VICESSE Working Paper
November 2021

Overview of AI Fairness and Ethics Guidelines

Victoria Kontrus

Abstract
Artificial intelligence (AI) systems have pervaded almost every aspect of our daily lives and the effects of this development are tangible. To address the ethical risks such ground-breaking technologies typically entail, a rapidly growing number of fairness and ethics guidelines have been published in recent years. This paper provides an exemplary overview of the diverse landscape of AI ethics guidelines and comparative analyses in secondary literature. It examines this recent surge of guidelines, characterizes and categorizes various sets of guidelines and discusses popular principles contained therein with an emphasis on fairness and explainability. It also points to a series of challenges that arise from the principled approach to AI ethics, such as virtue signaling and especially practical implementation.

Introduction

The development of AI systems, especially in the domains of machine learning and deep learning, has seen immense progress throughout recent years [1], [2]. Due to this significant increase and diversification of capabilities, AI can now be used to perform an impressive variety of increasingly complex tasks in an ever-growing number of fields of application. Examples include social media, healthcare, finance, education, work, security or transportation to name only a few [3], [4]. Although it is not always evident or we are not actively conscious of it, it is safe to say that by now AI systems have pervaded almost every aspect of our everyday lives [5], [6]. As a result of this proliferation, AI is said to be one of the most impactful technologies of the present and future [7], with the potential to revolutionize the foundations of human life as a whole [8].

As is usually the case with such groundbreaking technologies, advancements entail major challenges and severe risks that need to be anticipated, surmounted, managed and mitigated appropriately. One of these core concerns relating to AI is ethics. While the discussion around AI ethics originated in the 1960s, shortly after the term "AI" had been coined in 1956 [6], it has by now transcended academic circles [5] and receives broad media coverage as well as attention from a variety of stakeholders such as policymakers, industry and the non-profit sector [1]. This heightened awareness among the public mainly stems from the fact that – as a result of AI proliferation – people have begun to really feel the impact of AI in their lives, be it in filter bubbles, their credit score or predictive policing [6], [9].

A key theme in the broader AI ethics discussion is fairness. Sometimes fairness is mentioned as one of several ethical principles meaning fairness is seen a sub-category of ethics (e.g. see [10], [11]) and at other times fairness and ethicality are treated as separate categories (e.g. in [7], [12]). This means that the precise relation of the two categories in this context remains in the dark. This incongruency is also symptomatic of the vagueness inherent to many of the principles set forth in such guidelines. Nonetheless, the main objective is very straightforward: how to ensure that AI systems act ethically and fairly towards individuals, groups and society. Both the number and types of efforts aiming to contribute to this question are vast. They often come in the form of guidelines or principles [1] meant to serve as conducive directive for fair and ethical AI.

Before presenting our own guidelines for the development of fair AI systems, this chapter will provide an overview of existent sets of principles and guidelines for fair and ethical AI. This overview is by no means exhaustive as such an undertaking would go far beyond the scope of this paper. Instead, it intends to map an exemplary outline of the landscape of guidelines, the purpose of which to highlight its diversity as well as address a series of pertinent issues. This chapter will first examine the recent popularity of fair AI guidelines and discuss its potentials and drawbacks. This is followed by a brief overview of characteristics frequently used to categorize and distinguish guidelines. The next section focuses on the principles contained in such guidelines, their structure and interrelations between sets of guidelines. Finally, the issues of fairness and explainability will be discussed in more detail.

The rise of fair AI guidelines as a recent phenomenon

With the proliferation of AI came a proverbial flood of guidance documents on fair AI. The terminology used varies and documents that contain this sort of guidance may be called guidelines [10], [13], principles [8], [14], declaration [15], [16], charter [17], [18] or code[19]. While they all generally aim to ensure that AI systems adhere to and realize ethical principles, they place their emphasis in different areas such as fairness, trustworthiness, responsibility or reliability, which are reflected in their respective titles. This chapter will use the terms "guidelines", "principles" or "framework" to refer to all of these documents, regardless of their exact names.

It seems that this upsurge of fair AI guidelines is a very recent phenomenon that emerged in the later 2010s. The AI Ethics Guidelines Global Inventory [20] counted a total of 173 documents as of April 2020, and around 160 of them were published in 2018-19. Jobin et al. [21] reviewed 84 documents

and state that more than 70 of them were published after 2016. Ryan & Stahl [1] reviewed even more. All of these sources can be seen as examples of efforts to keep track of the fast-growing number of guidelines, and it has become increasingly difficult to keep an accurate and complete record. Fair AI guidelines have also received considerable attention in academic literature (for comprehensive analyses see e.g. [1], [21], [22]; for more selective analyses see e.g. [3], [6], [23], [24]) and even the upsurge of guidelines itself has been named. Mittelstadt [25, p. 2] uses the term "principlism" to refer to the popular attempt of issuing principles in order to make AI (more) ethical. Floridi & Cowls [6, p. 2] call this phenomenon the "problem of principle proliferation" which places an emphasis on the downside of this development. The rise of fair AI guidelines is certainly connected to the realization that AI *is* not fair but needs to be *made* so. However, apart from honest attempts to improve AI from an ethical perspective, it must not be forgotten that different stakeholders may also use the tool of fair AI guidelines to cater to their own interests, such as shaping AI narratives and discourses in a way that benefits them [7].

Similar to the proliferation of AI systems, the proliferation of guidelines also has its advantages and drawbacks. On the upside, the mere publication of guidelines already helps raise further awareness of the need to ensure that AI contributes to the value of fairness in all stages of its lifecycle [25]. Awareness is often a conducive first step towards practical implementation, as the public may exert pressure both on policy-makers and industry. Moreover, guidelines clarify (to some extent) what is meant by "fair AI" which "can help focus public debate" [25, p. 1]. This creates leverage, as a very abstract principle is translated into several more specific objectives, which are harder to evade. Although most of the present guidelines are not legally binding and/or lack enforcement mechanisms [1] stakeholders can subject themselves voluntarily, which may also have a positive impact. In the age of social media, many companies have come to the conclusion that it is inopportune to displease the internet [5] and substantive accusations of "bluewashing" – a term generally used to describe deceitful PR practices used to create the image of a socially responsible company – may significantly damage a reputation. "Ethics bluewashing", a similar term, as defined by Floridi [26, p. 187] refers to making oneself "appear more digitally ethical than one is" and was specifically coined for digital spaces and technologies.

These potential advantages of principle proliferation are complemented by a series of drawbacks. Although voluntary commitment to guidelines may contribute to their implementation, it often results in nothing more than virtue signaling [25]. While it has been argued that principles help focus discussions around specific issues, the opposite may be true as well: they appear to concretize concepts, but may in fact simply replace one abstract notion with four or five others, merely creating an appearance of specification [25]. This seems to pose a problem, as principles have frequently been criticized as too vague and thus impractical [1], [7]. Moreover, such instruments of self-regulation provide legislators and other entities with binding regulative powers with an excuse to avoid this sensitive issue [25]. In conclusion, they create a sense of achievement which tempts to rest on one's laurels, but in reality they cloud the fact that little to no progress has been made.

Distinguishing features to categorize guidelines by

All fair AI guidelines are not created equal, although they appear to be similar both in their purpose and their content. They differ from each other in a number of aspects which shall be discussed below. A key distinction lies in who issued the guidelines, which is generally wise to bear in mind when studying them, as different issuers are motivated by different intentions. While this distinction is so important that it is made by a number of papers (e.g. [21]–[23]) there seems to be no consensus on which and how many issuer categories there are exactly. This may be a result of the immense variety of initiatives that produce such guidelines. The categorization by issuer is further complicated by the fact that some guidelines are filed under different categories in different publications. The distinction presented in the following is a rough overview of the most common categories that draws on publications cited in this chapter as well as the AI Ethics Guidelines Global Inventory [20]. The most

commonly found issuer category in literature is “governments” and it also shows the highest degree of homogeneity of categorization across different publications, presumably because governments are clearly identifiable actors. This category mostly refers to national governments or authorities [11], [27], [28], but sometimes also includes EU institutions [10], [18], [29] or inter-governmental/international organizations [30]–[32]. Another important category is “non-governmental/ NGO” that greatly overlaps with “non-profit” and “civil society” [8], [14], [15]. Guidelines are also published by the “private” or “corporate” sector [33]–[37], sometimes referred to as “industry”. Due to obvious vested interests documents in this category do not enjoy the best reputation. Guidelines that fall on the technical side are published by professional organizations [38], industry associations [19] and certification institutions [2]. Similarly, standards organizations (e.g. ISO and IEC) are currently developing standards for fair [39] and ethical [40] AI. Although these issuers are closely linked to “private/corporate/industry”, the types of documents produced are quite distinct. Finally, there is “academia/research” [12], [16] and the residual category “multistakeholder / mixed cooperation / other” [2], [41] for all those initiatives that cannot be placed within one of the above categories.

Fair AI guidelines can also be distinguished by their addressees. However, the value of this distinction is dubitable, as most currently available documents are directed at a very broad audience or even “anyone” [16, p. 6]. This may be interpreted as inclusivity as well as an attempt to reach anyone who is capable of and willing to put principles into practice. The breadth of audience is a main point of criticism such guidelines are faced with [1]. With regard to successful implementation, it is often beneficial to be as specific as possible in terms of which tasks are assigned to whom, or which stakeholder is responsible to ensure AI systems comply with a certain requirement. Lack of specificity may result in no one feeling responsible. The Malta Framework [11] serves as a positive example here. Moreover, the practices and areas of competency among AI stakeholders differ significantly [25]. Guidelines that address everyone therefore often have to remain vague, which complicates their implementation. Frequently mentioned addressees include AI practitioners and industry (e.g. software developers and operators, computer scientists), AI researchers from fields other than IT (e.g. social sciences, data science, law, ethics), policy-makers and regulators, watchdog organizations and interested members of the general public. Our own guidelines are directed at developers and designers of AI systems.

Another, related distinction differentiates between the stages of an AI system’s lifecycle. While it has been argued that broad stakeholder participation should occur throughout an AI system’s lifecycle [42], the reality is that different stages are usually dominated by different stakeholder groups. Guidelines for AI certification naturally focus on the stages that precede deployment [2], but such limitations of scope are rare. Instead, many guidelines emphasize that the principles are to be followed throughout the entire lifecycle of an AI system [10], [11]. Our own fair AI guidelines focus on the initial development stage, but cover the entire lifecycle of an AI system as they are “ontogenetic in nature” [4, p. 18] meaning they are always in (re-)development and constantly evolve and adapt [10].

Fair AI guidelines can furthermore be distinguished by their geographical scope, by their degree of commitment (ranging from mere words to legally enforceable [43]) and whether they are directed at oneself (e.g. corporate value statements [33]–[37]) or at others. It has been noted that differences also lie in length, tone, topic emphasis and level of technicality [1].

Principles commonly found in guidelines

The rise of fair AI guidelines has made it increasingly difficult to keep track of their content. This is why, as indicated above, many authors have undertaken comparative studies of guidelines. These studies have found a high degree of convergence between guidelines [1], [6], [21], which does not come as a surprise, as many guidelines are explicitly based on one another, e.g. the Malta Framework [11] is based on the EU Guidelines for Trustworthy AI [10] and the G20 statement [30] builds on the OECD guidelines [31]. Moreover, guidelines typically endorse concepts of rather abstract nature. The more

abstract a concept, the more likely it is to overlap with closely related notions, as proper definition and thus delimitation of such "essentially contested concepts" [25, p. 5] is practically infeasible.

Although the convergence may be read as a promising foundation for further efforts, this seeming "consensus" is not to be overestimated, as conflicts will most certainly arise, the more such efforts advance towards substantiation [7]. In light of their overlap and vagueness, many guidelines have also been criticized as reiterative and confusing [6]. It has instead been proposed to direct efforts more specifically towards "translating" principles into practice [7], [10], [25], as even stakeholders willing to follow them experience difficulty. However, this high degree of convergence also entails the benefit of preventing "cherry picking". If guidelines contain largely the same principles, stakeholders committing to them have little choice. Otherwise, they could simply pick and choose whichever principles align best with their purpose [6].

Many guidelines follow a similar structure, but obviously, differences can be found especially between academic and private sector publications, as the latter are not bound by standards of scientific practice and seem to focus more on visual design. An introductory section typically lays out purpose and target audience of the guidelines, explains their development process and who was involved in it. Some also provide instructions on how to read and interpret principles. Lengthy publications often contain an executive summary. The main body contains the principles and there are different ways to present them. Very short publications [8], [37] simply list them and provide one or two explanatory sentences each. Longer publications [16], [34], [35] devote a section or chapter – often a page – to each principle and give more detailed explanations, specifications and even examples of what is meant. More comprehensive guidelines [10], [11] typically progress from abstract to concrete and include some sort of foundation from which the principles are deduced at the outset. Subcategories with more detailed explanations are presented for each principle, which are further translated into requirements and sub-requirements. The most meticulous guidelines [12], [38] even provide resources and guidance on how to fulfill these requirements. Some guidelines also include their own glossary [16], [18], [38], [44], whereas others refer to the glossaries contained in other documents[2], [32]. These glossaries often explain terms from the field of IT (e.g. explainability, machine learning, generative adversarial network) but also outside of it (e.g. reliability, sustainability or fairness).

Before discussing the issues of fairness and explainability in more detail, a short overview of popular principles contained in such guidelines will be given. The number of principles set forth in guidelines varies and is not too informative, as it depends on the level of abstraction. Presumably for the purpose of clarity, many guidelines contain around four to ten top-level principles, and comparative studies of secondary literature typically find roughly around five [6], [7] to eleven [1], [21]. This discrepancy results from different terminology and levels of abstraction being used as well as from merging and rearranging top- and mid-level principles (the latter may also be called requirements or practices). Secondary literature often uses tables to illustrate these interrelations [1, p. 65], [3, p. 392], [6, p. 9], [7, p. 2146], [21, p. 7]. The following overview of principles (Table 1) is based on the two most comprehensive, recent contributions of secondary literature, both of which have analyzed more than 80 guidelines [1], [21]. Arranged in descending order according to their frequency, the eleven top-level principles these publications have identified are: transparency, justice & fairness, non-maleficence, responsibility, privacy, beneficence, freedom & autonomy, trust, sustainability, dignity and solidarity.

Ethical principle	**Codes**
Transparency	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & fairness	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution

Non-maleficence	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	Responsibility, accountability, liability, acting with integrity
Privacy	Privacy, personal or private information
Beneficence	Benefits, beneficence, well-being, peace, social good, common good
Freedom & Autonomy	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	Trust
Sustainability	Sustainability, environment (nature), energy, resources (energy)
Dignity	Dignity
Solidarity	Solidarity, social security, cohesion

Table 1. Ethical principles and associated codes (adapted from [1], [21])

As can be seen in Table 1, each top-level principle is associated with one or more codes. In some cases, various codes show mere incongruency in terminology with little to no difference in meaning (e.g. understandability and interpretability under transparency) whereas in other cases, the codes refer to lower-level principles that differ in meaning both from each other as well as from the top-level principle they are filed under (e.g. diversity and redress under justice & fairness). It can also be seen that principles that occur in most publications are represented by more codes than those that occur less frequently. While the difference in number of associated codes surely is a result of principle frequency, it may also be that the number of codes is influenced by the amount of controversy that surrounds the respective principles.

A coarser guideline analysis condenses their content down to five main principles, four of which are common in bioethics (beneficence, non-maleficence, autonomy and justice), whereas the fifth (explainability) is a novelty specific to the field of AI [6]. Using this set of principles, Morley et al. [7] have devised a typology of publicly available AI ethics tools, methods and research [45] structured into different stages of an AI system's lifecycle.

The interrelation of these principles is not conceptualized equally in all guidelines. This is reflected in different approaches to interpreting principles, which some guidelines provide. Many guidelines view conflicts between principles as inevitable and some therefore provide guidance on how to handle such cases appropriately [10]–[12]. The European High-Level Expert Group on AI proposes to "identify, evaluate, document and communicate" [10, p. 24] trade-offs that have to be made as a result of tensions between principles. Proper reasoning is crucial with regard to accountability and in extreme cases – where no acceptable trade-off can be found – abstention from further development or deployment is in order. In contrast, the Montréal Declaration [16] orders its readers to interpret the principles in such a way they do not conflict, as conflicts between principles result from misinterpretation of their respective scopes of application, which is to be avoided. Unfortunately, no further explanation or example is provided, which gives rise to confusion.

Fairness

The vast majority of guidelines contain the principle of fairness. Fairness is often linked to justice and (based on Jobin et al [21]) it is the principle that has the greatest variety of associated codes and subcategories, which also include consistency, inclusion, equality, equity, (non-)bias, (non-) discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution. Being an "essentially contested concept" [25] p. 5) this multitude of terms does not come as a surprise. The following account highlights different approaches to fairness found in the various guidelines.

Many guidelines simply mention the term without further ado [2], [30], [31], whereas others provide examples or short explanations of what is considered fair or not [19], [33], [36], often using one or

more of the codes listed above. One set of guidelines that explicitly defines the concept of fairness in its glossary is the Assessment List for Trustworthy AI compiled by the High-Level Expert Group on Trustworthy AI: "Fairness refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally referred to as 'substantive' fairness. But fairness also encompasses a procedural perspective, that is the ability to seek and obtain relief when individual rights and freedoms are violated." [44, p. 27]. Both the Assessment List [44] and the corresponding, preceding Ethics Guidelines[10] lay out three main components of fairness: avoidance of unfair bias (including prejudice, marginalization, exploitation, incompleteness), accessibility & universal design and stakeholder participation (including longer term mechanisms). Notably, this definition includes a few codes not listed above.

A very detailed chapter on fairness can be found in the guidelines introduced by The Alan Turing Institute [12]. As there are many different conceptions of fairness, the guidelines refrain from defining it and advise to use the "principle of discriminatory non-harm" as a minimum […] threshold of fairness" instead. These guidelines break the concept of fairness down into four separate dimensions: Data Fairness, Design Fairness, Outcome Fairness and Implementation Fairness. Data Fairness means that AI systems must be "trained and tested on properly representative, relevant, accurate, and generalizable datasets". To achieve Design Fairness, AI systems must "have model architectures that do not include target variables, features, processes, or analytical structures (correlations, interactions, and inferences) which are unreasonable, morally objectionable, or unjustifiable"[12, p. 14]. Outcome fairness suggests that they "do not have discriminatory or inequitable impacts on the lives of the people they affect" and Implementation Fairness means that they should be "deployed by users sufficiently trained to implement them responsibly and without bias" [12, p. 14]. For each of these dimensions, the guidelines also contain sub-dimensions with explanations and/or definitions. The IEEE guidelines [38] take a very similar approach in that they expressly do not provide a definition of fairness, as no one definition can be suitable for all purposes. Instead, they recommend to use the principles of effectiveness, competence, accountability and transparency as measures for any criteria of fairness. They generally advise to use specific norms as opposed to abstract values such as fairness.

Yet another way to address the concept of fairness is to actively invite the readers and users of guidelines to find their own working definitions. This approach [10], [11] aims to assure that a definition used is adequate for its purpose and domain. Such a definition should be chosen after careful consideration of several alternatives and undergo a quantitative analysis. Moreover, it should be commonly used and accepted among impacted communities.

Explainability

Explainability is either treated as a principle in and of itself or filed under transparency. The latter seems to be the most popular principle in AI ethics guidelines [21]. This may be due to the fact that explainability is both a particularity of the field of AI [6] and a research field that has received considerable academic attention in recent years. Codes typically associated with transparency are explainability, explicability, understandability, interpretability, communication, disclosure and showing [21]. Other related codes include traceability, auditability [10], accountability [31] – although the latter is also frequently listed as an independent principle.

As is the case with fairness, some guidelines use the term without providing a definition [30], [46]. Some of them give examples [15], [33], but explainability and transparency are rarely delimited from each other in these. However, explainability is not a moral principle [7] (such as fairness), which means that in general there seems to be less hesitation to define it and no express advice against it was found. Nonetheless, it has been noted that the precise terminology is not fully consistent [47]. The glossary in the Assessment List for Trustworthy AI defines explainability as follows: "Feature of an AI system that is intelligible to non-experts. An AI system is intelligible if its functionality and operations can be

explained non technically to a person not skilled in the art." [44, p. 26]. The Malta Framework also specifies its sub-requirement explainability: "Ensure that end-users and other affected individuals can understand the operation of the AI system."[11, p. 27]. A definition of explainability can furthermore be found in the Singapore Framework: "ensure that automated and algorithmic decisions and any associated data driving those decisions can be explained to end-users and other stakeholders in non-technical terms" [48, p. 64]. A definition of intelligibility is given in the Montréal Declaration: "An AIS [Artificial Intelligence System] is intelligible when a human being with the necessary knowledge can understand its operations, meaning its mathematical model and the processes that determine it." [16, p. 19].

All the above definitions largely agree that explainability/intelligibility means that the workings of an AI system should be of such quality that they can be understood and explained. However, they disagree on the level of knowledge needed to understand such explanations. The first three set the bar low using the terms "non-experts", "end-users" and "other stakeholders". One even specifies the language to be used in such explanations, stating explanations must be given in "non-technical terms". The last definition clearly sets the bar higher as it requires "necessary knowledge" to understand the "mathematical model" of the AI system, which is most likely not to be found in every end-user, non-expert or other stakeholder. Due to inconsistent terminology being used in this field, it is unclear whether this disparity of target audience is a result of different opinions on one subject (i.e. explainability/intelligibility) or rather implies that explainability and intelligibility are two distinct requirements meant for different audiences.

Apart from being a specialty in the field of AI ethics [6], explainability also stands out for another reason. It has been ascribed a complementary function [6], [10] and called a "second order principle" [7, p. 2155]: Explainability is said to be necessary to implement other AI ethics principles. While some hold that explainability is key for the implementation of *all* other ethical principles [6], [7], the European High-Level Expert Group [10] links explainability most closely to the principle of fairness, both in its substantive and procedural dimension. Without knowing which data, features or operations led to a certain output, it may be impossible to assess whether the output produced is fair or not. Furthermore, such information is also necessary to hold people accountable in cases found to be unfair.

Despite this consensus that explainability is a key requirement for implementation of other principles, the extent to which this is feasible and how is surrounded by controversy. The most sophisticated methods such as neural networks, ensemble methods (e.g. random forests) or support vector machines are often referred to as "black boxes" as their "innerworkings and rationale are opaque or inaccessible to human understanding." [12, p. 46]. For AI systems using algorithms of this kind it has been suggested to rely on "traceability, auditability and transparent communication on system capabilities" [10, p. 13] as substitute measures. Contrarily, other guidelines argue that neither machine learning nor deep learning systems are real black boxes, as humans are in fact able to see inside but the processes are too complex to be understood by the human mind [2]. The difference here lies in why explainability is deemed infeasible with different resulting implications for future feasibility. The limitations of the human mind are a rather permanent obstacle, which is mentioned in both of the above conceptions of a black box algorithm. However, the first one also refers to system opacity, which may have a greater chance of being overcome with new methods being invented in this rapidly developing field. This progress is reflected in the abovementioned typology of AI ethics tools and methods [45]. Yet, this typology also shows that the great majority of these tools were devised for the later stages of an AI system's lifecycle and provide little guidance for the development stage. Considering the vital role of explainability with regard to other principles, it is evident that further research needs to be devoted to explainable AI methods, focusing especially on the early stages of the development process.

A promising approach to user-centric explainability uses so-called “counterfactuals” [49, p. 844]. Instead of disclosing the complete logic of an algorithm, counterfactuals briefly explain how input data would have to be different to achieve a desired algorithmic output (e.g. “If your annual salary was 5.000€ higher, you would have received the loan.”). This method is convincing not only because it caters to the interests of AI companies (i.e. trade secrets) and users alike, but also because technical methods to generate counterfactuals automatically already exist in principle, although they would have to be adapted for this specific purpose.

From principles to practice: implementable guidelines

Principles are a necessary first step to clarify the “what”. However, the “what” is of little value when no one knows “how”. As has been shown, considerable effort has been invested to answer the first question, but little to the second. Despite establishing ambitious goals and being progressive in thought, many ethical AI principles have received harsh criticism due to lack of impact. Some argue that principles need to become more specific, practical and actionable [1]. Others hold that these principles should be better tailored to both the needs of those who are meant to implement them and different contexts [5]. Still others advocate the development of more and better strategies and well-documented tools ready for use at lower skill-levels instead of further principles [7]. Finally, some others doubt the suitability of the principled approach for the field of ethical AI in general [25].

Ensuring that AI is developed, deployed and operated in a fair and ethical manner is an interdisciplinary task. Communication across disciplinary boundaries requires translation both conceptually and linguistically. Guidelines that aim to answer the question of “how” need to bear this in mind. Building on previous works and our own research, we have developed a set of fair AI guidelines. We aspire to overcome some of the drawbacks previous guidelines suffered from and focus on practical implementability.

References

[1] M. Ryan and B. C. Stahl, "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications," *JICES*, vol. 19, no. 1, pp. 61–86, Mar. 2021, doi: 10.1108/JICES-12-2019-0138.

[2] TÜV AUSTRIA Group and Johannes Kepler University Linz - Institute for Machine Learning, "Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications." 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://www.tuv.at/fileadmin/user_upload/White_Paper_Trusted_Artificial_Intelligence/White_Paper_-_Trusted_Artificial_Intelligence_-_Towards_Certification_of_Machine_Learning_Applications_web_s.pdf

[3] B. Buruk, P. E. Ekmekci, and B. Arda, "A critical perspective on guidelines for responsible and trustworthy artificial intelligence," *Med Health Care and Philos*, vol. 23, no. 3, pp. 387–399, Sep. 2020, doi: 10.1007/s11019-020-09948-1.

[4] R. Kitchin, "Thinking critically about and researching algorithms," *Information, Communication & Society*, vol. 20, no. 1, pp. 14–29, Jan. 2017, doi: 10.1080/1369118X.2016.1154087.

[5] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach, "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk, May 2019, pp. 1–16. doi: 10.1145/3290605.3300830.

[6] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," *Harvard Data Science Review*, Jun. 2019, doi: 10.1162/99608f92.8cd550d1.

[7] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Sci Eng Ethics*, vol. 26, no. 4, pp. 2141–2168, Aug. 2020, doi: 10.1007/s11948-019-00165-5.

[8] Future of Life Institute, "Asilomar AI Principles," 2017. Accessed: Oct. 11, 2021. [Online]. Available: https://futureoflife.org/ai-principles/

[9] F. Eyert, F. Irgmaier, and L. Ulbricht, "Extending the framework of algorithmic regulation. The Uber case," *Regulation & Governance*, p. rego.12371, Nov. 2020, doi: 10.1111/rego.12371.

[10] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI." 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

[11] Malta.AI Taskforce, "Malta towards Trustworthy AI - Malta Ethical AI Framework." 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://malta.ai/wp-content/uploads/2019/08/Malta_Towards_Ethical_and_Trustworthy_AI.pdf

[12] D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector," Zenodo, Jun. 2019. doi: 10.5281/ZENODO.3240529.

[13] Japanese Society for Artificial Intelligence, "The Japanese Society for Artificial Intelligence Ethical Guidelines." 2017. Accessed: Oct. 11, 2021. [Online]. Available: http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf

[14] UNI Global Union, "Top 10 Principles for Ethical Artificial Intelligence." 2017. Accessed: Oct. 11, 2021. [Online]. Available: http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

[15] Amnesty International/Access Now, "The Toronto declaration: protecting the rights to equality and non-discrimination in machine learning systems." 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

[16] Université de Montréal, "Montreal Declaration for a responsible development of Artificial Intelligence." 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://www.montrealdeclaration-responsibleai.com/the-declaration

[17] OpenAI, "OpenAI Charter." 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://openai.com/charter/

[18] European Commission for the Efficiency of Justice (CEPEJ), "European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment." 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

[19] Association for Computing Machinery, "ACM Code of Ethics and Professional Conduct." 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf

[20] AlgorithmWatch, "AI Ethics Guidelines Global Inventory." 2020. Accessed: Oct. 11, 2021. [Online]. Available: https://inventory.algorithmwatch.org/

[21] A. Jobin, M. Ienca, and E. Vayena, "Artificial Intelligence: the global landscape of ethics guidelines," 2019.

[22] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI," 2020.

[23] S. Fukuda-Parr and E. Gibbons, "Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines," *Glob Policy*, vol. 12, no. S6, pp. 32–44, Jul. 2021, doi: 10.1111/1758-5899.12965.

[24] D. Greene, A. L. Hoffmann, and L. Stark, "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning," 2019.

[25] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nat Mach Intell*, vol. 1, no. 11, pp. 501–507, Nov. 2019, doi: 10.1038/s42256-019-0114-4.

[26] L. Floridi, "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical," *Philos. Technol.*, vol. 32, no. 2, pp. 185–193, Jun. 2019, doi: 10.1007/s13347-019-00354-x.

[27] House of Lords, Select Committee on Artificial Intelligence, "AI in the UK: ready, willing and able." 2017. Accessed: Oct. 11, 2021. [Online]. Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

[28] Smart Dubai, “Artificial Intelligence Ethics and Principles, and toolkit for implementation.” 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?sfvrsn=d4184f8d_6
[29] European Group on Ethics in Science and New Technologies to the European Commission, “Statement on artificial intelligence, robotics and ‘autonomous’ systems.” 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1
[30] G20, “G20 Ministerial Statement on Trade and Digital Economy.” 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://www.mofa.go.jp/files/000486596.pdf
[31] OECD, “Recommendation of the Council on Artificial Intelligence.” 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#mainText
[32] Council of Europe, “Artificial Intelligence and Data Protection.” 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://rm.coe.int/2018-lignes-directrices-sur-l-intelligence-artificielle-et-la-protecti/168098e1b7
[33] IBM, “Everyday Ethics for Artificial Intelligence.” 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
[34] SAP AI Ethics Steering Committee, “SAP’s Guiding Principles for Artificial Intelligence.” 2018. [Online]. Available: https://d.dam.sap.com/m/zKaSxze/SAPs%20Guiding%20Principles%20for%20Artificial%20Intelligence.pdf
[35] Deutsche Telekom, “Digitale Ethik KI-Leitlinien.” 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://www.telekom.com/resource/blob/532444/87e1e54df08cce6f4483985bd25250b6/dl-180710-ki-leitlinien-data.pdf
[36] PwC, “A practical guide to Responsible Artificial Intelligence (AI).” 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf
[37] Telia Company, “Guiding Principles on Trusted AI Ethics.” 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf
[38] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition.” 2019. Accessed: Oct. 11, 2021. [Online]. Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf
[39] ISO/IEC JTC 1/SC 42 Artificial intelligence, “ISO/IEC TR 24027 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making.” ISO/IEC JTC 1/SC 42 Artificial intelligence (accessed Oct. 13, 2021).
[40] ISO/IEC JTC 1/SC 42 Artificial intelligence, “ISO/IEC DTR 24368 Information technology — Artificial intelligence — Overview of ethical and societal concerns.” https://www.iso.org/standard/78507.html (accessed Oct. 13, 2021).
[41] Partnership on AI, “Tenets.” 2016. Accessed: Oct. 11, 2021. [Online]. Available: https://partnershiponai.org/about/#tenets
[42] G. Baxter and I. Sommerville, “Socio-technical systems: From design methods to systems engineering,” *Interacting with Computers*, vol. 23, no. 1, pp. 4–17, Jan. 2011, doi: 10.1016/j.intcom.2010.07.003.
[43] European Commission, *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. 2021. Accessed: Oct. 11, 2021. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF
[44] High-Level Expert Group on Artificial Intelligence, “Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment.” 2020. Accessed: Oct. 11, 2021. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342
[45] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, “Applied AI Ethics Typology.” 2020. [Online]. Available: https://docs.google.com/document/d/1h6nK9K7qspG74_HyVlT0Lx97URM0dRoGbJ3ivPxMhaE/edit
[46] Access Now, “Human Rights in the Age of Artificial Intelligence.” 2018. Accessed: Oct. 11, 2021. [Online]. Available: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
[47] A. Rosenfeld and A. Richardson, “Explainability in human–agent systems,” *Auton Agent Multi-Agent Syst*, vol. 33, no. 6, pp. 673–705, Nov. 2019, doi: 10.1007/s10458-019-09408-y.
[48] Infocomm Media Development Authority and Personal Data Protection Commission Singapore, “Model Artificial Intelligence Governance Framework Second Edition.” 2020. Accessed: Oct. 11, 2021. [Online]. Available: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf
[49] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR,” *Harvard journal of law & technology*, vol. 31, no. 2, p. 841, 2018.