A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data

Loi, Michele ; Heitz, Christoph ; Christen, Markus

DOI: https://doi.org/10.1109/sds49233.2020.00015

A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data

Michele Loi
Digital Society Initiative
University of Zurich
Zurich, Switzerland
michele.loi@uzh.ch

Christoph Heitz
School of Engineering
Zurich University of Applied Sciences
Winterthur, Switzerland
christoph.heitz@zhaw.ch

Markus Christen
Digital Society Initiative
University of Zurich
Zurich, Switzerland
christen@ethik.uzh.ch

***Abstract*— Up to date, more than 80 codes exist for handling ethical risks of artificial intelligence and big data. In this paper, we analyse where those codes converge and where they differ. Based on an in-depth analysis of 20 guidelines, we identify three procedural action types (1. control and document, 2. inform, 3. assign responsibility) as well as four clusters of ethical values whose promotion or protection is supported by the procedural activities. We achieve a synthesis of previous approaches with a framework of seven principles, combining the four principles of biomedical ethics with three distinct procedural principles: control, transparency and accountability.**

Keywords—data ethics, ethical guidelines, artificial intelligence

I. INTRODUCTION

It is generally acknowledged that the enormous business opportunities of the digital transformation require an ethical framing for avoiding unwanted developments such as undermining privacy through mass surveillance, or loss of meaningful human control upon increasingly intelligent systems. This has led to a plethora of codes for handling ethical risks of artificial intelligence (AI) and big data. The widest global review of guidelines on AI covers 84 guidelines [1], where AI is defined broadly, including guidelines on big data in general. This analysis highlights eleven distinct clusters of value that are found across the corpus (while no single value is found in every document). These are: 1) Transparency; 2) Justice & Fairness; 3) Nonmaleficence; 4) Responsibility; 5) Privacy; 6) Beneficence; 7) Freedom & Autonomy; 8) Trust; 9) Sustainability; 10) Dignity; and 11) Solidarity. These 11 values are derived inductively (bottom up) and empirically (*a posteriori*) from the thematic analysis of existing guidelines. They can be compared with the five principle list proposed by two important scholars in the ethics of information, Luciano Floridi and Josh Cowls [2], which is based on the comparison of six very influential guidelines [3]–[8]. The five principles are: 1) Beneficence (promoting well-being, preserving dignity, and sustaining the planet); 2) Nonmaleficence (privacy, security and capability caution); 3) Autonomy (the power to decide); 4) Justice (promoting prosperity, preserving solidarity, avoiding unfairness); and 5) Explicability (enabling the other principles through intelligibility and accountability). This list corresponds to the widely influential four principles of biomedical ethics [9], plus one other, namely explicability.

Those two lists exemplarily show the difficulties when generating a "value ontology" that guide ethical AI: some lists cluster values that others keep separated. Furthermore, many definitions of values remain rather vague and are thus hard to operationalize in concrete business settings. In our contribution, we propose an alternative way to analyse those guidelines by focusing on the kinds of activities they prescribe. By an in-depth analysis of 20 selected codes (research strategy illustrated in Section II), we will work out three procedural action types (Section III) and four axiological action types (Section IV) representing four value clusters. We outline (Section V) how this approach is better suited to generate guidelines that match the needs of businesses in search of AI based products and services that fulfil ethical requirements. In this way, our contribution aims to contribute to decreasing the complexity of the high number of ethics codes and make them better suitable for everyday use.

II. RESEARCH STRATEGY

A. Sample selection

We have thoroughly examined the content of 20 guidelines, a subset of the guidelines examined in the review by Jobin et al [1]. With the term "guideline" we summarize different types of documents, e.g. named as "principles", "recommendations", "vision", or "declaration". However, all of them contain a set of *prescriptions*, typically in the form of principles. To avoid bias, we selected the guidelines in the order in which they were listed on a third-party website: https://aiethics.herokuapp.com, excluding those not in English and those not accessible with the link provided. The full list of analysed guidelines corresponds to the one in [65], pp. 7-9. Our sample contains less geographic diversity than the original set (only European and Global/International organizations are represented). However, the diversity in the type of stakeholder groups issuing the guidelines and to which the guidelines are addressed is preserved. The issuing organizations include Think Tanks, Companies, NGO/NPO, Community of researchers and practitioners, Private sector alliance, Professional Association/Society, Mixed, IGO/supranational, Governmental agencies/organization, Federation/Union. The addressee stakeholder list includes: multiple stakeholders, the private sector, the public sector, self (the guidelines are written by the issuing organization for itself, e.g. a data ethics code written and endorsed by the same company) or is left unspecified. Like other scholars who have been working on subsets of the Jobin et al sample (e.g. [10]), our focus is on theoretical elaboration.

B. Research Question

The question we asked ourselves is if there is a structured way of analysing the guidelines with respect to the kinds of activities they prescribe, and a best way to reduce their complexity by transforming them into an actionable framework.

C. Research Framework

The thematic analysis of these activities has shown that all the activities mentioned in these guidelines can be grouped in four macro action types:

978-1-7281-7177-7/20/$31.00 ©2020 IEEE
DOI 10.1109/SDS49233.2020.00015

1. To know and to document

2. To inform and communicate

3. To take or assign responsibility

4. To improve results in an ethical dimension (e.g. to mitigate injustice, protect privacy, promote trust, etc…).

Although the actions types appearing in the different guidelines coincide to a large degree, the relation of action types to specific values or ethical principles differs between different guidelines. This is justifiable because the same action (e.g. documenting the unfairness of an algorithm) may be considered a building block of transparency (when the unfairness is communicated), of accountability (when the data scientist takes responsibility for it), and of justice (when actions are undertaken to reduce the documented unfairness). As a result, it is not easy to compare the different guidelines with respect to these activities. We thus propose to associate the first three action types with dedicated ethical principles as follows:

1. *Control*: knowledge and documentation
2. *Transparency*: knowledge and documentation + communication
3. *Accountability*: knowledge and documentation + taking and assigning responsibility

We justify this association in section III. We refer to these action types as "procedural" because they relate to procedures that can be realized irrespective of the substantive goal (e.g. security, privacy, customer satisfaction) that one pursues.

We associate the fourth type of action, improving results, by specifying the ethical dimension as belonging to the value cluster of: *beneficence, non-maleficence, autonomy and justice.* We refer to this action type as "axiological" because, in the guidelines, specific value terms are used when specifying their goals. E.g. the prescription to *remove unfair biases* refers to the value of *fairness* (value cluster of *justice*). These actions are described in section IV.

III. Procedural Action Types

A. Control

The first class of activity prescribed is to know and document. Different entities can be documented, e.g. software objects, human goals, engineering processes and human outcomes. For example, documentation is recommended for the legitimacy of data collection (e.g. with respect to informed consent or data minimization [11], [12]), the data used for training machine learning algorithms [13], the data potential for bias [14], the type of machine learning algorithm used, the optimization goal/loss function/reward function [6], the features used to train an algorithm / make decisions [15], the weightings of these features (if known) [15], the algorithm type, the extent of its opacity [15, 16], its procedure of validation [15, 16], the goal and purpose of the service [14], i.e. "intended use" [6], or the performance of an algorithm[15, 16], measured by key parameters such as accuracy [14, 16], bias [14, 17], direct-indirect discrimination (US disparate impact) [11, 13, 18], fairness [13, 16], sometimes combined (e.g. unfair bias [14], unfair discrimination [19, 20]), and robustness, understood as reliability/reproducibility [14, 16].

Finally, a large number of guidelines require documenting the impact on humans, and the (dis)enhancement of human skills [14] (see [21] for why AI can *dis*enhance as well as enhance human capacities). The definition of human impact is often left unspecified [19], [15], with generic terms such as benefit and empowerment [7, 17], "serv[ing] the planet" [17], "reasonably predictable misuse" [23], being used. When more concrete categories are mentioned (e.g. "public safety risk" [22], "arms race" [17], support for "Fundamental Freedoms and Rights" [17], "well-being" [6]), often no concrete documentation actions are defined. Some guidelines focus on the impact on concrete individuals, e.g. workers, customers, other people affected by decision-making. They recommend "ongoing monitoring" [12], especially for robustness, e.g. "verify how your system behaves in unexpected situations and environments?" [14] with emphasis on considering the human impact of the full life-cycle of the data [19].

B. Transparency

In addition to knowledge and documentation actions, transparency essentially involves a second action type, namely adequately and effectively communicating such knowledge. Depending on the circumstances, this involves actions such as explaining, disclosing, making something accessible. Transparency can have different targets, e.g. to end-users, auditors, the broad public. An organization may know that an algorithm has a bias problem and hide the information in a folder accessible only to the heads of data science: they have *control (knowledge)*, but they do not have *transparency.* The adequacy of communication is relative to the context, in particular effective communication is tuned to the needs of its audience, as emphasized in [16].

Some guidelines directly or indirectly refer to the explainability of so-called black boxes. For example, one guideline recommends a "why did you do that" button in AIs interacting with humans [6], one suggests that "[t]he data provided by the black box could also assist robots in explaining their actions in language human users can understand" [17], one that "[i]n some cases it may be appropriate to develop an automated explanation for each decision" [16], and another that "public sector organisations" should be mandated to explain "how the decision is reached and what would need to change for individuals to get a different outcome" [15]. These proposals echo approaches of explainable AI (X-AI) known as *post-hoc* explanations [24], and, in the last quote, counterfactual ones [25]. A distinct approach to transparency requires *accessibility of documented knowledge*, i.e. the documented knowledge described in (A. Control) must be made accessible to and understandable by auditors [16]. Another approach gives auditors more direct access to the algorithm, e.g. demand an "API" to "allow the research community to perform automated auditing" [16], or access to the source code [17].

C. Accountability

Accountability presupposes a degree of knowledge and control of the design, deployment, and, sometimes, specific output, of an AI system or a data-driven product, by individuals who are responsible for their quality and/or use. Hence, accountability presupposes the actions prescribed by the principle of control (A). More characteristically, accountability requires identifying the person or organization that is responsible if problems occur. Responsibility can be legal, moral or organizational. Responsibility in the legal sense is

associated with, for example, liability for damages. Moral responsibility determines who is to blame. Organizational responsibility means being the person formally in charge, within an organization, for problem detection and solutions. Accountability is not reducible to transparency, not even to transparency to external auditors. Auditors can determine a dangerous flaw in a data driven product. But if no one is legally liable, and no one is formally in charge to fix the problem, the governance falls short of accountability. In short, accountability involves all actions that help answering the question "who is responsible?", for all issues listed under control.

For AI guidelines specifically, the recurring idea is that it should always be possible to identify human responsibilities behind the decisions of autonomous/intelligent systems [6, 12, 22]. This action category includes, for example, actions of "registration and record-keeping" [6] and "governance programs […] to detect and remedy any possible discriminatory effects of the data and models" [11].

IV. AXIOLOGICAL ACTION TYPES: CLUSTERS

Many other actions are characterized by a value goal (e.g. privacy, or fairness) that ought to be promoted. If we exclude responsibility (which belongs to accountability) and transparency (already mentioned) from the 11 value list of [1], we are left with: Justice & fairness, Non-maleficence, Privacy, Beneficence, Freedom & autonomy, Trust, Sustainability, Dignity, and Solidarity. These nine values are also included in the first four value clusters of the AI ethics framework proposed by Floridi and Cowls. In the following, we describe clusters corresponding to the axiological action types adopting the first four value clusters of Floridi and Cowls and revising them slightly.

A. Beneficence

Floridi and Cowls use "beneficence" for actions of promoting well-being, found for example in [6, 14], of preserving dignity [6, 13, 17, 23, 26, 27], and of sustaining the planet [14, 18]. We propose to revise this value cluster as follows: *sustainability* should be moved from beneficence to non-maleficence, because if a product is sustainable, it does not harm the environment in which it operates in the long term. *Dignity* should be under both non-maleficence *and* autonomy, rather than beneficence. Trust does not appear anywhere in the 5-value cluster, which is a problem, since many guidelines mention trust [6, 7, 13, 17, 19, 22, 23, 26, 28–30] and trustworthiness [7, 14]. Trust is a lubricant of social systems [31], which facilitates exchange and cooperation, when invested in trustworthy parties [32, 33], so it is closest to the value cluster of benevolence (for it indirectly promotes well-being) although it cannot be subsumed perfectly under it.

B. Non-maleficence

Under non-maleficence (which means avoiding harm), Floridi and Cowls list the goals of privacy, security and "capability caution". We revise this classification as follows: privacy is as related with *both* non-maleficence (e.g. avoiding reputational harm) *and* autonomy (e.g. promoting the growth of authentic individuality [34, 35], as well social intimacy within closed or private groups which is necessary for "group agency" [35–37]). In the guidelines we have analysed, privacy is mentioned both as "privacy" and as control of personal information by the data subject [11, 12, 38,19], e.g. when "informational self-determination" is mentioned [19]. In other instances [28], privacy is understood as a condition in which information about the person is inaccessible by third parties [39–41]. In some cases, it is not clear which meaning (self-determination or privacy) is in question and both interpretation are equally plausible, e.g. the claim that AI's "ability to […] 'derive the intimate from the available'"[42] is a threat to privacy (for a scholarly treatment see [43]).

Security, mentioned in the guidelines [17, 28, 30], belongs to the non-maleficence value cluster, according to both [1] and Floridi and Cawls. By "capability caution", Floridi and Cawls mean avoiding the misuse of the enhanced capabilities of AIs [11, 12, 14, 15, 20, 22, 23, 26, 29, 30, 44, 45], but the concept of capability caution is so general that it may be extended to all data-driven services.

C. Autonomy

In contrast to Floridi and Cawls, we associate the guideline prescriptions to promote human dignity found in our analysis with the principle of autonomy. In the guideline texts, dignity is usually mentioned in contexts in which human rights are also mentioned, an in this tradition, human dignity is connected to respect for human agency [46], which creates a strong link to the *autonomy* value cluster. Autonomy is sometimes mentioned explicitly, e.g. by the guidelines on Trustworthy AI [14], which is concerned about AI "interfering with the (end) user's decision-making process in an unintended way" and the Algorithm Impact Assessments proposed by the Women in AI [15], which aim to ensure that "the algorithm does not become policy, thus removing human autonomy in wider decision making".

Jobin et al. combine freedom and autonomy in the same value cluster [1], i.e. they include "freedom of expression or informational self-determination, […] privacy-protecting user controls, [….] freedom […], empowerment […]".[1, p. 395]. We agree with their analysis. Our findings, based on a subset of the guidelines they have examined, are congruent with theirs. We found guidelines addressing freedom of expression [42]. We also count in the autonomy/freedom cluster all guidelines prescribing respect for human rights [6, 13, 19], because the fundamental civic liberties and political freedoms are standard items in the most widely ratified human rights documents [47].

D. Justice

Finally, many actions prescribe advancing/promoting justice, or fairness, explicitly or by implications. The most common recommendation prescribes avoiding "unfair biases" [14] in machine learning models and intelligent systems. One often finds that fairness is mentioned explicitly [13-16, 22, 42]. Other guidelines address the problem of discrimination [13, 18, 19, 42], which also belongs to this cluster, since unjustified discrimination (or simply "discrimination" in the more common, moralized use) is unjust. Guidelines addressing AI and machine learning are quite clear about the fact that the relevant discrimination is not direct (systems are rarely explicitly designed with the goal of giving a worse treatment to women, blacks or minorities) but an emergent properties of systems that may have biases that happen to be more harmful for individuals of specific groups. The "discrimination" in question is "indirect discrimination", i.e. unequal impact (or "disparate impact", in US legal jargon) on different groups [11, 16]. Guidelines refer to discrimination definition in anti-discrimination laws [18], human rights

[19], international law [29], others, beside the usual suspects ("race, sex, religion"), mention also features not included in anti-discrimination laws, e.g. "[…], gender identity, ability status, socio-economic status, education level, […], country of origin" [16]. Some guidelines prescribe eliminating bias from data [22, 28], or elimination biased data from training [12]. Others are more careful: they qualify the target of elimination as "*unfair* bias" [14, our italics] or "unlawful" [19]. In statistics and machine learning, "bias" can mean many different things, e.g. confounding bias, selection bias and measurement bias, bias as unequal accuracy (or even more specifically, unequal recall rate, or unequal precision), etc. In journalist parlance, bias often means unequal impact or, for the more sociologically and morally sophisticated, something like the highly political concept of "unequal impact due to societal biases and other social facts that are unjust where injustice ultimately explains the unequal impact". Machine learning scholars have studied the logical/mathematical (in)compatibility of different concepts [48, 49]. Computer scientists are gradually coming to grip with the normativity and complexity of the search for biases in data and algorithmic decisions [50–56]. Perhaps as a result, some guidelines explicitly prescribe that developers of data driven products should "consider which definition of fairness best applies to their context and application" [13]. Some guidelines also recognize the existence of a trade-off between fairness and accuracy, and still recommend that fairness may also be pursued at the expense of accuracy, achieving a reasonable compromise [11].

Some guidelines consider the impact of AI, algorithms, and more broadly data driven services on social justice more broadly. They require avoiding "self-fulfilling markers of success and reinforce patterns of inequality"[18], reproducing and aggravating "[e]xisting patterns of structural discrimination" [18], harmful stereotypes [29]. As positive objectives, they promote gender equality [18, 29, 45], diversity [14] and the inclusion of women [45]. Solidarity was mentioned as a value to promote and to not undermine in six guidelines [6, 8, 14, 20, 27, 44] and once in an ethical principle [8]. Like Floridi and Cowls before us [2], we consider solidarity as a member of this value cluster, because equality-based [58] and fraternity-based [59] ideas of justice lead to recommendations similar to those based on solidarity.

V. Towards a Unified Framework for the Assessment of Ethical Guidelines and Codes

How can the content of ethical guidelines and codes be expressed through the simplest possible framework? In this section, we suggest a novel framework for the systematic evaluation of ethical guidelines and codes in terms of their recommendations for the practice of data science.

The list of eleven values found in Jobin et al's analysis [1] is excellent for research purposes, but it may be too complex for the practice of data science. For it contains too many values, for data scientists to constantly consider in their practice. Therefore, we start with the minimal set proposed by Floridi and Cowls [2], and expand their frameworks until we reach the minimal level of complexity, suitable to capture the directives and recommendations we have found in the twenty guidelines analysed here.

While the first four principles of Floridi and Cowls correspond to the principles of bioethics and with the main axiological action types we found in our analysis, we disagree that explicability should be the fifth, fundamental additional principle for AI in society. We object to their proposal of using a single value term (explicability) to connect "the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?')" and "the ethical sense of 'accountability' (as an answer to the question 'who is responsible for the way it works?')" [2]. These are not merely distinct questions, but also substantively different ones. We propose to distinguish these two aspects: the epistemic aspect of explicability is an aspect of *transparency* and the ethical aspect refers to a distinct value: *accountability*. This treats transparency, which is the "the most prevalent principle in the current literature" [1, p. 391], as a *distinct* procedural value. Thus, we conclude that the fifth principle of Floridi and Cowls relates to two distinct procedural value-clusters: transparency and accountability. Finally, since not *all* forms of *control* – an action type found in guidelines – are instrumental to transparency or accountability, we introduce *Control* as a distinct procedural principle.

We therefore propose replacing Floridi and Cowls' fifth principle (explicability), with 3 distinct principles:

5. Control principle: you shall control the entities, goals, process, and outcomes of data-driven services affecting people, and generate the knowledge necessary for such control.

6. Transparency principle: you shall communicate your knowledge of the entities, goals, process, and outcomes of data-driven services affecting people, in an adequate and effective way, to the relevant stakeholders.

7. Accountability principle: you shall assign moral, legal, and organizational responsibilities to the individuals who control the entities, goals, process, and outcomes of data-driven services affecting people.

We argue that not only the value clusters of Floridi and Cowls are covered by this 7 value framework, but that also all values in the 11 value list of [1] can be mapped into this framework. There is still some non-congruence, e.g. with respect to the trust principle. But it is not clear that "promote trust" is an ethical principle, because trust contributes to well-being only when it is well-placed (placed on trustworthy entities). More sensibly, what really matters is to promote well-being, individually and collectively, i.e. the principle beneficence. We place sustainability in non-maleficence cluster. Privacy can be associated with equal merit to both the non-maleficence and autonomy cluster (e.g. depending on whether one emphasizes harmful access to the self, or control of information). We also include dignity and freedom (including liberty rights) in autonomy. Justice includes fairness, non-discrimination, and solidarity. Responsibility characterizes accountability.

From Floridi and Cowls we retain the idea that the fifth principle plays a different role from the first four. This is also true for the three principles replacing it in our framework (control, transparency, and accountability). The best way to understand the principles of control, transparency, and accountability is not as indicating ends that, when realized, make the world an intrinsically morally better place. They are, rather, enablers of the other values. Second, they can be seen as deontological principles, not as teleological ones. Teleological principles characterize ethical action in terms of its *goals*. Deontological principles involve duties to perform actions for reasons *other* than their being instrumental to specific goals [60], [61]. In deontological morality, control,

transparency, and accountability are best conceived as adverbs, rather than nouns. The principles 5-7 can be expressed, in their adverbial form, as follows: whatever your goal, act, (5) in control, (6) transparently, and (7) accountably.

Thus, by extending the framework of Floridi and Cowl to distinguish the different aspects of what they call "explainability". After subsuming the values inductively discovered by [1] into such set, we obtain a framework consisting of the following seven principles:

1. Beneficence: do the good (promote individual and community well-being, and preserve trust in trustworthy agents).
2. Non-maleficence: avoid harm (also by protecting security, privacy, dignity, and sustainability).
3. Autonomy: promote the capabilities of individuals and groups (also by protecting civic and political freedoms, privacy, and dignity).
4. Justice: be fair, avoid discrimination, promote social justice and solidarity
5. Control: knowledgeably control entities, goals, process, and outcomes affecting people.
6. Transparency: communicate your knowledge of entities, goals, process, and outcomes, in an adequate and effective way, to the relevant stakeholders.
7. Accountability: assign moral, legal, and organizational responsibilities to the individuals who control entities, goals, process, and outcomes affecting people.

In 5, 6, 7 "entities, goals, processes, and outcomes" refers to elements, goals, processes and outcomes of AIs and other data-driven products.

VI. Conclusions

We performed an in-depth analysis of 20 guidelines for ethical usage of AI and big data, focusing on the kind of activities prescribed by these guidelines. As common elements, we found three distinct procedural principles (control, transparency, and accountability), in addition to the promotion of four value clusters (beneficience, non-maleficience, autonomy, and justice). The ethical usage of AI and big data should promote the values indicated by the traditional four-value clusters, while respecting the three procedural/deontological principles indicated here.

Conceptually, the distinction of procedural principles and value-oriented principles helps to clarify the role of specific recommendations appearing in the different guidelines. We have argued that, in ethical terms, the procedural principles may be understood as deontological principles, while the value clusters describe teleological principles. As the analysed guidelines feature prescriptions of either type, both seem necessary when providing ethical guidelines for AI.

Our framework is closely related to the one by Floridi and Cowl while decomposing "explainability" into transparency and accountability, and introducing control as an additional, irreducible procedural value. Furthermore, the framework can be used to simplify and shorten the 11-value list of Jobin et al [1].

We found that our set of seven principles is extremely useful as a tool to map and compare the recommendations of different codes and guidelines. The framework covers most of the recommendations found in the analysed guidelines.

The framework may also be used to improve existing guidelines because it allows detecting uncovered areas where recommendations might be missing. As our framework has been developed by systematically identifying the common core of recommendations of codes and guidelines for AI, one might use it to systematically check the completeness of a given code or guideline. Our framework allows for an ever more specific analysis: since each of the three procedural principles can be applied to all activities of the four value clusters (the principles of control, transparency or accountability can be applied to each activity intended to promote, e.g., autonomy), one might check whether all 4x3 *combinations* of procedural principles and value clusters are covered by a given code or guideline.

The authors have recently applied this methodology during the development of an ethical code for data-based business, intended to support companies in developing data-based products and services with high ethical standards [64]. The systematic approach proved to be a great help in finding weak spots and missing elements of this code.

Acknowledgment

This work was partly financed by Innosuisse, under its NTN grant for the Swiss Alliance for Data-Intensive Services. Preliminary results have also been used for an ethical analysis of people analytics, not in a peer-reviewed publication, for a research project funded by Algorithmwatch and the Hans Böckler Stiftung [65].

References

[1] A. Jobin, M. Ienca, and E. Vayena, "Artificial Intelligence: the global landscape of ethics guidelines," Nat. Mach. Intell., vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[2] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," Jun. 2019, doi: 10.1162/99608f92.8cd550d1.

[3] European Group on Ethics in Science and New Technologies, "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems," European Commission Directorate-General for Research and Innovation, Mar. 2018.

[4] Future of Life Institute, "Asilomar AI principles." .

[5] House of Lords - Artificial Intelligence Committee, "AI in the UK: ready, willing and able?" [Online]. Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm. [Accessed: 12-Feb-2020].

[6] Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically aligned design. A vision for prioritizing human wellbeing with autonomous and intelligent systems, version 2." 12-Dec-2017.

[7] Partnership on AI, "Tenets." 26-Sep-2016.

[8] Université de Montreal, "Montreal Declaration for a responsible development of AI," 2017. [Online]. Available: https://www.montrealdeclaration-responsibleai.com/the-declaration. [Accessed: 29-Jan-2019].

[9] T. L. Beauchamp and J. F. Childress, Principles of Biomedical Ethics, 6. ed. New York: Oxford University Press, 2008.

[10] T. Hagendorff, "The Ethics of AI Ethics -- An evaluation of guidelines," Minds Mach., Feb. 2020, doi: 10.1007/s11023-020-09517-8.

[11] Software & Information Industry Association (SIIA), Public Policy Division, "Ethical principles for artificial intelligence and data analytics." 15-Sep-2017.

[12] Internet Society, "Artificial intelligence and Machine Learning: Policy Paper." 18-Apr-2017.

[13] WEF, Global Future Council on Human Rights 2016-2018, "White Paper: How to prevent discriminatory outcomes in machine learning." 12-Mar-2018.

[14] Independent High-Level Expert Group On Artificial Intelligence Set Up By The European Commission, “Ethics guidelines for trustworthy AI.” European Commission - Digital Single Market, 08-Apr-2019.

[15] Women leading in AI, “10 Principles of responsible AI.” .

[16] Fairness, Accountability, and Transparency in Machine Learning (FATML), “Principles for accountable algorithms and a social impact statement for algorithms.” 26-Nov-2016.

[17] UNI Global Union, “Top 10 principles for ethical artificial intelligence.” 17-Dec-2017.

[18] Access Now; Amnesty International, “The Toronto Declaration: Protecting the right to equality and nondiscrimination in machine learning systems.” 16-May-2018.

[19] ICDPPC, “Declaration on ethics and data protection in Artificial Intelligence.” 23-Oct-2018.

[20] Mission Villani, “For a meaningful Artificial Intelligence. Towards a French and European strategy.” 29-Mar-2018.

[21] M. Loi, “Technological unemployment and human disenhancement,” Ethics Inf. Technol., vol. 17, no. 65, pp. 1–10, Jul. 2015, doi: 10.1007/s10676-015-9375-8.

[22] The Public Voice, “Universal guidelines for artificial intelligence.” 23-Oct-2018.

[23] Information Technology Industry Council (ITI), “ITI AI Policy Principles.” 24-Oct-2017.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” ArXiv160204938 Cs Stat, Feb. 2016.

[25] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR.(2017),” Harv. J. Law Technol., vol. 31, p. 841, 2017.

[26] COMEST/UNESCO, “Report of COMEST on Robotics Ethics.” 14-Sep-2017.

[27] Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Ethically aligned design. A vision for prioritizing human wellbeing with autonomous and intelligent systems, version 1.” 25-Mar-2019.

[28] Deutsche Telekom, “AI Guidelines.” 11-May-2018.

[29] Leaders of the G7, “Charlevoix Common Vision for the Future of Artificial Intelligence.” 09-Jun-2018.

[30] Future of Humanity Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI, “The malicious use of Artificial Intelligence: Forecasting, prevention, and mitigation.” 20-Feb-2018.

[31] K. J. Arrow, The limits of organization, 1st edition. New York, NY: W. W. Norton & Company, 1974.

[32] R. Hardin, “Trustworthiness,” Ethics, vol. 107, no. 1, pp. 26–42, 1996.

[33] A. Ferrario, M. Loi, and E. Viganò, “In AI we trust incrementally: a multi-layer model of trust to analyze human-Artificial Intelligence interactions,” Philos. Technol., Oct. 2019, doi: 10.1007/s13347-019-00378-3.

[34] J. S. Mill, On Liberty. London: Penguin Classics, 1859.

[35] E. J. Bloustein, Individual and Group Privacy, 2nd ed. New Brunswick, N.J., U.S.A.: Routledge, 2003.

[36] M. Loi and M. Christen, “Two concepts of group privacy,” Philos. Technol., May 2019, doi: 10.1007/s13347-019-00351-0.

[37] L. Floridi, “Group Privacy: a defence and an interpretation,” in Group Privacy: New Challenges of Data Technologies, L. Taylor, L. Floridi, and B. Van der Sloot, Eds. Cham: Springer, 2016, pp. 83–100.

[38] A. F. Westin, Privacy and Freedom, 1st ed. New York: Atheneum, 1967.

[39] S. D. Warren and L. D. Brandeis, “The Right to Privacy,” Harv. Law Rev., vol. 4, no. 5, pp. 193–220, 1890, doi: 10.2307/1321160.

[40] R. Gavison, “Privacy and the limits of the law,” in Philosophical Dimensions of Privacy: An Anthology, F. D. Shoeman, Ed. Cambridge MA: Cambridge University Press, 1984.

[41] J. Reiman, “Privacy, intimacy and personhood,” Philos. Public Aff., vol. 6, no. 1, pp. 26–44, 1976.

[42] Privacy International & Article 19, “Privacy and freedom of expression in the age of Artificial Intelligence.” 25-Apr-2018.

[43] S. Wachter, “Affinity profiling and discrimination by association in online behavioural advertising,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3388639, May 2019.

[44] AI4People, “Ethical framework for a good AI society.” Atomium European Institute for Science, Media, and Democracy, 2019.

[45] W20, “Artificial Intelligence: open questions about gender inclusion.” 02-Jul-2018.

[46] J. Griffin, On Human Rights. Oxford: Oxford University Press, 2008.

[47] “International Covenant on Civil and Political Rights.” [Online]. Available: http://www2.ohchr.org/english/law/ccpr.htm. [Accessed: 22-Mar-2012].

[48] T. M. Mitchell, Machine Learning, 1 edition. McGraw-Hill Education, 1997.

[49] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, “On the (im)possibility of fairness,” ArXiv160907236 Cs Stat, Sep. 2016.

[50] J. Sánchez-Monedero, L. Dencik, and L. Edwards, “What does it mean to ‘solve’ the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 2020, pp. 458–468, doi: 10.1145/3351095.3372849.

[51] I. D. Raji et al., “Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.

[52] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, “How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness,” in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 2019, pp. 99–106, doi: 10.1145/3306618.3314248.

[53] A. Lundgard, “Measuring justice in machine learning,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 2020, p. 680, doi: 10.1145/3351095.3372838.

[54] J. Herington, “Measuring fairness in an unfair World,” in Proceedings of 2020 AAAI-ACM Conference on Artificial Intelligence, Ethics, and Society, New York, NY, USA, 2020, doi: 10.1145/3375627.3375854.

[55] H. Heidari, M. Loi, K. P. Gummadi, and A. Krause, “A moral framework for understanding fair ML through economic models of equality of opportunity,” in Proceedings of the Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 2019, pp. 181–190, doi: 10.1145/3287560.3287584.

[56] R. D. P. Binns, “Fairness in machine learning: Lessons from political philosophy,” J. Mach. Learn. Res., 2018.

[58] G. A. Cohen, Rescuing Justice and Equality. Harvard University Press, 2008.

[59] J. Rawls, A Theory of Justice, 2nd ed. Cambridge, MA: Harvard University Press, 1999.

[60] I. Kant, Groundwork for the Metaphysics of Morals. Yale University Press, 2002.

[61] J. L’Etang, “A Kantian approach to codes of ethics,” J. Bus. Ethics, vol. 11, no. 10, 1992.

[62] M. Loi, M. Christen, N. Kleine, and K. Weber, “Cybersecurity in health – disentangling value tensions,” J. Inf. Commun. Ethics Soc., May 2019, doi: 10.1108/JICES-12-2018-0095.

[63] M. Brundage et al., “The malicious use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” ArXiv180207228 Cs, Feb. 2018.

[64] M. Loi, C. Heitz, A. Ferrario, A. Schmid and M. Christen, "Towards an Ethical Code for Data-Based Business," 2019 6th Swiss Conference on Data Science (SDS), Bern, Switzerland, 2019, pp. 6-12, doi: 10.1109/SDS.2019.00-15.

[65] M. Loi, “People Analytics must benefit the people. An ethical analysis of data-driven algorithmic systems in human resources management,” Algorithmwatch, Mar. 2020. Available: https://algorithmwatch.org/wp-content/uploads/2020/03/AlgorithmWatch_AutoHR_Study_Ethics_Loi_2020.pdf.