Minds and Machines (2020) 30:99–120
https://doi.org/10.1007/s11023-020-09517-8

The Ethics of AI Ethics: An Evaluation of Guidelines

Thilo Hagendorff[1]

Received: 1 October 2019 / Accepted: 21 January 2020 / Published online: 1 February 2020
© The Author(s) 2020

Abstract
Current advances in research, development and application of artificial intelligence (AI) systems have yielded a far-reaching discourse on AI ethics. In consequence, a number of ethics guidelines have been released in recent years. These guidelines comprise normative principles and recommendations aimed to harness the "disruptive" potentials of new AI technologies. Designed as a semi-systematic evaluation, this paper analyzes and compares 22 guidelines, highlighting overlaps but also omissions. As a result, I give a detailed overview of the field of AI ethics. Finally, I also examine to what extent the respective ethical principles and values are implemented in the practice of research, development and application of AI systems—and how the effectiveness in the demands of AI ethics can be improved.

Keywords Artificial intelligence · Machine learning · Ethics · Guidelines · Implementation

1 Introduction

The current AI boom is accompanied by constant calls for applied ethics, which are meant to harness the "disruptive" potentials of new AI technologies. As a result, a whole body of ethical guidelines has been developed in recent years collecting principles, which technology developers should adhere to as far as possible. However, the critical question arises: Do those ethical guidelines have an actual impact on human decision-making in the field of AI and machine learning? The short answer is: No, most often not. This paper analyzes 22 of the major AI ethics guidelines and issues recommendations on how to overcome the relative ineffectiveness of these guidelines.

AI ethics—or ethics in general—lacks mechanisms to reinforce its own normative claims. Of course, the enforcement of ethical principles may involve

✉ Thilo Hagendorff
thilo.hagendorff@uni-tuebingen.de

[1] Cluster of Excellence "Machine Learning: New Perspectives for Science", University of Tuebingen, Tübingen, Germany

reputational losses in the case of misconduct, or restrictions on memberships in certain professional bodies. Yet altogether, these mechanisms are rather weak and pose no eminent threat. Researchers, politicians, consultants, managers and activists have to deal with this essential weakness of ethics. However, it is also a reason why ethics is so appealing to many AI companies and institutions. When companies or research institutes formulate their own ethical guidelines, regularly incorporate ethical considerations into their public relations work, or adopt ethically motivated "self-commitments", efforts to create a truly binding legal framework are continuously discouraged. Ethics guidelines of the AI industry serve to suggest to legislators that internal self-governance in science and industry is sufficient, and that no specific laws are necessary to mitigate possible technological risks and to eliminate scenarios of abuse (Calo 2017). And even when more concrete laws concerning AI systems are demanded, as recently done by Google (2019), these demands remain relatively vague and superficial.

Science- or industry-led ethics guidelines, as well as other concepts of self-governance, may serve to pretend that accountability can be devolved from state authorities and democratic institutions upon the respective sectors of science or industry. Moreover, ethics can also simply serve the purpose of calming critical voices from the public, while simultaneously the criticized practices are maintained within the organization. The association "Partnership on AI" (2018) which brings together companies such as Amazon, Apple, Baidu, Facebook, Google, IBM and Intel is exemplary in this context. Companies can highlight their membership in such associations whenever the notion of serious commitment to legal regulation of business activities needs to be stifled.

This prompts the question as to what extent ethical objectives are actually implemented and embedded in the development and application of AI, or whether merely good intentions are deployed. So far, some papers have been published on the subject of teaching ethics to data scientists (Garzcarek and Steuer 2019; Burton et al. 2017; Goldsmith and Burton 2017; Johnson 2017) but by and large very little to nothing has been written about the tangible implementation of ethical goals and values. In this paper, I address this question from a theoretical perspective. In a first step, 22 of the major guidelines of AI ethics will be analyzed and compared. I will also describe which issues they omit to mention. In a second step, I compare the principles formulated in the guidelines with the concrete practice of research and development of AI systems. In particular, I critically examine to what extent the principles have an effect. In a third and final step, I will work out ideas on how AI ethics can be transformed from a merely discursive phenomenon into concrete directions for action.

2 Guidelines in AI Ethics

2.1 Method

Research in the field of AI ethics ranges from reflections on how ethical principles can be implemented in decision routines of autonomous machines (Anderson and

Anderson 2015; Etzioni and Etzioni 2017; Yu et al. 2018) over meta-studies about AI ethics (Vakkuri and Abrahamsson 2018; Prates et al. 2018; Boddington 2017; Greene et al. 2019; Goldsmith and Burton 2017) or the empirical analysis on how trolley problems are solved (Awad et al. 2018) to reflections on specific problems (Eckersley 2018) and comprehensive AI guidelines (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019). This paper mainly deals with the latter issue. The list of ethics guidelines considered in this article therefore includes compilations that cover the field of AI ethics as comprehensively as possible. To the best of my knowledge, a few preprints and papers are currently available, which also deal with the comparison of different ethical guidelines (Zeng et al. 2018; Fjeld et al. 2019; Jobin et al. 2019). While especially the paper from Jobin et al. (2019) is a systematic scoping review of all the existing literature on AI ethics, this paper does not aim at a full analysis of every available soft-law or non-legal norm document on AI, algorithm, robot, or data ethics, but rather a semi-systematic overview of issues and normative stances in the field, demonstrating how the details of AI ethics relate to a bigger picture.

The selection and compilation of 22 major ethical guidelines were based on a literature analysis. This selection was undertaken in two phases. In the first phase, I searched different databases, namely Google, Google Scholar, Web of Science, ACM Digital Library, arXiv, and SSRN for hits or articles on "AI ethics", "artificial intelligence ethics", "AI principles", "artificial intelligence principles", "AI guidelines", and "artificial intelligence guidelines, following every link in the first 25 search results, while at the same time ignoring duplicates in the search process. During the analysis of the search results, I also sifted through the references in order to manually find further relevant guidelines. Furthermore, I used Algorithm Watch's AI Ethics Guidelines Global Inventory, a crowdsourced, comprehensive list of ethics guidelines, to check whether I missed relevant guidelines. Via the list, I found three further guidelines that meet the criteria for the selection. In this context, a shortcoming one has to consider is that my selection is biased towards documents which are western/northern in nature, excluding guidelines which are not written in English.

I rejected all documents older than 5 years in order to only take guidelines into account that are relatively new. Documents that only refer to a national context—such as for instance position papers of national interest groups (Smart Dubai Smart Dubai 2018), the report of the British House of Lords (Bakewell et al. 2018), or the Nordic engineers' stand on Artificial Intelligence and Ethics (Podgaiska and Shklovski)—were excluded from the compilation. Nevertheless, I included the European Commission's "Ethics Guidelines for Trustworthy AI" (Pekka et al. 2018), the Obama administration's "Report on the Future of Artificial Intelligence" (Holdren et al. 2016), and the "Beijing AI Principles" (Beijing Academy of Artificial Intelligence 2019), which are backed by the Chinese Ministry of Science and Technology. I have included these three guidelines because they represent the three largest AI "superpowers". Furthermore, I included the "OECD Principles on AI" (Organisation for Economic Co-operation and Development 2019) due to their supranational character. Scientific papers or texts that fall into the category of AI ethics but focus on one or more specific aspects of the topic were not considered either. The same applies to guidelines or toolkits,

which are not specifically about AI but rather about big data, algorithms or robotics (Anderson et al. 2018; Anderson and Anderson 2011). I further excluded corporate policies, with the exception of the "Information Technology Industry AI Policy Principles" (2017), the principles of the "Partnership on AI" (2018), the IEEE first and second version of the document on "Ethically Aligned Design" (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2016, 2019), as well as the brief principle lists of Google (2018), Microsoft (2019), DeepMind (DeepMind), OpenAI (2018), and IBM (Cutler et al. 2018) which have become well-known through media coverage. Other large companies such as Facebook or Twitter have not yet published any systematic AI guidelines, but only isolated statements of good conduct. Paula Boddington's book on ethical guidelines (2017) funded by the Future of Life Institute was also not considered as it merely repeats the Asilomar principles (2017).

The decisive factor for the selection of ethics guidelines was not the depth of detail of the individual document, but the discernible intention of a comprehensive mapping and categorization of normative claims with regard to the field of AI ethics. In Table 1, I only inserted green markers if the corresponding issues were explicitly discussed in one or more paragraphs. Isolated mentions without further explanations were not considered, unless the analyzed guideline is so short that it consists entirely of brief mentions altogether.

Table 1 Overview of AI ethics guidelines and the different issues they cover

	The European Commission's High-Level Expert Group on Artificial Intelligence	Report on the Future of Artificial Intelligence	Beijing AI Principles	OECD Recommendation of the Council on Artificial Intelligence	The Montreal Declaration of Responsible AI	Asilomar AI Principles	The Malicious Use of Artificial Intelligence	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	AI Now 2019 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Montréal Declaration for Responsible Development of Artificial Intelligence	OpenAI Charter	IEEE Ad Hoc Committee 'Ethically Aligned Design'	IEEE Ad Hoc Committee 'Ethically Aligned Design, A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition)'	IT Industry Principles	Microsoft AI principles	DeepMind Ethics & Society Principles	Artificial Intelligence at Google	Everyday Ethics for Artificial Intelligence	Partnership on AI	number of mentions
authors	(Pekka et al. 2018)	(Stoddart et al. 2016)	(Beijing Academy of Artificial Intelligence 2019)	(Organisation for Economic Co-operation and Development 2019)	(Brundage et al. 2018)	(Floridi et al. 2018)	(Future of Life Institute 2017)	(Crawford et al. 2016)	(Campolo et al. 2017)	(Whittaker et al. 2018)	(Crawford et al. 2019)	(Diakopoulos et al.)	(Abrassart et al. 2018)	(OpenAI 2018)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2016)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019)	(Information Technology Industry Council 2017)	(Microsoft Corporation 2019)	(DeepMind)	(Google 2018)	(Cutler et al. 2018)	(Partnership on AI 2018)	
key issue	AI principles of the EU	AI principles of the US	AI principles of China	AI principles of the OECD	analysis of abuse scenarios of AI	meta-analysis about principles for the beneficial use of AI	large collection of different principles	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	principles of the FAT ML community	code of ethics released by the Université de Montréal	several short principles for the ethical use of AI	detailed description of ethical aspects in the context of AI	detailed description of ethical aspects in the context of AI	brief guideline about basic ethical principles	short list of keywords for the ethical use of AI	several short principles for the ethical use of AI	several short principles for the ethical use of AI	IBM's short list of keywords for the ethical use of AI	principles of an association between several industry leaders	
privacy protection																							18
fairness, non-discrimination, justice																							18
accountability																							17
transparency, openness																							16
safety, cybersecurity																							16
common good, sustainability, well-being																							16
human oversight, control, auditing																							12
solidarity, inclusion, social cohesion																							11
explainability, interpretability																							10
science-policy link																							10
legislative framework, legal status of AI systems																							10
future of employment/worker rights																							9
responsible/intensified research funding																							8
public awareness, education about AI and its risks																							8
dual-use problem, military, AI arms race																							8
field-specific deliberations (health, military, mobility etc.)																							8
human autonomy																							7
diversity in the field of AI																							7
certification for AI products																							4
protection of whistleblowers																							3
cultural differences in the ethically aligned design of AI systems																							2
hidden costs (labeling, clickwork, contend moderation, energy, resources)																							2
notes on technical implementations	yes, but very few	none	none	none	yes	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	
proportion of women among authors (f/m)	(8/10)	(2/3)	ns	ns	(5/21)	(5/8)	ns	(6/2)	(3/1)	(6/4)	(12/4)	(1/12)	(8/10)	ns	varies in each chapter	varies in each chapter	ns	ns	ns	ns	(1/2)	ns	(55/77)
length (number of words)	16546	22787	766	3249	34017	8809	646	11530	18273	25759	38870	1359	4754	461	40915	108.092	2272	73	417	882	4488	1481	
affiliation (government, industry, science)	government	government	science/gov./ind.	government	science	science	science	science	science	science	science	science	science	non-profit	industry	industry	industry	industry	industry	industry	industry	industry	
number of ethical aspects	9	12	13	12	8	14	12	13	9	12	13	5	11	4	14	18	9	6	6	6	6	6	

2.2 Multiple Entries

As shown in Table 1, several issues are unsurprisingly recurring across various guidelines. Especially the aspects of *accountability*, *privacy* or *fairness* appear altogether in about 80% of all guidelines and seem to provide the minimal requirements for building and using an "ethically sound" AI system. What is striking here is the fact that the most frequently mentioned aspects are those for which technical fixes can be or have already been developed. Enormous technical efforts are undertaken to meet ethical targets in the fields of *accountability* and *explainable AI* (Mittelstadt et al. 2019), *fairness* and *discrimination aware data mining* (Gebru et al. 2018), as well as *privacy* (Baron and Musolesi 2017). Many of those endeavors are unified under the FAT ML or XAI community (Veale and Binns 2017; Selbst et al. 2018). Several tech-companies already offer tools for *bias mitigation* and *fairness* in machine learning. In this context, Google, Microsoft and Facebook have issued the "AI Fairness 360" tool kit, the "What-If Tool", "Facets", "fairlern.py" and "Fairness Flow", respectively (Whittaker et al. 2018).

Accountability, *explainability*, *privacy*, *justice*, but also other values such as *robustness* or *safety* are most easily operationalized mathematically and thus tend to be implemented in terms of technical solutions. With reference to the findings of psychologist Carol Gilligan, one could argue at this point that the way AI ethics is performed and structured constitutes a typical instantiation of a male-dominated justice ethics (Gilligan 1982). In the 1980s, Gilligan demonstrated in empirical studies that women do not, as men typically do, address moral problems primarily through a "calculating", "rational", "logic-oriented" ethics of justice, but rather interpret them within a wider framework of an "empathic", "emotion-oriented" ethics of care. In fact, no different from other parts of AI research, the discourse on AI ethics is also primarily shaped by men. My analysis of the distribution of female and male authors of the guidelines, as far as authors were indicated in the documents, showed that the proportion of women was 41.7%. This ratio appears to be close to balance. However, it should be considered that the ratio of female to male authors is reduced to a less balanced 31.3% if the four AI Now Reports are discarded, which come from an organization that is deliberately led by women. The proportion of women is lowest at 7.7% in the FAT ML community's guidelines which are focused predominantly on technical solutions (Diakopoulos et al.). Accordingly, the "male way" of thinking about ethical problems is reflected in almost all ethical guidelines by way of mentioning aspects such as *accountability*, *privacy* or *fairness*. In contrast, almost no guideline talks about AI in contexts of care, nurture, help, welfare, social responsibility or ecological networks. In AI ethics, technical artefacts are primarily seen as isolated entities that can be optimized by experts so as to find technical solutions for technical problems. What is often lacking is a consideration of the wider contexts and the comprehensive relationship networks in which technical systems are embedded. In accordance with that, it turns out that precisely the reports of AI Now (Crawford et al. 2016, 2019; Whittaker et al. 2018; Campolo et al. 2017), an organization primarily led by women, do not conceive AI applications in isolation, but within a larger network of social and ecological dependencies and relationships

(Crawford and Joler 2018), corresponding most closely with the ideas and tenets of an ethics of care (Held 2013).

What are further insights from my analysis of the ethics guidelines, as summarized in Table 1? On the one hand, it is noticeable that guidelines from industrial contexts name on average 9.1 distinctly separated ethical aspects, whereas the average for ethics codes from science is 10.8. The principles of Microsoft's AI ethics are the most brief and minimalistic (Microsoft Corporation 2019). The OpenAI Charta names only four points and is thus situated at the bottom of the list (OpenAI 2018). Conversely, the IEEE guideline contains the largest volume with more than 100.000 words (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019). Finally, yet importantly, it is noteworthy that almost all guidelines suggest that technical solutions exist for many of the problems described. Nevertheless, there are only two guidelines which contain genuinely technical explanations at all—albeit only very sparsely. The authors of the guideline on the "Malicious Use of AI" provide the most extensive commentary here (Brundage et al. 2018).

2.3 Omissions

Despite the fact that the guidelines contain various parallels and several recurring topics, what are issues the guidelines do not discuss at all or only very occasionally? Here, I want to give a (non-exhaustive) overview of issues that are missing. Two things should be considered in this context. First, the sampling method used to select the AI ethics guidelines has an effect on the list of issues and omissions. When deliberately excluding for instance robot ethics guidelines, this has the effect that the list of entries lacks issues that are connected with robotics. Second, not all omissions can be treated equally. There are omissions which are missing or severely underrepresented without any good reason—for instance the aspect of political abuse or "hidden" social and ecological costs of AI systems—, and omissions that can be justified—for instance deliberations on artificial general intelligence or machine consciousness, since those technologies are purely speculative.

Nevertheless, in view of the fact that significant parts of the AI community see the emergence of *artificial general intelligence* as well as associated *dangers for humanity* or *existential threats* as a likely scenario (Müller and Bostrom 2016; Bostrom 2014; Tegmark 2017; Omohundro 2014), one could argue that those topics could be discussed in ethics guidelines under the umbrella of potential prohibitions to pursue certain research strands in this area (Hagendorff 2019). The fact that artificial general intelligence is not discussed in the guidelines may be due to the fact that most of the guidelines are not written by research groups from philosophy or other speculative disciplines, but by researchers with a background directly in computer science or its application. In this context, it is noteworthy that the fear of the emergence of superintelligence is more frequently expressed by people who lack technical experience in the field of AI—one just has to think of people like Stephen Hawking, Elon Musk or Bill Gates—while "real" experts generally regard the idea of a strong AI as rather absurd (Calo 2017, 26). Perhaps the same holds true for the question of *machine consciousness* and the ethical problems associated with

it (Lyons 2018), as this topic is also omitted from all examined ethical guidelines. What is also striking is the fact that only the Montréal Declaration for Responsible Development of Artificial Intelligence (2018) as well as the AI Now 2019 Report (2019) explicitly addresses the aspect of democratic control, governance and political deliberation of AI systems. The mentioned documents are also the only guidelines that explicitly prohibits imposing certain lifestyles or concepts of "good living" on people by AI systems, as it is for example demonstrated in the Chinese scoring system (Engelmann et al. 2019). The former document further criticizes the application of AI systems for the reduction of *social cohesion*, for example by isolating people in echo chambers (Flaxman et al. 2016). In addition, hardly any guideline discusses the possibility for *political abuse of AI systems* in the context of automated propaganda, bots, fake news, deepfakes, micro targeting, election fraud, and the like. What is also largely absent from most guidelines is the issue of a *lack in diversity* within the AI community. This lack of diversity is prevailing in the field of artificial intelligence research and development, as well as in the workplace cultures shaping the technology industry. In the end, a relatively small group of predominantly white men determines how AI systems are designed, for what purposes they are optimized, what is attempted to realize technically, etc. The famous AI startup "nnaisense" run by Jürgen Schmidhuber, which aims at generating an artificial general intelligence, to name just one example, employs only two women—one scientist and one office manager—in its team, but 21 men. Another matter, which is not covered at all or only very rarely mentioned in the guidelines, are aspects of *robot ethics*. As mentioned in the methods chapter, specific guidelines for robot ethics exist, most prominently represented by Asimov's three laws of robotics (Asimov 2004), but those guidelines were intentionally excluded from the analysis. Nonetheless, advances in AI research contribute, for instance, to increasingly anthropomorphized technical devices. The ethical question that arises in this context echoes Immanuel Kant's "brutalization argument" and states that the abuse of anthropomorphized agents—as, for example, is the case with language assistants (Brahnam 2006)—also promotes the likelihood of violent actions between people (Darling 2016). Apart from that, the examined ethics guidelines pay little attention to the rather popular *trolley problems* (Awad et al. 2018) and their alleged relation to ethical questions surrounding self-driving cars or other autonomous vehicles. In connection to this, no guideline deals in detail with the obvious question where systems of *algorithmic decision making* are superior or inferior, respectively, to human decision routines. And finally, virtually no guideline deals with the *"hidden" social and ecological costs* of AI systems. At several points in the guidelines, the importance of AI systems for approaching a sustainable society is emphasized (Rolnick et al. 2019). However, it is omitted—with the exception of the AI Now 2019 Report (2019)—that producer and consumer practices in the context of AI technologies may in themselves contradict sustainability goals. Issues such as lithium mining, e-waste, the one-way use of rare earth minerals, energy consumption, low-wage "clickworkers" creating labels for data sets or doing content moderation are of relevance here (Crawford and Joler 2018; Irani 2016; Veglis 2014; Fang 2019; Casilli 2017). Although "clickwork" is a necessary prerequisite for the application of methods of supervised machine learning, it is associated with numerous social problems (Silberman et al. 2018; Irani

2015; Graham et al. 2017), such as low wages, work conditions and psychological work consequences, which tend to be ignored by the AI community. Finally, yet importantly, not a single guideline raises the issue of *public–private partnerships* and *industry-funded research* in the field of AI. Despite the massive lack of transparency regarding the allocation of research funds, it is no secret that large parts of university AI research are financed by corporate partners. In light of this, it remains questionable to what extent the ideal of freedom of research can be upheld—or whether there will be a gradual "buyout" of research institutes.

3 AI in Practice

3.1 Business Versus Ethics

The close link between business and science is not only revealed by the fact that all of the major AI conferences are sponsored by industry partners. The link between business and science is also well illustrated by the AI Index 2018 (Shoham et al. 2018). Statistics show that, for example, the number of corporate-affiliated AI papers has grown significantly in recent years. Furthermore, there is a huge growth in the number of active AI startups, each supported by huge amounts of annual funding from Venture Capital firms. Tens of thousands of AI-related patents are registered each year. Different industries are incorporating AI applications in a broad variety of fields, ranging from manufacturing, supply-chain management, and service development, to marketing and risk assessment. All in all, the global AI market comprises more than 7 billion dollars (Wiggers 2019).

A critical look at this global AI market and the use of AI systems in the economy and other social systems sheds light primarily on unwanted side effects of the use of AI, as well as on directly malevolent contexts of use. These occur in various areas (Pistono and Yampolskiy 2016; Amodei et al. 2017). Leading, of course, is the military use of AI in cyber warfare or regarding weaponized unmanned vehicles or drones (Ernest and Carroll 2016; Anderson and Waxman 2013). According to media reports, the US government alone intends to invest two billion dollars in military AI projects over the next 5 years (Fryer-Biggs 2018). Moreover, governments can use AI applications for automated propaganda and disinformation campaigns (Lazer et al. 2018), social control (Engelmann et al. 2019), surveillance (Helbing 2019), face recognition or sentiment analysis (Introna and Wood 2004), social sorting (Lyon 2003), or improved interrogation techniques (McAllister 2017). Notwithstanding the above, companies can cause massive job losses due to AI implementation (Frey and Osborne 2013), conduct unmonitored forms of AI experiments on society without informed consent (Kramer et al. 2014), suffer from data breaches (Schneier 2018), use unfair, biased algorithms (Eubanks 2018), provide unsafe AI products (Sitawarin et al. 2018), use trade secrets to disguise harmful or flawed AI functionalities (Whittaker et al. 2018), rush to integrate and put immature AI applications on the market and many more. Furthermore, criminal or black-hat hackers can use AI to tailor cyberattacks, steal information, attack IT infrastructures, rig elections, spread misinformation for example through deepfakes, use voice synthesis

technologies for fraud or social engineering (Bendel 2017), or disclose personal traits that are actually secret or private via machine learning applications (Kosinski and Wang 2018; Kosinski et al. 2013, 2015). All in all, only a very small number of papers is published about the misuse of AI systems, even though they impressively show what massive damage can be done with those systems (Brundage et al. 2018; King et al. 2019; O'Neil 2016).

3.2 AI Race

While the United States currently has the largest number of start-ups, China claims to be the "world leader in AI" in 2030 (Abacus 2018). This claim is supported by the sheer amount of data that China has at its disposal to train its own AI systems, as well as by the large label companies that take over the manual preparation of data sets for supervised machine learning (Yuan 2018). Conversely, China is seen to have a weakness vis-à-vis the USA in that the investments of the market leaders Baidu, Alibaba and Tencent are too application-oriented comprising areas such as autonomous driving, finance or home appliances, while important basic research on algorithm development, chip production or sensor technology is neglected (Hao 2019). The constant comparison between China, the USA and Europe renders the fear of being inferior to each other an essential motive for efforts in the research and development of artificial intelligence.

Another justification for competitive thinking is provided by the military context. If the own "team", framed in a nationalist way, does not keep pace, so the consideration, it will simply be overrun by the opposing "team" with superior AI military technology. In fact, potential risks emerge from the AI race narrative, as well as from an actual competitive race to develop AI systems for technological superiority (Cave and ÓhÉigeartaigh 2018). One risk of this rhetoric is that "impediments" in the form of ethical considerations will be eliminated completely from research, development and implementation. AI research is not framed as a cooperative global project, but as a fierce competition. This competition affects the actions of individuals and promotes a climate of recklessness, repression, and thinking in hierarchies, victory and defeat. The race for the best AI, whether a mere narrative or a harsh reality, reduces the likelihood of the establishment of technical precaution measures as well as of the development of benevolent AI systems, cooperation, and dialogue between research groups and companies. Thus, the AI race stands in stark contrast to the idea of developing an "AI4people" (Floridi et al. 2018). The same holds true for the idea of an "AI for Global Good", as was proposed at the 2017's ITU summit, or the large number of leading AI researchers who signed the open letter of the "Future of Life Institute", embracing the norm that AI should be used for prosocial purposes.

Despite the downsides, in less public discourses and in concrete practice, an AI race has long since established itself. Along with that development, in- and outgroup-thinking has intensified. Competitors are seen more or less as enemies or at least as threats against which one has to defend oneself. Ethics, on the other hand, in its considerations and theories always stresses the danger of an artificial differentiation between in- and outgroups (Derrida 1997). Constructed outgroups are subject

to devaluation, are perceived de-individualized and in the worst case can become victims of violence simply because of their status as "others" (Mullen and Hu 1989; Vaes et al. 2014). I argue that only by abandoning such thinking in- and outgroups may the AI race be reframed into a global cooperation for beneficial and safe AI.

3.3 Ethics in Practice

Do ethical guidelines bring about a change in individual decision-making regardless of the larger social context? In a recent controlled study, researchers critically reviewed the idea that ethical guidelines serve as a basis for ethical decision-making for software engineers (McNamara et al. 2018). In brief, their main finding was that the effectiveness of guidelines or ethical codes is almost zero and that they do not change the behavior of professionals from the tech community. In the survey, 63 software engineering students and 105 professional software developers were scrutinized. They were presented with eleven software-related ethical decision scenarios, testing whether the influence of the ethics guideline of the Association for Computing Machinery (ACM) (Gotterbarn et al. 2018) in fact influences ethical decision-making in six vignettes, ranging from responsibility to report, user data collection, intellectual property, code quality, honesty to customer to time and personnel management. The results are disillusioning: "No statistically significant difference in the responses for any vignette were found across individuals who did and did not see the code of ethics, either for students or for professionals." (McNamara et al. 2018, 4).

Irrespective of such considerations on the microsociological level, the relative ineffectiveness of ethics can also be explained at the macrosociological level. Countless companies are eager to monetize AI in a huge variety of applications. This strive for a profitable use of machine learning systems is not primarily framed by value- or principle-based ethics, but obviously by an economic logic. Engineers and developers are neither systematically educated about ethical issues, nor are they empowered, for example by organizational structures, to raise ethical concerns. In business contexts, speed is everything in many cases and skipping ethical considerations is equivalent to the path of least resistance. Thus, the practice of development, implementation and use of AI applications has very often little to do with the values and principles postulated by ethics. The German sociologist Ulrich Beck once stated that ethics nowadays "plays the role of a bicycle brake on an intercontinental airplane" (Beck 1988, 194). This metaphor proves to be particularly true in the context of AI, where huge sums of money are invested in the development and commercial utilization of systems based on machine learning (Rosenberg 2017), while ethical considerations are mainly used for public relations purposes (Boddington 2017, 56).

In their AI Now 2017 Report, Kate Crawford and her team state that ethics and forms of soft governance "face real challenges" (Campolo et al. 2017, 5). This is mainly due to the fact that ethics has no enforcement mechanisms reaching beyond a voluntary and non-binding cooperation between ethicists and individuals working in research and industry. So what happens is that AI research and development takes place in "closed-door industry settings", where "user consent, privacy and transparency are often overlooked in favor of frictionless functionality that supports profit-driven business models"

(Campolo et al. 2017, 31 f.). Despite this dispensation of ethical principles, AI systems are used in areas of high societal significance such as health, police, mobility or education. Thus, in the AI Now Report 2018, it is repeated that the AI industry "urgently needs new approaches to governance", since, "internal governance structures at most technology companies are failing to ensure accountability for AI systems" (Whittaker et al. 2018, 4). Thus, ethics guidelines often fall into the category of a "'trust us' form of [non-binding] corporate self-governance" (Whittaker et al. 2018, 30) and people should "be wary of relying on companies to implement ethical practices voluntarily" (Whittaker et al. 2018, 32).

The tension between ethical principles and wider societal interests on the one hand, and research, industry, and business objectives on the other can be explained with recourse to sociological theories. Especially on the basis of system theory it can be shown that modern societies differ in their social systems, each working with their own codes and communication media (Luhmann 1984, 1997, 1988). Structural couplings can lead decisions in one social system to influence other social systems. Such couplings, however, are limited and do not change the overall autonomy of social systems. This autonomy, which must be understood as an exclusive, functionalist orientation towards the system's own codes is also manifested in the AI industry, business and science. All these systems have their own codes, their own target values, and their own types of economic or symbolic capital via which they are structured and based upon which decisions are made (Bourdieu 1984). Ethical intervention in those systems is only possible to a very limited extent (Hagendorff 2016). A certain hesitance exists towards every kind of intervention as long as these lie beyond the functional laws of the respective systems. Despite that, unethical behavior or unethical intentions are not solely caused by economic incentives. Rather, individual character traits like cognitive moral development, idealism, or job satisfaction play a role, let alone organizational environment characteristics like an egoistic work climate or (non-existent) mechanisms for the enforcement of ethical codes (Kish-Gephart et al. 2010). Nevertheless, many of these factors are heavily influenced by the overall economic system logic. Ethics is then, so to speak, "operationally effectless" (Luhmann 2008).

And yet, such system-theoretical considerations apply only on a macro level of observation and must not be generalized. Deviations from purely economic behavioral logics in the tech industry occur as well, for example when Google withdrew from the military project "Maven" after protests from employees (Statt 2018) or when people at Microsoft protested against the company's cooperation with Immigration and Customs Enforcement (ICE) (Lecher 2018). Nevertheless, it must also be kept in mind here that, in addition to genuine ethical motives, the significance of economically relevant reputation losses should not be underestimated. Hence, the protest against unethical AI projects can in turn be interpreted in an economic logic, too.

3.4 Loyalty to Guidelines

As indicated in the previous sections, the practice of using AI systems is poor in terms of compliance with the principles set out in the various ethical guidelines. Great progress has been made in the areas of privacy, fairness or explainability. For example,

many privacy-friendly techniques for the use of data sets and learning algorithms have been developed, using methods where AI systems' "sight" is "darkened" via cryptography, differential or stochastic privacy (Ekstrand et al. 2018; Baron and Musolesi 2017; Duchi et al. 2013; Singla et al. 2014). Nevertheless, this contradicts the observation that AI has been making such massive progress for several years precisely because of the large amounts of (personal) data available. Those data are collected by privacy-invasive social media platforms, smartphone apps, as well as Internet of Things devices with its countless sensors. In the end, I would argue that the current AI boom coincides with the emergence of a post-privacy society. In many respects, however, this post-privacy society is also a black box society (Pasquale 2015), in which, despite technical and organizational efforts to improve explainability, transparency and accountability, massive zones of non-transparency remain, caused both by the sheer complexity of technological systems and by strategic organizational decisions.

For many of the issues mentioned in the guidelines, it is difficult to assess the extent to which efforts to meet the set objectives are successful or whether conflicting trends prevail. This is the case in the areas of safety and cybersecurity, the science-policy link, future of employment, public awareness about AI risks, or human oversight. In other areas, including the issue of hidden costs and sustainability, the protection of whistleblowers, diversity in the field of AI, the fostering of solidarity and social cohesion, the respect for human autonomy, the use of AI for the common good or the military AI arms race, it can certainly be stated that the ethical goals are being massively underachieved. One only has to think of the aspect of gender diversity: Even though ethical guidelines clearly demand its improvement, the state of affairs is that on average 80% of the professors at the world's leading universities such as Stanford, Oxford, Berkeley or the ETH are male (Shoham et al. 2018). Furthermore, men make up more than 70% of applicants for AI jobs in the U.S. (Shoham et al. 2018). Alternatively, one can take human autonomy: As repeatedly demanded in various ethical guidelines, people should not be treated as mere data subjects, but as individuals. In fact, however, countless examples show that computer decisions, regardless of their susceptibility to error, are attributed a strong authority which results in the ignorance of individual circumstances and fates (Eubanks 2018). Furthermore, countless companies strive for the opposite of human autonomy, employing more and more subtle techniques for manipulating user behavior via micro targeting, nudging, UX-design and so on (Fogg 2003; Matz et al. 2017). Another example is that of cohesion: Many of the major scandals of the last years would have been unthinkable without the use of AI. From echo chamber effects (Pariser 2011) to the use of propaganda bots (Howard and Kollanyi 2016), or the spread of fake-news (Vosoughi et al. 2018), AI always played a key role to the effect of diminishing social cohesion, fostering instead radicalization, the decline of reason in public discourse and social divides (Tufekci 2018; Brady et al. 2017).

4 Advances in AI Ethics

4.1 Technical Instructions

Given the relative lack of tangible impact of the normative objectives set out in the guidelines, the question arises as to how the guidelines could be improved to make them more effective. At first glance, the most obvious potential for improvement of the guidelines is probably to supplement them with more detailed technical explanations—if such explanations can be found. Ultimately, it is a major problem to deduce concrete technological implementations from the very abstract ethical values and principles. What does it mean to implement justice or transparency in AI-systems? What does a "human-centered" AI look like? How can human oversight be obtained? The list of questions could easily be continued.

The ethics guidelines examined refer exclusively to the term "AI". They never or very seldom use more specific terminology. However, "AI" is just a collective term for a wide range of technologies or an abstract large-scale phenomenon. The fact that not a single prominent ethical guideline goes into greater technical detail shows how deep the gap is between concrete contexts of research, development, and application on the one side, and ethical thinking on the other. Ethicists must partly be capable of grasping technical details with their intellectual framework. That means reflecting on the ways data are generated, recorded, curated, processed, disseminated, shared, and used (Bruin and Floridi 2017), on the ways of designing algorithms and code, respectively (Kitchin 2017; Kitchin and Dodge 2011), or on the ways training data sets are selected (Gebru et al. 2018). In order to analyze all this in sufficient depth, ethics has to partially transform to "microethics". This means that at certain points, a substantial change in the level of abstraction has to happen insofar as ethics aims to have a certain impact and influence in the technical disciplines and the practice of research and development of artificial intelligence (Morley et al. 2019). On the way from ethics to "microethics", a transformation from ethics to technology ethics, to machine ethics, to computer ethics, to information ethics, to data ethics has to take place. As long as ethicists refrain from doing so, they will remain visible in a general public, but not in professional communities.

A good example of such a microethical work which can be implemented easily and concretely in practice is the paper by Gebru et al. (2018). The researchers propose the introduction of standardized datasheets listing the properties of different training data sets, so that machine learning-practitioners can check to what extent certain data sets are best suitable for their purposes, what the original intention was when the data set was created, what data the data set is composed of, how the data was collected and pre-processed, etc. The paper by Gebru et al. makes it possible for practitioners to obtain a more informed decision on the selection of certain training data sets, so that supervised machine learning ultimately becomes fairer, and more transparent, and avoids cases of algorithmic discrimination (Buolamwini and Gebru 2018). Such work is, however, an exception.

In general, ethical guidelines postulate very broad, overarching principles which are then supposed to be implemented in a widely diversified set of

scientific, technical and economic practices, and in sometimes geographically dispersed groups of researchers and developers with different priorities, tasks and fragmental responsibilities. Ethics thus operates at a maximum distance from the practices it actually seeks to govern. Of course, this does not remain unnoticed among technology developers. In consequence, the generality and superficiality of ethical guidelines in many cases not only prevents actors from bringing their own practice into line with them, but rather encourages the devolution of ethical responsibility to others.

4.2 Virtue Ethics

Regardless of the fact that normative guidelines should be accompanied by in-depth technical instructions—as far as they can reasonably be identified—, the question still arises how the precarious situation regarding the application and fulfillment of AI ethics guidelines can be improved. To address this question, one needs to take a step back and look at ethical theories in general. In ethics, several major strands of theories were created and shaped by various philosophical traditions. Those theories range from deontological to contractualistic, utilitarian, or virtue ethical approaches (Kant 1827; Rawls 1975; Bentham 1838; Hursthouse 2001). In the following, two of these approaches—deontology and virtue ethics—will be selected to illustrate different approaches in AI ethics. The deontological approach is based on strict rules, duties or imperatives. The virtue ethics approach, on the other hand, is based on character dispositions, moral intuitions or virtues—especially "technomoral virtues" (Vallor 2016). In the light of these two approaches, the traditional type of AI ethics can be assigned to the deontological concept (Mittelstadt 2019). Ethics guidelines postulate a fixed set of universal principles and maxims which technology developers should adhere to (Ananny 2016). The virtue ethics approach, on the other hand, focuses more on "deeper-lying" structures and situation-specific deliberations, on addressing personality traits and behavioral dispositions on the part of technology developers (Leonelli 2016). Virtue ethics does not define codes of conduct but focusses on the individual level. The technologists or software engineers and their social context are the primary addressees of such an ethics (Ananny 2016), not technology itself.

I argue that the prevalent approach of deontological AI ethics should be augmented with an approach oriented towards virtue ethics aiming at values and character dispositions. Ethics is then no longer understood as a deontologically inspired tick-box exercise, but as a project of advancing personalities, changing attitudes, strengthen responsibilities and gaining courage to refrain from certain actions, which are deemed unethical. When following the path of virtue ethics, ethics as a scientific discipline must refrain from wanting to limit, control, or steer (Luke 1995). Very often, ethics or ethical guidelines are perceived as something whose purpose is to stop or prohibit activity, to hamper valuable research and economic endeavors (Boddington 2017, 8). I want to resign this negative notion of ethics. It should not be the objective of ethics to stifle activity, but to do the exact opposite, i.e. broadening

the scope of action, uncovering blind spots, promoting autonomy and freedom, and fostering self-responsibility.

In view of AI ethics, approaches that focus on virtues aim at cultivating a moral character, expressing technomoral virtues such as honesty, justice, courage, empathy, care, civility, or magnanimity, to name just a few (Vallor 2016). Those virtues are supposed to raise the likelihood of ethical decision-making practices in organizations that develop and deploy AI applications. Cultivating a moral character, in terms of virtue ethics, means to educate virtues in families, schools, communities, as well as companies. At best, every individual, every member of a society should encourage this cultivation, by generating the motivation to adopt and habituate practices that influence technology development and use in a positive manner. Especially the subject of responsibility diffusion can only be circumvented when virtue ethics is adopted on a broad and collective level in communities of tech professionals. Simply every person involved in data science, data engineering and data economies related to applications of AI has to take at least some responsibility for the implications of their actions (Leonelli 2016). This is why researchers such as Floridi argue that every actor who is causally relevant for bringing about the collective consequence or impacts in question, has to be held accountable (Floridi 2016). Interestingly, Floridi uses the backpropagation method known from Deep Learning to describe the way in which responsibilities can be assigned, except that here backpropagation is used in networks of distributed responsibility. When working in groups, actions that are on first glance allegedly morally neutral can nevertheless have consequences or impacts—intended or non-intended—that are morally wrong. This means that practitioners from AI communities always need to discern the overarching, short- and long-term consequences of the technical artefacts they are building or maintaining, as well as to explore alternative ways of developing software or using data, including the option of completely refraining from carrying out particular tasks, which are considered unethical.

In addition to the endorsement of virtue ethics in tech communities, several institutional changes should take place. They include the adoption of legal framework conditions, the establishment of mechanisms for an independent auditing of technologies, the establishment of institutions for complaints, which also compensate for harms caused by AI systems, and the expansion of university curricula in particular through content from ethics of technology, media, and information (Floridi et al. 2018; Cowls and Floridi 2018; Eaton et al. 2017; Goldsmith and Burton 2017). So far, however, hardly any of these demands have been met.

5 Conclusion

Currently, AI ethics is failing in many cases. Ethics lacks a reinforcement mechanism. Deviations from the various codes of ethics have no consequences. And in cases where ethics is integrated into institutions, it mainly serves as a marketing strategy. Furthermore, empirical experiments show that reading ethics guidelines has no significant influence on the decision-making of software developers. In practice, AI ethics is often considered as extraneous, as surplus or some kind

of "add-on" to technical concerns, as unbinding framework that is imposed from institutions "outside" of the technical community. Distributed responsibility in conjunction with a lack of knowledge about long-term or broader societal technological consequences causes software developers to lack a feeling of accountability or a view of the moral significance of their work. Especially economic incentives are easily overriding commitment to ethical principles and values. This implies that the purposes for which AI systems are developed and applied are not in accordance with societal values or fundamental rights such as beneficence, non-maleficence, justice, and explicability (Taddeo and Floridi 2018; Pekka et al. 2018).

Nevertheless, in several areas ethically motivated efforts are undertaken to improve AI systems. This is particularly the case in fields where technical "fixes" can be found for specific problems, such as accountability, privacy protection, anti-discrimination, safety, or explainability. However, there is also a wide range of ethical aspects that are significantly related to the research, development and application of AI systems, but are not or very seldomly mentioned in the guidelines. Those omissions range from aspects like the danger of a malevolent artificial general intelligence, machine consciousness, the reduction of social cohesion by AI ranking and filtering systems on social networking sites, the political abuse of AI systems, a lack of diversity in the AI community, links to robot ethics, the dealing with trolley problems, the weighting between algorithmic or human decision routines, "hidden" social and ecological costs of AI, to the problem of public–private-partnerships and industry-funded research. Again, as mentioned earlier, the list of omissions is not exhaustive and not all omissions can be justified equally. Some omissions, like deliberations on artificial general intelligence, can be justified by pointing at their purely speculative nature, while other omissions are less valid and should be a reason to update or improve existing and upcoming guidelines.

Checkbox guidelines must not be the only "instruments" of AI ethics. A transition is required from a more deontologically oriented, action-restricting ethic based on universal abidance of principles and rules, to a situation-sensitive ethical approach based on virtues and personality dispositions, knowledge expansions, responsible autonomy and freedom of action. Such an AI ethics does not seek to subsume as many cases as possible under individual principles in an overgeneralizing way, but behaves sensitively towards individual situations and specific technical assemblages. Further, AI ethics should not try to discipline moral actors to adhere to normative principles, but emancipate them from potential inabilities to act self-responsibly on the basis of comprehensive knowledge, as well as empathy in situations where morally relevant decisions have to be made.

These considerations have two consequences for AI ethics. On the one hand, a stronger focus on technological details of the various methods and technologies in the field of AI and machine learning is required. This should ultimately serve to close the gap between ethics and technical discourses. It is necessary to build tangible bridges between abstract values and technical implementations, as long as these bridges can be reasonably constructed. On the other hand, however, the consequence of the presented considerations is that AI ethics, conversely, turns away from the description of purely technological phenomena in order to focus more strongly on genuinely social and personality-related aspects. AI ethics then deals less with AI

as such, than with ways of deviation or distancing oneself from problematic routines of action, with uncovering blind spots in knowledge, and of gaining individual self-responsibility. Future AI ethics faces the challenge of achieving this balancing act between the two approaches.

Acknowledgements Open Access funding provided by Projekt DEAL.

Funding This research was supported by the Cluster of Excellence "Machine Learning – New Perspectives for Science" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – Reference Number EXC 2064/1 – Project ID 390727645.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abacus. (2018). China internet report 2018. Retrieved July 13, 2018. https://www.abacusnews.com/china-internet-report/china-internet-2018.pdf.

Abrassart, C., Bengio, Y., Chicoisne, G., de Marcellis-Warin, N., Dilhac, M.-A., Gambs, S., Gautrais, V., et al. (2018). *Montréal declaration for responsible development of artificial intelligence* (pp. 1–21).

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2017). Concrete problems in AI safety. *arXiv* (pp. 1–29).

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values, 41*(1), 93–117.

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge: Cambridge University Press.

Anderson, M., Anderson, S. L. (2015). Towards ensuring ethical behavior from autonomous systems: A case-supported principle-based paradigm. In *Artificial intelligence and ethics: Papers from the 2015 AAAI Workshop* (pp. 1–10).

Anderson, D., Bonaguro, J., McKinney, M., Nicklin, A., Wiseman, J. (2018). *Ethics & algorithms toolkit*. Retrieved February 01, 2019. https://ethicstoolkit.ai/.

Anderson, K., Waxman, M. C. (2013). Law and ethics for autonomous weapon systems: Why a ban won't work and how the laws of WAR can. *SSRN Journal*, 1–32.

Asimov, I. (2004). *I, Robot*. New York: Random House LLC.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Bakewell, J. D., Clement-Jones, T. F., Giddens, A., Grender, R. M., Hollick, C. R., Holmes, C., Levene, P. K. et al. (2018). *AI in the UK: Ready, willing and able?*. Select committee on artificial intelligence (pp. 1–183).

Baron, B., Musolesi, M. (2017). Interpretable machine learning for privacy-preserving pervasive systems. *arXiv* (pp. 1–10).

Beck, U. (1988). *Gegengifte: Die organisierte Unverantwortlichkeit*. Frankfurt am Main: Suhrkamp.

Beijing Academy of Artificial Intelligence. (2019). *Beijing AI principles*. Retrieved June 18, 2019. https://www.baai.ac.cn/blog/beijing-ai-principles.

Bendel, O. (2017). The synthetization of human voices. *AI & SOCIETY - Journal of Knowledge, Culture and Communication, 82,* 737.

Bentham, J. (1838). *The Works of Jeremy Bentham*. With the assistance of J. Bowring. 11 vols. 1. Edinburgh: William Tait. Published under the Superintendence of his Executor.
Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Cham: Springer.
Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.
Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Cambridge: Harvard University Press.
Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proc Natl Acad Sci USA, 114*(28), 7313–7318.
Brahnam, S. (2006). Gendered bots and bot abuse. In Antonella de Angeli, Sheryl Brahnam, Peter Wallis, & Peter Dix (Eds.), *Misuse and abuse of interactive technologies* (pp. 1–4). Montreal: ACM.
Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A. et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv* (pp. 1–101).
Buolamwini, J., Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. In Sorelle and Wilson 2018 (pp. 1–15).
Burton, E., Goldsmith, J., Koening, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. *Artificial Intelligence Magazine, 38*(2), 22–36.
Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. *SSRN Journal*, 1–28.
Campolo, A., Sanfilippo, M., Whittaker, M., Crawford, K. (2017). *AI now 2017 report*. Retrieved October 02, 2018. https://assets.ctfassets.net/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf.
Casilli, A. A. (2017). Digital labor studies go global: Toward a digital decolonial turn. *International Journal of Communication, 11,* 1934–3954.
Cave, S., ÓhÉigeartaigh, S. S. (2018). *An AI race for strategic advantage: Rhetoric and risks* (pp. 1–5).
Cowls, J., Floridi, L., (2018). Prolegomena to a white paper on an ethical framework for a good AI society. *SSRN Journal*, 1–14.
Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A. et al. (2019). *AI now 2019 report*. Retrieved December 18, 2019. https://ainowinstitute.org/AI_Now_2019_Report.pdf.
Crawford, K., Joler, V. (2018). *Anatomy of an AI system*. Retrieved February 06, 2019. https://anatomyof.ai/.
Crawford, K., Whittaker, M., Clare Elish, M., Barocas, S., Plasek, A., Ferryman, K. (2016). *The AI now report: The social and economic implications of artificial intelligence technologies in the near-term*.
Cutler, A., Pribić, M., Humphrey, L. (2018). *Everyday ethics for artificial intelligence: A practical guide for designers & developers*. Retrieved February 04, 2019. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf: 1–18.
Darling, K. (2016). Extending legal protection to social robots: The effect of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. M. Froomkin, & I. Kerr (Eds.), *Robot law* (pp. 213–234). Cheltenham: Edward Elgar.
de Bruin, B., & Floridi, L. (2017). The ethics of cloud computing. *Science and Engineering Ethics, 23*(1), 21–39.
DeepMind. *DeepMind ethics & society principles.* Retrieved July 17, 2019. https://deepmind.com/applied/deepmind-ethics-society/principles/.
Derrida, J. (1997). *Of grammatology*. Baltimore: Johns Hopkins Univ. Press.
Diakopoulos, N., Friedler, S. A., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V. et al. Principles for accountable algorithms and a social impact statement for algorithms. Retrieved July 31, 2019. https://www.fatml.org/resources/principles-for-accountable-algorithms.
Duchi, J. C., Jordan, M. I., Wainwright, M. J. (2013). Privacy aware learning. *arXiv* (pp. 1–60).
Eaton, E., Koenig, S., Schulz, C., Maurelli, F., Lee, J., Eckroth, J., Crowley, M. et al. (2017). Blue sky ideas in artificial intelligence education from the EAAI 2017 new and future AI educator program. *arXiv* (pp. 1–5).
Eckersley, P. (2018). Impossibility and uncertainty theorems in AI value alignment or why your AGI should not have a utility function. *arXiv* (pp. 1–13).
Ekstrand, M. D., Joshaghani, R., Mehrpouyan, H. (2018). Privacy for all: Ensuring fair and equitable privacy protections. In Sorelle and Wilson 2018 (pp. 1–13).
Engelmann, S., Chen, M., Fischer, F., Kao, C., Grossklags, J. (2019). Clear sanctions, vague rewards: How China's social credit system currently defines "Good" and "Bad" behavior. In *Proceedings of the conference on fairness, accountability, and transparency—FAT* '19* (pp. 69–78).

Ernest, N., & Carroll, D. (2016). Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions. *Journal of Defense Management*. https://doi.org/10.4172/2167-0374.1000144.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics, 21*(4), 403–418.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Marting's Press.

Fang, L. (2019). *Google hired gig economy workers to improve artificial intelligence in controversial drone-targeting project*. Retrieved February 13, 2019. https://theintercept.com/2019/02/04/google-ai-project-maven-figure-eight/.

Fjeld, J., Hilligoss, H., Achten, N., Daniel, M. L., Feldman, J., Kagay, S. (2019). *Principled artificial intelligence: A map of ethical and rights-based approaches*. Retrieved July 17, 2019. https://ai-hr.cyber.harvard.edu/primp-viz.html.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *PUBOPQ, 80*(S1), 298–320.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 374*(2083), 1–13.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do*. San Francisco: Morgan Kaufmann Publishers.

Frey, C. B., Osborne, M. A. (2013). *The future of employment: How susceptible are jobs to computerisation:* Oxford Martin Programme on Technology and Employment (pp. 1–78).

Fryer-Biggs, Z. (2018). The pentagon plans to spend $2 billion to put more artificial intelligence into its weaponry. Retrieved January 25, 2019. https://www.theverge.com/2018/9/8/17833160/pentagon-darpa-artificial-intelligence-ai-investment.

Future of Life Institute. (2017). *Asilomar AI principles.* Retrieved October 23, 2018. https://futureoflife.org/ai-principles/.

Garzcarek, U., Steuer, D. (2019). Approaching ethical guidelines for data scientists. *arXiv* (pp. 1–18).

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé, III, H., Crawford, K. (2018). Datasheets for datasets. *arXiv* (pp. 1–17).

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge: Harvard University Press.

Goldsmith, J., Burton, E. (2017). *Why teaching ethics to AI practitioners is important. ACM SIGCAS Computers and Society* (pp. 110–114).

Google. (2018). *Artificial intelligence at Google: Our principles.* Retrieved January 24, 2019. https://ai.google/principles/.

Google. (2019). *Perspectives on issues in AI governance* (pp. 1–34). Retrieved February 11, 2019. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

Gotterbarn, D., Brinkman, B., Flick, C., Kirkpatrick, M. S., Miller, K., Vazansky, K., Wolf, M. J. (2018). *ACM code of ethics and professional conduct: Affirming our obligation to use our skills to benefit society* (pp. 1–28). Retrieved February 01, 2019. https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf.

Graham, M., Hjorth, I., & Lehdonvirta, V. (2017). Digital labour and development: Impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European Review of Labour and Research, 23*(2), 135–162.

Greene, D., Hoffman, A. L., Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Hawaii international conference on system sciences* (pp. 1–10).

Hagendorff, T. (2016). Wirksamkeitssteigerungen Gesellschaftskritischer Diskurse. *Soziale Probleme. Zeitschrift für soziale Probleme und soziale Kontrolle, 27*(1), 1–16.

Hagendorff, T. (2019). Forbidden knowledge in machine learning: Reflections on the limits of research and publication. *arXiv* (pp. 1–24).

Hao, K. (2019). Three charts show how China's AI Industry is propped up by three companies. Retrieved January 25, 2019. https://www.technologyreview.com/s/612813/the-future-of-chinas-ai-indus

try-is-in-the-hands-of-just-three-companies/?utm_campaign=Artificial%2BIntelligence%2BWeekly&utm_medium=email&utm_source=Artificial_Intelligence_Weekly_95.
Helbing, D. (Ed.). (2019). *Towards digital enlightment: Essays on the darf and light sides of the digital revolution*. Cham: Springer.
Held, V. (2013). Non-contractual society: A feminist view. *Canadian Journal of Philosophy, 17*(Supplementary Volume 13), 111–137.
Holdren, J. P., Bruce, A., Felten, E., Lyons, T., & Garris, M. (2016). *Preparing for the future of artificial intelligence* (pp. 1–58). Washington, D.C: Springer.
Howard, P. N., Kollanyi, B. (2016). Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU Referendum. *arXiv* (pp. 1–6).
Hursthouse, R. (2001). *On virtue ethics*. Oxford: Oxford University Press.
Information Technology Industry Council. (2017). *ITI AI policy principles*. Retrieved January 29, 2019. https://www.itic.org/public-policy/ITIAIPolicyPrinciplesFINAL.pdf.
Introna, L. D., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society, 2*(2/3), 177–198.
Irani, L. (2015). The cultural work of microwork. *New Media & Society, 17*(5), 720–739.
Irani, L. (2016). The hidden faces of automation. *XRDS, 23*(2), 34–37.
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
Johnson, D. G. (2017). Can engineering ethics be taught? *The Bridge, 47*(1), 59–64.
Kant, I. (1827). *Kritik Der Praktischen Vernunft*. Leipzig: Hartknoch.
King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2019). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics, 26,* 89–120.
Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *The Journal of Applied Psychology, 95*(1), 1–31.
Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14–29.
Kitchin, R., & Dodge, M. (2011). *Code/space: Software and everyday life*. Cambridge: The MIT Press.
Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70*(6), 543–556.
Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802–5805.
Kosinski, M., & Wang, Y. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*(2), 246–257.
Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8788–8790.
Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.
Lecher, C. (2018). The employee letter denouncing Microsoft's ICE contract now has over 300 signatures. Retrieved February 11, 2019. https://www.theverge.com/2018/6/21/17488328/microsoft-ice-employees-signatures-protest.
Leonelli, S. (2016). Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 374*(2083), 1–12.
Luhmann, N. (1984). *Soziale Systeme: Grundriß einer allgemeinen Theorie*. Frankfurt A.M: Suhrkamp.
Luhmann, N. (1988). *Die Wirtschaft der Gesellschaft*. Frankfurt A.M: Suhrkamp.
Luhmann, N. (1997). *Die Gesellschaft der Gesellschaft*. Frankfurt am Main: Suhrkamp.
Luhmann, N. (2008). *Die Moral der Gesellschaft*. Frankfurt AM: Suhrkamp.
Luke, B. (1995). Taming ourselves or going Feral? Toward a nonpatriarchal metaethic of animal liberation. In Carol J. Adams & Josephine Donovan (Eds.), *Animals & women: Feminist theoretical explorations* (pp. 290–319). Durham: Duke University Press.
Lyon, D. (2003). Surveillance as social sorting: Computer codes and mobile bodies. In David Lyon (Ed.), *Surveillance as social sorting: Privacy, risk, and digital discrimination* (pp. 13–30). London: Routledge.

Lyons, S. (2018). *Death and the machine*. Singapore: Palgrave Pivot.
Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America, 114,* 12714–12719.
McAllister, A. (2017). Stranger than science fiction: The rise of A.I. interrogation in the dawn of autonomous robots and the need for an additional protocol to the U.N. convention against torture. *Minnesota Law Review, 101,* 2527–2573.
McNamara, A., Smith, J., Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development?" In G. T. Leavens, A. Garcia, C. S. Păsăreanu (Eds.) *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering—ESEC/FSE 2018* (pp. 1–7). New York: ACM Press.
Microsoft Corporation. (2019). Microsoft AI principles. Retrieved February 01, 2019. https://www.microsoft.com/en-us/ai/our-approach-to-ai.
Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507.
Mittelstadt, B., Russell, C., Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency—FAT* '19* (pp. 1–10).
Morley, J., Floridi, L., Kinsey, L., Elhalal, A. (2019). From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. *arXiv* (pp. 1–21).
Mullen, B., & Hu, L.-T. (1989). Perceptions of ingroup and outgroup variability: A meta-analytic integration. *Basic and Applied Social Psychology, 10*(3), 233–252.
Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In Vincent C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 555–572). Cham: Springer International Publishing.
Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence, 26*(3), 303–315.
O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishers.
OpenAI. (2018). *OpenAI Charter*. Retrieved July 17, 2019. https://openai.com/charter/.
Organisation for Economic Co-operation and Development. (2019). *Recommendation of the Council on Artificial Intelligence* (pp. 1–12). Retrieved June 18, 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. New York: The Penguin Press.
Partnership on AI. (2018). *About us*. Retrieved January 25, 2019. https://www.partnershiponai.org/about/.
Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge: Harvard University Press.
Pekka, A.-P., Bauer, W., Bergmann, U., Bieliková, M., Bonefeld-Dahl, C., Bonnet, Y., Bouarfa, L. et al. (2018). *The European Commission's high-level expert group on artificial intelligence: Ethics guidelines for trustworthy ai*. Working Document for stakeholders' consultation. Brussels (pp. 1–37).
Pistono, F., Yampolskiy, R. (2016). Unethical research: How to create a malevolent artificial intelligence. *arXiv* (pp. 1–6).
Podgaiska, I., Shklovski, I. *Nordic engineers' stand on artificial intelligence and ethics: Policy recommendations and guidelines* (pp. 1–40).
Prates, M., Avelar, P., Lamb, L. C. (2018). On quantifying and understanding the role of ethics in AI research: A historical account of flagship conferences and journals. *arXiv* (pp. 1–13).
Rawls, J. (1975). *Eine Theorie Der Gerechtigkeit*. Frankfurt am Main: Suhrkamp.
Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S. et al. (2019). Tackling climate change with machine learning. *arXiv* (pp. 1–97).
Rosenberg, S. (2017) Why AI is still waiting for its ethics transplant."Retrieved January 16, 2018. https://www.wired.com/story/why-ai-is-still-waiting-for-its-ethics-transplant/.
Schneier, B. (2018). *Click here to kill everybody*. New York: W. W. Norton & Company.
Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., Vertesi, J. (2018). Fairness and abstraction in Sociotechnical Systems. In *ACT conference on fairness, accountability, and transparency (FAT)* (vol. 1, No. 1, pp. 1–17).
Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., Lyons, T., Etchemendy, J., Grosz, B., Bauer, Z. (2018). *The AI index 2018 annual report.* Stanford, Kalifornien (pp. 1–94).

Silberman, M. S., Tomlinson, B., LaPlante, R., Ross, J., Irani, L., & Zaldivar, A. (2018). Responsible research with crowds. *Communications of the ACM, 61*(3), 39–41.
Singla, A., Horvitz, E., Kamar, E., White, R. W. (2014). Stochastic Privacy. *arXiv* (pp. 1–10).
Sitawarin, C., Bhagoji, A. N., Mosenia, A., Chiang, M., Mittal, P. (2018). DARTS: Deceiving autonomous cars with toxic signs. *arXiv* (pp. 1–27).
Smart Dubai. 2018. *AI ethics principles & guidelines*. Retrieved February 01, 2019. https://smartdubai.ae/pdfviewer/web/viewer.html?file=https://smartdubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?Status=Master&sfvrsn=d4184f8d_6.
Statt, N. (2018). Google reportedly leaving project maven military AI program after 2019. Retrieved February 11, 2019. https://www.theverge.com/2018/6/1/17418406/google-maven-drone-imagery-ai-contract-expire.
Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.
Tegmark, A. (2017). *Life 3.0: Being human in the age of artificial intelligence*. New York: Alfred A. Knopf.
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2016). *Ethically aligned design: A vision for prioritizing human well-being with artificial intelligence and autonomous systems* (pp. 1–138).
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (pp. 1–294).
Tufekci, Z. (2018). YouTube, the great Radicalizer. Retrieved March 19, 2018. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.
Vaes, J., Bain, P. G., & Bastian, B. (2014). Embracing humanity in the face of death: why do existential concerns moderate ingroup humanization? *The Journal of Social Psychology, 154*(6), 537–545.
Vakkuri, V., Abrahamsson, P. (2018). The key concepts of ethics of artificial intelligence. In *Proceedings of the 2018 IEEE international conference on engineering, technology and innovation* (pp. 1–6).
Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. New York: Oxford University Press.
Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), 1–17.
Veglis, A. (2014). Moderation techniques for social media content. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, et al. (Eds.), *Social computing and social media* (pp. 137–148). Cham: Springer International Publishing.
Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.
Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., Schwartz, O. (2018). *AI now report 2018* (pp. 1–62).
Wiggers, K. (2019). CB insights: Here are the top 100 AI companies in the world. Retrieved February 11, 2019. https://venturebeat.com/2019/02/06/cb-insights-here-are-the-top-100-ai-companies-in-the-world/.
Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., Yang, Q. (2018). Building ethics into artificial intelligence. *arXiv* (pp. 1–8).
Yuan, L. (2018). *How cheap labor drives China's A.I. ambitions*. Retrieved November 30, 2018. https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html.
Zeng, Y., Lu, E., Huangfu, C. (2018). Linking artificial intelligence principles. *arXiv* (pp. 1–4).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.