IRIE
International Review of Information Ethics Vol. 29 (03/2021)

Catharina Rudschies, Ingrid Schneider, Judith Simon

Value Pluralism in the AI Ethics Debate – Different Actors, Different Priorities

Abstract:
In the current debate on the ethics of Artificial Intelligence (AI) much attention has been paid to find some "common ground" in the numerous AI ethics guidelines. The divergences, however, are equally important as they shed light on the conflicts and controversies that require further debate. This paper analyses the AI ethics landscape with a focus on divergences across actor types (public, expert, and private actors). It finds that the differences in actors' priorities for ethical principles influence the overall outcome of the debate. It shows that determining "minimum requirements" or "primary principles" on the basis of frequency excludes many principles that are subject to controversy, but might still be ethically relevant. The results are discussed in the light of value pluralism, suggesting that the plurality of sets of principles must be acknowledged and can be used to further the debate.

Keywords: Artificial Intelligence, Deliberation, Ethics, Value Pluralism

Outline:

Authors:

Catharina Rudschies:

- Universität Hamburg, Department of Informatics, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany
- ✉ rudschies@informatik.uni-hamburg.de

Prof. Dr. Ingrid Schneider:

- Universität Hamburg, Department of Informatics, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany
- ✉ ingrid.schneider@uni-hamburg.de

Prof. Dr. Judith Simon:

- Universität Hamburg, Department of Informatics, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany
- ✉ simon@informatik.uni-hamburg.de

ISSN 1614-1687

1. Introduction

In reaction to the criticism Artificial Intelligence (AI) technologies have received for issues such as discrimination, opaqueness, or a lack of accountability, a multitude of stakeholders have engaged in a debate on the ethics of AI. As a result, there are now numerous AI ethics guidelines that are to guide the design, deployment and utilisation of the technology. For the effectiveness of such ethical regimes a scattered landscape of guidelines, however, proves problematic, since "it is difficult for individuals involved in the development or use of AI to determine which ethical issues they should be aware of" (Ryan & Stahl, 2).

In order to make sense of the abundance of ethics guidelines, several scholars have tried to synthesise the ethical principles put forward and, thereby, to find some "common ground", "overarching themes", or "minimum requirements" for the ethics of AI (Floridi *et al.*; Hagendorff; Jobin *et al*; Fjeld *et al.*). They have, hence, mainly focused on the convergences in the AI ethics landscape. Most scholars have determined a "core set of principles" on basis of frequency by which ethical principles were mentioned. Interestingly, however, they mostly ended up with different results. While the datasets and methods differed and certainly contributed to the incongruent results, it shows that "core principles" are not easily determined. Furthermore, it shows that the debate around the governance of AI is still in its infancy.

In this paper it is argued that the emphasis on convergences hides the conflicts and controversies that are still existent in the AI ethics debate. It builds upon another analysis recently conducted which found that various stakeholders have varying perceptions of "ethics" and these perceptions translate into diverging formal characteristics of AI ethics guidelines (Rudschies *et al.*, forthcoming). Hence, it is hypothesised here that the diversity in the sets of principles that can be found for the ethics of AI could be the result of diverging priorities of the various stakeholders involved in the AI ethics debate. Consequently, this paper examines AI ethics guidelines with a focus on the divergences of ethical principles put forward by actor types (public, expert, and private actors) in order to find out where conflicts and controversies lie and further discussion is needed. Such endeavour will show in how far the priorities of varying actor groups differ and shape the debate around "core principles" for the ethics of AI. The findings are then discussed in light of value pluralism, suggesting that the plurality of sets of principles must be acknowledged and, more importantly, can be used to further the debate.

2. Shortcomings of the Current AI Ethics Debate

The current state of the AI ethics landscape is marked by high heterogeneity. According to Jobin *et al.* there are at least 84 documents that deal with the ethical design and deployment of AI technologies[1]. Due to this sheer amount of different documents, several scholars have tried to synthesise the ethical principles proposed by the various actors (Floridi; Hagendorff; Jobin *et al.*; Fjeld *et al.*). What is striking though is that despite similar aims to find a "common ground", "overarching themes", or "minimum requirements" for the ethical design and deployment of AI technologies, the reviews come to diverging results. Hagendorff comes to the conclusion that there are three principles (privacy, fairness, and accountability) that can be called the minimum requirements for an ethical AI, for they were mentioned by at least 90 percent of the guidelines he examined. Floridi *et al.* take the bioethical principles nonmaleficence, beneficence, autonomy, and justice as a basis and complements them with the principle of explicability, including intelligibility and accountability. Jobin *et al.* call transparency, privacy, justice and fairness, responsibility, and non-maleficence the "core ethical principles" for AI. And Fjeld *et al.* group the ethical principles under the following themes: transparency and explainability, privacy, nondiscrimination and fairness, safety and security, accountability, human

[1] It should be noted, however, that the documents differ considerably in nature and not all can strictly speaking be called an AI ethics guideline. We explain in the methodology section what we consider to be an AI ethics guideline.

control of technology, and the promotion of human values. While there are some overlaps across all of these sets of principles, it is only one principle which all reviews uniformly declare to be a "core principle" for AI: justice and fairness.

While one of the reasons for these diverging results lies in the variation of datasets and methods employed in the different studies, the findings show that "core principles" or "minimum requirements" are not easily determined. Moreover, what the existing reviews do not sufficiently account for is that the landscape of AI ethics guidelines is an ensemble of documents from a variety of actors, each potentially having their own particular viewpoint on the issue at hand. Jobin *et al*. have found in their analysis at least in a few cases that interpretations of ethical principles were dependent on the actor type that published the guideline at hand. Nonetheless, they focused this (rather limited) analysis merely on those principles where actors converged. While it is certainly important to determine where concurrences already exist, it is the divergences that shed light on the issues of controversy and conflict. Just because stakeholders might not all agree on an ethical principle at hand, it does not necessarily mean that it is ethically irrelevant. For instance, Fjeld *et al.* report:

> *"[…], we were frequently asked why sustainability and environmental responsibility did not appear more prominently. While the authors are sensitive to the significant impact AI is having, and will have, on the environment, we did not find a concentration of related concepts in this area that would rise to the level of a theme, […]."*[2]

Hence, declaring particular principles as the core ethical issues only on basis of the highest frequency might undermine the importance of some ethical considerations that either have not found their way into the mainstream AI ethics debate or do not find consensus (yet) among all stakeholder groups. Examining the divergences of diverse actors can help to shed light on how the actors' diverse preferences shape the debate on the ethics of AI and where further discussion is needed. Fjeld *et al*. indicated that further mapping projects that concentrate on the different perspectives of stakeholder groups would be compelling (66). This paper aims to provide such an inquiry.

3. Methodology

In order to examine how far different actors diverge in their priorities of ethical principles for AI, we conducted a content analysis of 40 different AI ethics guidelines. The term "ethics guideline" is here used in a broader sense, meaning all kinds of documents that present suggestions for the ethical design and deployment of AI technologies. Therefore, we collected a variety of document types, comprising *inter alia* corporate statements, political strategy papers, declarations, White Papers and consultation or discussion papers. For the data collection we used four major databases and sources[3] as well as a web-based search and selected the relevant documents on basis of the purposive sampling method. Eligible for analysis were those documents that were dealing with AI or those technologies that fall under the broader term of AI, such as machine learning, automated decision-making or algorithmic systems with features of AI. We only selected sources from recent years (2016-2019) and included an approximate number of sources for each actor group: 14 sources of public actors, 14 sources of expert actors and 12 sources of private actors. Since we only selected guidelines in English language, our data sample is limited. Furthermore, the Global South is highly underrepresented, which can be explained by our restricted language criteria, but also the overall dominance of sources from Western countries in the AI ethics debate.

[2] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, 2020-1.

[3] *The AI Policy Sourcebook 2019* edited by Marc Rotenberg, 2019; a data collection on AI policy initiatives by the *European Union Agency for Fundamental Rights* (https://fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights/ai-policy-initiatives (last access January 23, 2021)); the *AI Governance Database* provided by *Nesta* (https://www.nesta.org.uk/data-visualisation-and-interactive/ai-governance-database/ (last access 23 Jan. 2021))*;* the AI Principles Map by the *AI Ethics Lab* (http://aiethicslab.com/big-picture/ (last access 23 Jan. 2021))

For the content analysis, all sources were viewed with a focus on which principles they proposed. While the term "principle" is commonly used in the AI ethics debate, the documents examined reveal that there is a variance on what can actually be referred to as an ethical principle. For instance, equality, responsibility, and auditability are mentioned as if they are of the same nature. However, whereas equality constitutes a value, responsibility would account for a principle, and auditability as a requirement that technologies and the people involved in their development and deployment should meet. A distinction between values and principles, on the one side, and requirements, on the other side, would be especially important, since the former are more focused on *what* should be achieved, while the latter are more instrumental in nature and are, hence, focused on the ways and processes *how* this can be done. For purposes of comparability, we nevertheless decided to use the common term "principles". It should, however, be kept in mind that the term is used in reference to values, principles, and requirements alike.

Due to the different nature and form of document types, some ethics guidelines provided a clear list of principles, while others mentioned them more implicitly in the text. We initially categorised the former type as "upper principle" and the latter type as "lower principle". However, since some documents only entailed one form or the other and some included both, it made a comparison across documents rather difficult. Consequently, we accounted both – upper and lower principles – for a "mentioned principle". Principle codes were identified using an inductive approach and entered into a database. All documents were reviewed twice in order to make sure that every code found during the process was identified in all documents alike.

In comparison to existing reviews of AI ethics guidelines, our analysis does not focus on the convergences but more on the divergences in ethical principles across the various sources. Nevertheless, it is important to summarise our findings also in regard to concurrences in order to make a comparison possible, first, between our results and the existing reviews, and second, among the actor types. Hence, we will shortly outline these results and will then concentrate in more detail on the divergences among actors.

4. Ethical Principles – Convergences, Divergences, and Priorities

Overall, the ethics guidelines analysed in this paper mention 49 principles. When examining the congruence of the guidelines, our analysis shows great overlap with the results of the study conducted by Fjeld *et al.* (*see Table 1 for an overview*). Ethics guidelines converge mostly on the following principles: transparency, privacy, responsibility, non-discrimination/fairness, explainability, security/safety, and accountability. Each of these principles are mentioned by at least 75% of the documents. All of them can be found as primary themes in the review by Fjeld *et al*. Two principles (human control and promotion of human values) mentioned by the Harvard scholars are proposed by less than three quarter of the documents analysed in this paper, however, still gain an acceptance rate above 60%. Non-maleficence and beneficence declared as core ethical principles by Floridi *et al*. are suggested in 67.5% of our sources.

Comparing the results of all of the reviews reveals that almost all principles that were identified as "common ground" show high frequency within the AI ethics guidelines of our dataset. Divergences in their ranking are likely to be caused by differences in the dataset, size of the dataset, and the methods used. However, authors may have also opted for different selections (priorities) or simply grouped certain principles under different overall themes. The only exception constitutes the principle of autonomy put forward by Floridi *et al.,* which is – in our dataset – only mentioned in 35% of all documents and is, therefore, not supported to have found convergence in most ethics guidelines. Nevertheless, it is still noticeable that Jobin *et al*., Hagendorff, and Floridi *et al*. have each refrained from mentioning certain principles that are in fact found in a majority of sources examined in this paper (*see Table 1*). Synthesising ethical principles to a minimum number has, hence, the potential to undermine the importance of several ethical issues not captured by such a set.

Our analysis	**Fjeld *et al.***	**Jobin *et al.***	**Hagendorff**	**Floridi *et al.***
transparency (97.5%)	Transparency and explainability	transparency	---	---
privacy (87.5%)	privacy	privacy	privacy	---
non-discrimination and fairness (82.5%)	non-discrimination/fairness	justice and fairness	fairness	(justice)
responsibility (77.5%)	professional responsibility	responsibility	---	---
explainability (77.5%)	transparency and explainability	---	---	explicability (including intelligibility and accountability)
safety and security (75%)	safety and security	---	---	---
accountability (75%)	accountability	---	accountability	(accountability included in explicability)
human control and determinism (70%)	human control of technology	---	---	---
promotion of human values (62.5%)	promotion of human values	---	---	---
non-maleficence (67.5%)	---	non-maleficence	---	non-maleficence
beneficence (67.5%)	---	---	---	beneficence
autonomy (35%)	---	---	---	autonomy

Table 1: Comparison of ethical principles mentioned the most in AI ethics guidelines according to reviews. Results may diverge due to differences in dataset, size of dataset, and method.

Whereas the convergences on certain principles propose that AI ethics guidelines have some common ground, other analyses have shown that ethics guidelines still show high heterogeneity, especially when examined according to actor type (Rudschies *et al.*; Jobin *et al.*). Looking not only at the convergences, but also on the divergences is, therefore, crucial to understand the dynamics within the AI ethics debate.

By examining our sources with a focus on divergences, the documents show the different preferences of actor types *(see table 2 for an overview)*. Most striking is that private actors completely refrain from specifically mentioning primary principles such as freedom, dignity, and autonomy, while many public and expert actors consider them to be of utmost importance. For instance, 71.4% of public actors and 64.3% of expert actors mention freedom as a significant ethical principle in the design and deployment of AI. However, private actors do not discuss the principle specifically. Consequently, taking the documents of all actors together, only 47.5% of all ethics guidelines consider freedom as a principle for ethical AI, leaving the principle under the "majority count threshold"[4].

Notably, five private actors mention human rights overall, which include *inter alia* human dignity as fundamental and foundational norm and the right to freedom. It remains questionable, however, why certain principles falling under the framework of human rights such as privacy are additionally singled out by most private actors as guiding principles for ethical AI, while others such as freedom are not. Expert actors and public actors most often simultaneously mention human rights overall and single out the significance of the principles of freedom, autonomy, and dignity. Thus, there seems to be a discrepancy between private actors on the one side and public and expert actors on the other, whereby the former only mention certain principles falling under human rights and the latter do both. One can only make assumptions here, why this is the case. However, it might indicate diverging priorities of the different actors at hand.

Similar results to those related to the principle of freedom also appear for principles in the realm of democratic rights as well as sustainability. 71.4% of public actors and almost half of the expert actors mention the respect for democratic principles and the rule of law in their ethics guidelines, whereas it is only one company (Unity Technologies) that declares to "not *knowingly* develop AI tools and experiences that interfere with normal, functioning democratic systems of government" (Unity Technologies, 1, emphasis added). Similarly, most of the public actors and almost half of the expert actors consider sustainability to be part of ethical AI. In contrast, it is only one private actor that proposes the principle in its full sense and another one who mentions sustainable social development. Although companies can have substantial influence on issues such as sustainability, they do not seem to be part of private actors' agenda for an ethical AI.

Ethics guidelines from private actors also consider the principles of non-maleficence as well as more practical requirements such as contestability, accessibility and accuracy/completeness considerably less than the sources from the other two actor groups. In all of these cases the principles are mentioned by a majority of sources from public and expert actors, but are discussed only in a few private actors' documents, hindering their consideration as "core" ethical principles and requirements for ethical AI.

The priorities of private actors seem to lie elsewhere. They tend to prescribe themselves ethical principles that are widely discussed within the ethical debate among computer scientists (e.g. the FAT ML community focuses on issues such as privacy, non-discrimination, transparency). Further, it is noteworthy that private entities mention not only those principles most often for which technical fixes exist, as Hagendorff (2019) suggested, but also principles for which legislation is already in place (e.g. privacy, non-discrimination, security/safety). As the French Data Protection Authority points out in its report: "[…], the notion of ethics has evolved to refer to something alongside the law, used by stakeholders such as companies. […] Often, their only purpose can be to restate – consciously or otherwise – legal standards." (CNIL, 25) According to our data, there is no principle suggested by private actors' sources that has not been mentioned similarly often in the guidelines from public and expert actors. *Vice versa* this was considerably more often the case.

[4] We are aware that ethical principles are not voted into a universal ethics guideline using the majority rule. However, existing reviews have mostly used the frequency of ethical principles mentioned in the various guidelines as an indication for their overall importance, basing their results on some sort of majority count. This analysis, however, shows that such a method might undermine important ethical principles in the discussion.

Contrary to private actors' documents, guidelines from public actors discuss the principle of non-discrimination and fairness as well democracy and rule of law considerably more than expert and private actors, showing different priorities across actor types. This holds also true for the principles of value-alignment and inclusiveness. For instance, 92.9% of sources from public actors stress that AI systems should be aligned with human, societal and/or ethical values. The principle is, however, only mentioned by 50% of guidelines from private actors and a mere 28.6% of guidelines from expert actors.

The principle of responsibility is most often put forward in the guidelines from expert actors. All except for one expert actor's guideline mentions responsibility as a necessary principle for ethical AI. By contrast, it is mentioned by 71.4% of sources from public actors and 66.7% of sources from private actors. Generally, expert actors propose principles such as controllability, non-maleficence, accountability, and contestability slightly more often than public actors, showing a light tendency for principles that are concerned with preventing harm from individuals by providing them with some form of control and opportunity to redress. Also here, it is the guidelines from private actors that show considerably more reluctance towards such principles.

Notably, different priorities per actor type become also apparent in regard to those principles that are considered "common ground"[5]. While the principles of privacy and transparency have been mentioned by all actor types in similar frequency, this does not hold true for other "core" principles. For instance, 85.7% of expert actors and 78.6% of public actors considered accountability to be part of ethical AI. Among private actors the principle was mentioned by 58.3%. Similarly, the principle of explainability was put forward by 92.9% of all public actors and 85.7% of all expert actors. It was, however, discussed only by 50% of all private actors. Interestingly, the rate of occurrence of the principle intelligibility is even lower than that of explainability. While 64.3% of sources from public actors and 71.4% of sources from expert actors mention intelligibility often as an additional principle to transparency and explainability, it is only 25% of sources from the private sector that propose that AI systems and their decisions should be intelligible to individuals deploying, using and/or overseeing them[6].

Lastly, as Hagendorff has also found, certain principles are generally underrepresented. For instance, principles such as solidarity or trust are only rarely mentioned in the AI ethics guidelines. Yet, research has found that Big Data and AI technologies are likely to exacerbate social inequalities and a fragmentation of society (Eubanks; Henman; O'Neil). Hagendorff thus states that the principle of solidarity might help to foster social cohesion. Miguel Luengo-Oroz even claims that solidarity should be a core ethical principle for AI, for it can help to collectively share AI's benefits and risks for instance via redistributive measures. In relation to trust, the European Commission has recently published a White Paper on AI that puts great emphasis on the prevention of risks and the necessity of a regulatory framework that strengthens the public's and users' trust in AI technologies.

Against these backgrounds, there seems to be a need for the consideration of the principles of solidarity and trust, which has not been met by the AI ethics guidelines examined here. The same applies to the principle of sustainability, which has found especially little attention by private actors, as discussed above, even though the research on climate change postulates a climate crisis if measures against global warming aren't taken. Notably, sustainability, solidarity, and trust provide no immediate answer to how these principles should be interpreted in the context of AI. Stakeholders proposing a list of principles might therefore not be inclined to include such principles in their guidelines, since they are very difficult to conceptualise and, consequently, also to operationalise. Further research on these principles is, hence, desperately needed.

By concentrating on the varying preferences of the three actor groups (public, expert, private actors), several major findings were made: first, divergences across actor types exist also in relation to principles that were found to be "common ground"; second, while expert and public actors often appear to have relatively similar views on AI ethics, it is especially private actors who deviate from those views and thus "hinder" particular principles to become (at least quantitatively) "common ground"; third, while public and expert actors put additional emphasis on those values linked to fundamental rights

[5] The case of the responsibility principle described above is also one of these cases.

[6] Overall, the principle of intelligibility was mentioned in 55% of all sources.

and democratic principles like freedom, dignity, and autonomy, as well as principles that go beyond existing discussions and regulation, private actors tend to put forward rather those ethical principles for which technical solutions exist or legislation is already in place. We will turn to the discussion of these findings in the next part.

Principles	**Proportion of ethics guidelines mentioning principle X - all three actor groups combined**	**Proportion of ethics guidelines mentioning principle X - public actors**	**Proportion of ethics guidelines mentioning principle X - expert actors**	**Proportion of ethics guidelines mentioning principle X - private actors**
Principles of "common ground"				
transparency	95%	100%	92.9%	92.7%
privacy	87.5%	85.7%	92.9%	83.3%
non-discrimination and fairness	82.5%	92.9%	78.6%	75%
responsibility	77.5%	71.4%	92.9%	66.7%
explainability	77.5%	92.9%	85.7%	50%
safety and security	75%	78.6%	78.6%	66.7%
accountability	75%	78.6%	85.7%	58.3%
non-maleficence	67.5%	71.4%	85.7%	41.7%
Human and citizen rights	62.5%	78.6%	64.3%	41.7%
Principles without an overall "majority"				
freedom	47.5%	71.4%	64.3%	0%
dignity	32.5%	42.9%	50%	0%
sustainability	37.5%	57.1%	35.7%	16.7%
democracy and rule of law	42.5%	71.4%	42.9%	8.3%
contestability	40%	50%	57.1%	8.3%
accuracy and completeness	40%	50%	50%	16.7%

Table 2: Convergences and divergences on a selection of principles according to ethics guidelines overall and among different actor types.

5. Discussing the Ethical Priorities in Light of Value Pluralism

The analysis has shown that AI ethics guidelines encompass different preferences for ethical principles depending on actor type. It, hence, supports the findings of an earlier study we conducted that the three actor types follow different approaches when it comes to the ethics of AI. Similarly to the diverging natures and approaches of AI ethics guidelines (Rudschies *et al.*), public, expert, and private actors also seem to have diverging priorities which values and principles ought to guide the design, deployment, and utilisation of AI technologies. While certain convergences on ethical principles exist and mainly overlap with the results of existing reviews, it is especially the divergences that shed some light on the contemporary shortcomings of the AI ethics debate as well as the dynamics between actor types.

Expert and public actors have been found to have relatively similar priorities regarding the ethical principles for AI development and use. There are, nonetheless, slight differences. Public actors tend to put more emphasis on principles such as non-discrimination and fairness, fundamental rights such as freedom as well as democracy and the rule of law. In comparison, expert actors mention principles such as responsibility, accountability, non-maleficence, and contestability more often than public and private actors. The ethics guidelines from private actors reveal the most deviating preferences of all sources examined. They generally encompass considerably fewer principles than those from other actor types. The principles put forward are often those for which legislation and technical fixes already exist or are at least possible. Private actors' sources do not mention any principle more often than those from other actor types. Considering the frequency by which principles have been mentioned in the guidelines, it is the sources from private actors that most often hinder certain principles from becoming "common ground".

Interestingly, divergences in the preferences for ethical principles do not only appear in respect to those principles that have *not* been considered "common ground" (e.g. dignity, freedom), but also for those principles that *actually have*. For instance, even though being mentioned in the majority of all ethics guidelines examined, human and citizen rights as well as the principles of explainability, accountability, and non-maleficence have been proposed by *at least* 20% fewer guidelines from private actors compared with those from the other actor types (*see table 2*). Furthermore, some ethical principles have been found to be underrepresented in the ethics guidelines overall.

As the diverging preferences of actor types as well as the varying sets of synthesised principles of the other reviews reveal, it proves rather difficult to find one universal set of principles in the context of AI. This is especially true in a pluralist society, where individuals and subgroups have different backgrounds and, thus, value perspectives on the issue at hand. Hence, one must refute a monist view of an ethics of AI, which assumes that there is a single value or a small set of values that overrides all others or provides guidance on how to compare one value with another. Rather, we need to acknowledge that the AI ethics landscape is characterised by a plurality of sets of values. According to the theory of moral or value pluralism, there is, due to diverging understandings of a good life, a multitude of genuine human goods and values. Some of these values might even be incommensurable (Wendel). Pluralism implies, writes Crowder, that there is a wide range of legitimate and reasonable rankings (417).

Consequently, determining "minimum requirements" for ethical AI by quantitatively assessing which ethical principles have found convergence in most AI ethics guidelines is problematic, for it ignores the variety of legitimate ethical principles put forward and disregards conflicts and controversy. Synthesising ethical principles to a set of "primary principles" is a form of classification that attempts to rank them based on their frequency rather than ethical relevance. Given the sheer number of ethical principles proposed in the various AI ethics guidelines, it stands to reason that scholars and stakeholders within the debate are aiming to synthesise them and find some common ground. However, by doing so, the actors' preferences for ethical AI are simply aggregated in order to determine which "primary principles" ought to guide the design, deployment and utilisation of AI. In utilitarian terms, the approach, hence, aims to satisfy the preferences and increase the overall utility. However, such an approach undermines the fact that certain ethical principles are subject to controversy and conflict (between actor types, for instance), but might still be ethically relevant. Furthermore, the results of such an approach are highly dependent on the dynamics between the actors in play, for the principles with the highest frequency will only reflect the most dominant viewpoints. Ethical values put forward by minority groups might, however, be disregarded.

By acknowledging the plurality of values and principles in the AI ethics landscape, the analysis changes from a sole focus on convergences (meaning here the classification or ranking of values by frequency to arrive at some "primary principles" or "common ground") towards the examination of divergences and convergences alike. Such an approach is crucial, for a mere focus on convergences can mask the debate as uncontroversial and, hence, presents a picture of the AI ethics landscape that is more homogenous than it actually is. Such a picture potentially hinders the consideration of legitimate principles. As Crowder argues in a pluralist tradition: "If these are genuine human goods, we must not be indifferent to them, even when we have to choose against them" (421). Thus, a focus on convergences also hampers the open discussion of dissenting opinions, conflicts, and alternative courses of action. Notably, a consensus-driven approach is favourable because it shows that agreement among participants of the debate is sought. However, it usually represents the lowest common denominator and, hence, does not necessarily depict all relevant aspects of the issue. Looking at the divergences across actor types by taking a more pluralist view on the AI ethics debate is, therefore, a first step to shedding light on the different priorities and determining where further discussion is still needed.

6. Two Problems of Value Pluralism in the AI Ethics Debate

Whereas the recognition of the plurality of ethical principles is crucial to steer the attention towards the conflicts, controversies, and shortcomings of the debate, it also brings to the surface two problems. First, by focusing on the divergences in the AI ethics debate, one seems to be farther away from a unified framework that can guide the governance of AI. And second, acknowledging a plurality of (sets of) ethical principles can cloud the fact that a multitude of AI ethics guidelines is not based on purely moral considerations. We will turn to the latter issue first.

In traditional philosophical thought, an act can only be ethical when motivated by purely moral concerns, whereby "the word 'moral' refers to formal conditions, such as supremacy, concern for others, and universalizability, that characterize a particular kind of judgment" (Wendel, 124). Ethical judgement, hence, ought to be made on the basis of rational arguments remote from one's individual (or particular group) interest.

As we have argued in an earlier study (Rudschies *et al.*), the approaches taken by the three actor types in respect to the ethics of AI are likely to be influenced by the actor's function, perspective, mandate, and aims. For instance, ethics guidelines from private actors are rather vague, short, and do not specify any oversight or control. Ethics was presented as a form of voluntary self-commitment (ibid). The analysis of this paper shows that private actors are the ones most likely to hinder certain principles from becoming "common ground". Moreover, they tend to propose principles for which legislation is in place or technical fixes already exist. Given that the private actor's role is to develop and provide AI technologies with the aim to create some value for their customers and – in the form of profit – for themselves, they might be reluctant to impose further requirements and thus restraints on their work, because it conflicts with their interests. It follows that if it is true that the approaches in the ethics of AI are influenced by the actors' role and aims, then the choice of ethical principles is potentially not based on purely moral considerations. Rather, the principles might have been chosen to serve the interests of the actor proposing it and, hence, could not be called ethical.

In order to overcome this problem, actor groups need to find a common understanding of what constitutes ethical AI by focusing on the moral concern for others (or the common good) rather than on their self-interest. Referring to the first problem mentioned above, such a common understanding will also help to reach a social consensus that all actors find acceptable and that can serve as guidance for governance. We will explain in the next section, how one can approach this task from a pluralist point of view.

7. An Attempt to Overcome the Problems in the AI Ethics Debate

In order to find a more common understanding of all actor groups, a pluralist view on AI ethics offers a promising approach. Intuitively, this might sound contradictory, since pluralisms demands the acknowledgement of the diversity of ethical values and principles, leaving little room for a unified framework. Nonetheless, pluralism does not claim that all human values should be equally complied with. This is not only practically impossible, because values and principles can stand in conflict with each other. It is also ethically questionable, since one cannot simply impose the values of one individual or subcommunity onto the other. Rather, pluralism merely points out that "when we choose against a good, we should […] recognize that the good we forego is still valuable. It follows that pluralist choice should not be merely arbitrary or casual" (Crowder, 421). Rather, it is only through rational reflection that ethical judgements, including the choice of ethical principles, can be made. A unified framework must, hence, be "based on terms acceptable to all concerned and not just to dominant individuals or groups" (ibid, 422).

Hence, to arrive at a more unified scheme of ethical principles, the various actor groups and members of society need to enter – in a Habermasian tradition – into an active and meaningful process of rational deliberation, where the plurality of human values must be recognised and debated under purely moral considerations. Wendel argues, that just like in scientific debates also in ethical deliberation one ought to approach the "rival culture" with some degree of imagination and sympathy "to reach a sufficiently rich understanding of the other culture's practices to compare them with similar practices in our culture" (169). In order to arrive at such understandings, van der Wal and van Hout argue that it is necessary to clarify interpretations and meanings among each other in order to avoid ambiguity in semantics (224). Within the AI ethics debate Jobin *et al.* and Fjeld *et al.* have already pointed out that such clarifications are desperately needed, for ethical principles are often relatively vague and leave room for ample interpretation.

By entering into such a process of deliberation and understanding, it is possible to separate particular interests from true value judgements and, subsequently, find a social consensus on what constitutes ethical AI for our society overall. "Rival cultures" of the diverse individuals or subcommunities (e.g. the different actor types) and the values that arise from them are tested for their ethical relevance. Principles that are based on pure self-interest cannot withstand such a test, for they will most likely not be acceptable to all members of the society. The AI ethics landscape is currently highly heterogenous, but the pluralism underlying it can be used to further the debate and find a more common approach to the ethics of AI. This paper has tried to make a first step in this direction.

8. Conclusion

This paper has analysed the current AI ethics debate with a focus on divergences using the theory of moral or value pluralism. It has shown that current reviews of AI ethics guidelines put too much emphasis on the convergences in AI ethics guidelines, an approach that aims to find some "minimum requirements", "common ground" or "primary principles" for the ethical design, deployment and utilisation of AI technologies. Although certainly an important endeavour, the divergences should not be overlooked as they shed light on the conflicts and controversies as well as the issues that require further debate and consideration.

By conducting a content analysis of 40 AI ethics guidelines this paper has found that although there are some convergences, the priorities for ethical principles diverge across actor types. By discerning the varying priorities of different actor types, the study has shown that a universal ranking of principles for the ethics of AI is difficult to find, for different individuals and subcommunities can have varying understandings of an ethical AI. Rather, the AI ethics landscape is marked by a plurality of sets of principles that need to be recognised and debated among the diverse individuals and subgroups within society. Nevertheless, actors' priorities should be discussed and tested only on basis of their moral relevance, not their relevance in terms of self-interest. Pluralism does not imply that "anything goes" (Wendel, 209), but that diverse value claims might simultaneously be legitimate and, hence, need to be considered. By engaging stakeholders in a rational deliberation process, it is possible to reconcile conflicts and controversies. If the landscape of AI ethics guidelines remains as heterogeneous as it currently is, it is likely that parties involved in the design and utilisation of AI are torn between the different incongruent regimes. The insights on divergences in the actors' ethical preferences, as generated in this paper, can constructively be used to further the debate and, hence, reach a social consensus.

Acknowledgements

This research is part of the project “GOAL – Governance of and by algorithms” funded by the German Federal Ministry of Education and Research (BMBF) under reference number 01IS19020D. The authors bear full responsibility for the content of this research.

9. References

*Access Now. Human rights in the age of artificial intelligence, 2018, www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf. **

Amnesty International & Access Now. The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems, 2018, www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

Association for Computing Machinery (ACM). Statement on Algorithmic Transparency and Accountability. Washington: Association for Computing Machinery US Public Policy Council, 2017, www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

Austrian Council on Robotics and Artificial Intelligence. Shaping the Future of Austria with Robotics and Artificial Intelligence, 2018, www.acrai.at/wp-content/uploads/2020/03/ACRAI_White_Paper_EN.pdf.

CNIL - The French Data Protection Authority. How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence, Paris: Commission Nationale de l'Informatique et des Libertés, 2017, www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf.

Council of Europe, Committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence, 12 November 2018, Draft Recommendation of the Committee of Ministers to member States on human rights impacts of algorithmic systems. Strasbourg: Council of Europe, 2018. MSI-AUT(2018)06.

Crowder, George. Value pluralism and communitarianism. Contemporary Political Theory 5.4 (2006): 405-427.

Data Ethics Commission of the Federal Government of Germany (DEK), Federal Ministry of the Interior, Building and Community, Federal Ministry of Justice and Consumer Protection. Opinion of the Data Ethics Commission, Berlin: DEK, 2019. datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf.

Dawson, D. & Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J., and Hajkowicz, S. Artificial Intelligence: Australia's Ethics Framework. Data61 CSIRO, Australia, 2019, consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf.

Deutsche Telekom. AI Guidelines, 2018. www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366.

Eubanks, Virginia. Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press, 2018.

European Commission, High-Level Expert Group on AI. Ethics Guidelines for Trustworthy AI. Brussels: European Commission, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

European Commission, Directorate-General for Research and Innovation, European Group on Ethics in Science and New Technologies. Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems. Brussels: European Commission, 2018, ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

European Commission. White Paper on Artificial Intelligence – A European approach to excellence and trust. Brussels, European Commission, 2020, ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

FRA – European Union Agency for Fundamental Rights. AI policy initiatives (2016-2020), 2020, fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights/ai-policy-initiatives.

Fairness, Accountability, and Transparency in Machine Learning (FAT ML). Principles for Accountable Algorithms and a Social Impact Statement for Algorithms, 2018, www.fatml.org/resources/principles-for-accountable-algorithms.

Fjeld, Jessica, et al. "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI." Berkman Klein Center Research Publication 2020-1 (2020).

Floridi, Luciano, et al. "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations." Minds and Machines 28.4 (2018): 689-707.

Future of Life Institute. Asilomar AI Principles, 2017, futureoflife.org/ai-principles/.

Government of Japan, The Conference toward AI Network Society. Draft AI R&D GUIDELINES for International Discussions, 2017, www.soumu.go.jp/main_content/000507517.pdf.

Hagendorff, Thilo. "The ethics of AI ethics - an evaluation of guidelines." (2019), arXiv preprint arXiv:1903.03425.

Henman, Paul. "Improving public services using artificial intelligence: possibilities, pitfalls, governance." Asia Pacific Journal of Public Administration 42.4 (2020): 209-221.

House of Lords, Select Committee on Artificial Intelligence. AI in the UK: ready, willing and able? Report of Session 2017-19, London: House of Lords, 2018, publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

IBM. Everyday Ethics for Artificial Intelligence, 2019, www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

Intel. Artificial Intelligence - The Public Policy Opportunity, 2017, blogs.intel.com/policy/files/2017/10/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf.

International Conference of Data Protection and Privacy Commissioners (ICDPPC). Declaration on Ethics and Data Protection in Artificial Intelligence, 40th International Conference, Brussels, 2018, https://www.privacyconference2018.org/system/files/2018-10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf.

International Working Group on Data Protection in Telecommunications (IWGDPT), 64th Meeting, 29-30 November 2018, Queenstown (New Zealand). Working Paper on Privacy and Artificial Intelligence, 2018, www.datenschutz-berlin.de/fileadmin/user_upload/pdf/publikationen/working-paper/2018/2018-IWGDPT-Working_Paper_Artificial_Intelligence.pdf.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." Nature Machine Intelligence 1.9 (2019): 389-399.

Korea Artificial Intelligence Ethics Association (kaiea). The AI Ethics Charter. Seoul: keiea, 2019, drive.google.com/file/d/18fCgEhXKxTjB9uovFX9oVwqyz5XCEptb/view.

Malta.AI, Financial Services, Digital Economy and Innovation, Office of the Prime Minister. Malta. Towards Trustworthy AI – Malta Ethical AI Framework for Public Consultation. Valletta: Malta.AI, 2019, malta.ai/wp-content/uploads/2019/08/Malta_Towards_Ethical_and_Trustworthy_AI.pdf.

Microsoft. Microsoft AI principles, 2018, www.microsoft.com/en-us/ai/responsible-ai.

National Endowment for Science, Technology and the Arts (Nesta). AI Governance Database, n.d., www.nesta.org.uk/data-visualisation-and-interactive/ai-governance-database/.

OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449.

O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, 2016.

OP Financial Group. OP Financial Group's ethical guidelines for artificial intelligence, 2018, www.op.fi/op-financial-group/corporate-social-responsibility/commitments-and-principles.

Partnership on AI. Tenets, 2016, www.partnershiponai.org/tenets/.

Pichai, Sundar. AI at Google: Our Principles, 2018. www.blog.google/technology/ai/ai-principles/.

Rotenberg, Marc. The AI Policy Sourcebook 2019. Electronic Privacy Information Center, Washington DC, ed. 2019.

Rudschies, Catharina, Schneider, Ingrid & Simon, Judith. The Heterogeneity of AI Ethics Guidelines Examined: Varying Natures, Actors, and Perceptions. Forthcoming.

Sage. The Ethics of Code: Developing AI for Business with Five Core Principles, 2017, www.sage.com/~/media/group/files/business-builders/business-builders-ethics-of-code.pdf?la=en.

SAP. SAP's Guiding Principles for Artificial Intelligence, 2018, news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/.

Smart Dubai. AI Ethics Principles & Guidelines, 2018, www.smartdubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?sfvrsn=d4184f8d_6.

Sony. AI Engagement within Sony Group, 2018, www.sony.net/SonyInfo/csr_report/humanrights/hkrfmg0000007rtj-att/AI_Engagement_within_Sony_Group.pdf.

Task Force on Artificial Intelligence of the Agency for Digital Italy ai.italia.it. Artificial Intelligence at the service of citizens. The Agency for Digital Italy (AGID), 2018, ia.italia.it/assets/whitepaper.pdf.

Telefónica. AI Principles of Telefónica, 2018, www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles.

Telia. Guiding Principles on Trusted AI Ethics, 2019, www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf.

The Danish Government, Ministry of Finance and Ministry of Industry, Business and Financial Affairs. National Strategy for Artificial Intelligence. Copenhagen: The Danish Government, 2019, eng.em.dk/media/13081/305755-gb-version_4k.pdf.

The Institute of Electrical and Electronics Engineers (IEEE). Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition, 2019, standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.

The Public Voice. Universal Guidelines for Artificial Intelligence, Explanatory Memorandum and References, 2018, thepublicvoice.org/ai-universal-guidelines/memo/.

UNI Global Union. Top 10 principles for ethical artificial intelligence. Switzerland, UNI Global Union, 2016, www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.

United Nations Educational, Scientific and Cultural Organization (UNESCO). Preliminary study on the technical and legal aspects relating to the desirability of a standard-setting instrument on the ethics of artificial intelligence. Paris: UNESCO, 2019, unesdoc.unesco.org/ark:/48223/pf0000367422.

Unity Technologies. Introducing Unity's Guiding Principles for Ethical AI, 2018, blogs.unity3d.com/2018/11/28/introducing-unitys-guiding-principles-for-ethical-ai/.

University of Montréal. Montréal Declaration for a responsible development of artificial intelligence, 2018, 5dcfa4bd-f73a-4de5-94d8-c010ee777609.filesusr.com/ugd/ebc3a3_506ea08298cd4f8196635545a16b071d.pdf.

Wendel, W. Bradley. "Value pluralism in legal ethics." Wash. ULQ 78 (2000): 113.

**All online sources were last accessed on 23 Jan. 2021.*