Max Weber, Edward Shils, and Henry A Finch*, 49–112. Glencoe: Free Press.

West, Darrell M. 2018. *The Future of Work: Robots, AI, and Automation*. Washington, DC: Brookings Institution Press.

Zeng, Yi, Enmeng Lu, and Cunqing Huangfu. 2018. Linking Artificial Intelligence Principles. *ArXiv:1812.04814 [Cs]* (December) http://arxiv.org/abs/1812.04814.

Daniel S. Schiff PhD Candidate, School of Public Policy, Georgia Institute of Technology, USA; schiff@gatech.edu. Daniel Schiff studies issues related to the intersection of AI and policy, including research on education, labor, misinformation, governance of AI, corporate social responsibility, and other social and ethical implications of AI.

Kelly Laas Librarian and Ethics Instructor, Center for the Study of Ethics in the Professions, Illinois Institute of Technology, USA; laas@iit.edu. Her research interests include the history and use of codes of ethics in professional fields, ethics education in STEM, research ethics, and integrating ethics into technical curricula.

Justin B. Biddle Associate Professor, School of Public Policy, Georgia Institute of Technology, USA; justin.biddle@pubpolicy.gatech.edu. Justin Biddle works at the intersection of a number of fields, including philosophy of science and technology, ethics of emerging technologies, and science and technology policy. He is particularly interested in the ethics of artificial intelligence and the role of value judgments in the design and development of computing technologies.

Jason Borenstein Director of Graduate Research Ethics Programs, School of Public Policy and the Office of Graduate Studies, Georgia Institute of Technology, USA; borenstein@gatech.edu. Dr. Borenstein's teaching and research interests include engineering ethics, AI & robot ethics, research ethics, and bioethics. His ethics-related research includes topics such as autonomous vehicles, human-robot interaction, community engagement, and AI & bias.