AI and Ethics (2021) 1:41–47
https://doi.org/10.1007/s43681-020-00008-1

OPINION PAPER

Lessons learned from AI ethics principles for future actions

Merve Hickok[1]

Received: 29 August 2020 / Accepted: 2 September 2020 / Published online: 6 October 2020

Abstract
As the use of artificial intelligence (AI) systems became significantly more prevalent in recent years, the concerns on how these systems collect, use and process big data also increased. To address these concerns and advocate for ethical and responsible development and implementation of AI, non-governmental organizations (NGOs), research centers, private companies, and governmental agencies published more than 100 AI ethics principles and guidelines. This first wave was followed by a series of suggested frameworks, tools, and checklists that attempt a technical fix to issues brought up in the high-level principles. Principles are important to create a common understanding for priorities and are the groundwork for future governance and opportunities for innovation. However, a review of these documents based on their country of origin and funding entities shows that private companies from US-West axis dominate the conversation. Several cases surfaced in the meantime which demonstrate biased algorithms and their impact on individuals and society. The field of AI ethics is urgently calling for tangible action to move from high-level abstractions and conceptual arguments towards applying ethics in practice and creating accountability mechanisms. However, lessons must be learned from the shortcomings of AI ethics principles to ensure the future investments, collaborations, standards, codes or legislation reflect the diversity of voices and incorporate the experiences of those who are already impacted by the biased algorithms.

Keywords Artificial intelligence · AI and ethics · Applied ethics · Governance · Social justice · Diversity

1 Introduction

In the last few years advances in the capabilities of AI coupled with big data brought into focus the opportunities and risks of AI. These developments in response created a spike in the publication of AI ethics principles and guidelines from civil society organizations, research centers, private companies and governmental agencies as these parties made their commitments public, and positioned themselves on what they think are the most important values to embed in the development and implementation of AI products.

Principles and guidelines are intentionally provided as high-level, abstract documents as the application of them is case, time and context sensitive. They need to be applicable across multiple areas. These high-profile value statements in AI contribute to the formation of a moral background as they make the connection between values, ethics, and technologies explicit [1]. Principles are also treated as the guides on how the policymakers and professionals should prioritize and structure future legislation, standards, governance models, and investment. Soft and hard governance models, which include principles in former and regulation in latter, shape the power relations among private companies and governments and individuals, whether individuals are consumers or citizens. This makes the debate on principles crucial. However, their abstract nature also makes the principles less useful for practical purposes and for creating a common understanding of their meaning.

Algorithmic or automated decision-making systems are already used in every aspect of our private and public lives, from credit scoring to recruitment, college admissions to predictive policing, welfare benefit eligibility determinations to criminal justice decisions, online content personalization to military applications or border control. Algorithms are increasingly being marketed and adopted to improve efficiency, reduce costs, and enhance personalization of products and services. However, despite this ubiquitous state of algorithms and the high stakes and high impact of these decisions on the individual and society, the decisions about what is important, what should be priority or how and where these systems

✉ Merve Hickok
merve@lighthousecareerconsulting.com

[1] AIethicist.org, Ann Arbor, MI, USA

should be deployed are still made by a small number of stakeholders. These decisions reflect and shape the power relations among those who decide, and between those who decide and who are impacted. Even the field of AI ethics whose aim is to surface the ethical issues relevant to AI and suggest solutions is not immune to the forces of politics and power. The main criticisms with most of the principles and guidelines are that (1) they are too abstract to be practical, (2) they reflect mainly the values of the experts chosen to create them, (3) and they serve the priorities of the private entities which funded some of this work.

The main criticism with the technical proposals, such as tools and checklists, which followed the wave of principles are that (1) they try to bring technical fixes to structural issues; (2) they do not reflect on the fundamental social, racial, economical context of these issues; (3) they are not focused on governing the liability and responsibility regarding the impact of AI; and (4) they divert attention from asking the crucial questions of business culture and exploitation. As Ruha Benjamin cautions "the road to inequity is paved with technical fixes [2]."

This work argues that there are two priorities on which the actors involved in this debate should focus. First is to spotlight the shortcomings of AI ethics principles and make them explicit. Second is to ensure that future work builds up from lessons learned and has a truly inclusive and global voice. Tangible actions are urgently required to ensure that AI is ethically and responsibly developed and implemented and that relevant governance methods are in place. A deeper understanding of the current state concludes a more diverse and inclusive approach is necessary to counter the issues raised by AI. Focusing on the lessons learned does not impede the progress of moving principles to practice, in fact it opens future possibilities. It brings forward different AI cases, the consequences and allows for better responses. Communities shape the ways they will use AI and own the advanced technologies. However, this requires inclusion and mechanisms to contest, redress and reverse technological interventions if necessary. AI ethics principles so far failed to meaningfully engage with the people who would be the targets of those systems, cutting off these groups from ownership, inclusion, and justice [3] as part of the process.

The article asks the questions of (1) what is the current state of principles and guidelines in terms of representation? (2) What needs to change and how? It adds to the calls for an urgent action to move from principles to practices with a focus on the context in which AI is deployed and its impact. It also demands further diversity and inclusivity in the field.

2 Current state of ethical/responsible AI principles

A few online platforms [4–6] keep a current inventory of all AI ethics principles and guidelines published around the world, and several research papers have completed meta-analyses of these documents.

A review of these platforms, and published analysis research tells us that in terms of content, these documents focus on similar values and priorities, however, with different definitions. In their 2019 meta-analysis of "global" landscape of AI ethics guidelines, Jobin et al. [7] review 84 AI ethics guidelines publicly released. The findings suggest, although there was no single principle that was common in every 84 of these documents, over half of them included the themes of transparency, justice, non-maleficence, responsibility, and privacy. Additional themes that repeat in different guidelines include respect for human autonomy and dignity, explainability, beneficence, and trust. Fjeld et al. [8] 2020 study of a different sample of documents shows, while there are certainly points of convergence, by no means is there unanimity across the 36 principle documents reviewed. Fairness and non-discrimination principles were present in all documents in this review, and the other key themes were privacy, accountability, safety and security, transparency and explainability, human control of technology, professional responsibility, and promotion of human values. The convergence on similar principles, but differences in how each actor defines these same principles creates inconsistency and confusion in discussions. Similar terms mean different things, hence allowing for a variety of interpretation. This makes it harder to hold actors accountable to their commitments. It also allows for the possibility of "digital ethics shopping" where private and public actors mix and match ethical principles, guidelines, codes, frameworks for the kind of ethics that is best retrofitted to justify their current behaviors, rather than revising their behaviors by benchmarking them against public, ethical standards [9].

In terms of geographic distribution, Jobin et al. [10] data show a prominent representation of USA (25%) and the UK (15.5%), EU member states, Canada, Iceland, and Norway and their institutions (26%). These together account for 67% all AI ethics principles and show a heavy influence of US-West values. Japan is represented by (4.8%), and UAE, India, Singapore, South Korea, and Australia with one document each compromise a total of 6%. However, African, South and Central American and Central Asian countries (with exception of India) are not represented independently from international or supranational organizations (like World Economic Forum, G20, OECD, or UNESCO). China released its principles in 2019.

Although one could make an argument that the principles included in these analyses might have a language bias (for example in Jobin et al. analysis, only those written in English, German, French, Italian or Greek were selected; and in Fjeld et al. analysis those that were in English, Chinese, French, German, and Spanish.), then one needs to also accept there is then a barrier the field needs to correct and focus energy to ensure access to work in other language regions are accessible. The language barrier in accessing research is not a new obstacle. However, given AI's impact on all domains of life and its reach across borders, language remaining a barrier is unacceptable. It can be overcome through global partnerships, virtual connectivity and increased funding. In the absence of representation from these regions "more economically developed countries are shaping this debate more than others, which raises concerns about neglecting local knowledge, cultural pluralism and the demands of global fairness" [10]. The field needs to question how principles, guidelines, or practices can be 'global' if they do not include any ethical perspective, community involvement, or social and historical context from Africa, Latin America, or Central Asia. Mohamed et al. view the developments in data and AI practices from an algorithmic coloniality perspective and warn against the repeat of history. Practitioners are reminded to critically engage with that inheritance, to avert algorithmic colonialism and to reveal the relations of power that underlie technology deployment [11].

A breakdown of the AI principle documents by the entities that funded their creation shows that 23% were produced by private companies, 21% by governmental agencies, academic and research institutions then intergovernmental or supranational organizations making up 11% and 10% respectively, non-profit organizations and professional associations/scientific societies 8.3% each. The representation of private companies shows slightly similar, but just as concerning trends in different platforms. AI Ethics Lab shows that AI principles from private companies make up 34% of the 103 documents that inventoried in the platform [12]; Fjeld et al. [13] research shows private funding as 22% out of the 36 documents reviewed; and AlgoritmWatch's AI Ethics Guidelines Global Inventory [14] shows that private companies total to 26% of the 160 listed in their selection. These different reviews show principles from private companies make up for almost as many as those from governmental agencies. Guidelines for the use of artificial intelligence methods are not created in a vacuum, but are shaped by the decision-making processes, intentions, and limitations of the companies and organizations that develop and use them [14]. These efforts from private companies are criticized as their way of lobbying and marketing for self-governance and thus avoiding stricter regulations. "The word ethics is under siege in technology policy. Weaponized in support of deregulation, self-regulation, or hands-off governance, "ethics" is increasingly identified with technology companies' self-regulatory efforts and with shallow appearances of ethical behavior" [15]. Whittaker et al. caution that ethics guidelines often fall into the category of a "'trust us' form of corporate self-governance" and people should "be wary of relying on companies to implement ethical practices voluntarily" [16]. Concerns have also been raised that, as large parts of university AI research are financed by corporate partners, it might be questionable to what extent the ideal of freedom of research can be upheld [17]. Time and evidence will show if the rare call for regulation from these companies are genuine [18].

AI ethics principles in general and the technical solutions and checklists that followed mainly focus on how to improve these algorithmic systems. They rarely question the business culture, revenue models, or incentive mechanisms that continuously push these products into the markets. In their frame review of ethical AI principles, Greene et al. [19] confirm that ethical debate is largely limited to appropriate design and implementation and not whether these systems should be built in the first place. Business models and the culture of the developer companies are not positioned as needing the same level of scrutiny as design decisions. Rather than asking fundamental ethical and political questions about whether AI systems should be built, these documents implicitly frame technological progress as inevitable, calling for better building [20]. They also frame ethical design as a project of expert oversight and draw "a narrow circle of who can or should adjudicate ethical concerns around AI" on behalf of the rest of us [19]. The decisions rest within the power structures whether that is the government or the private companies.

When the principles and guidelines are analyzed in terms of the gender of the authors, the result also shows a skewed picture in favor of male dominance. Hagendorff's literature analysis of authors of 21 major ethics guidelines shows that the proportion of female authors was only 31% [21]. The numbers do not get any better when we expand our lens to diversity in AI academics, contributions to conferences, or number of AI job applicants or employees. On average, 80% of professors from UC Berkeley, Stanford, University of Illinois, Carnegie Mellon University, UC London, Oxford, and ETH Zurich are male [22]. Males made up 88% of contributions accepted to The Conference and Workshop on Neural Information Processing Systems (NeurIPS), The International Conference on Learning Representations (ICLR), and The International Conference on Machine Learning (ICML) in 2017 [23].

In 2018, the World Economic Forum identified the gender gap in the talent pool by surveying Linkedln users who self-identified as possessing AI skills. The results showed that 22% of them were female [24]. Women employed in the

software and IT services industry make up 7.4% of the AI talent pool [24]. Across the three countries in which AI talent is ranked as most prominent (the United States, India and Germany), share of female professionals with AI skills show as 23%, 22% and 16%, respectively [24]. When it comes to AI job applicants in US, men make up around 71% of the applicant pool. These analyses and research make a very strong case for the issue of dominance of one group. Unfortunately, they also assume a binary representation of gender and, therefore, stay blind to the contributions of non-binary scholars or employees.

Despite the claim of AI ethics principles and guidelines to be global and beneficial for all, they are still limited in their representation of values and perspectives from different regions, entities, and communities. The concern here is that if the field does not make any course correction, the same issues will be replicated in the future legislation, standards, investment, and education priorities.

3 What needs to change and how?

Principles are the outline of a shared foundation upon which one can build and use as a benchmark to communicate expectations and evaluate deliverables. Co-design in AI would be more difficult without this common framework [25]. Principles are a valuable part of any applied ethics and business ethics. They help to condense complex ethical issues into a few central elements which can allow widespread commitment to a shared set of values. However, without an understanding of why these issues and categories have been chosen and not others, it is difficult to be confident that all the relevant issues have been captured. Without the inclusion of different regions or members of society, it is not clear whose values and priorities are being promoted [26]. Understanding who sits at the table, what questions and concerns are sidelined and what power asymmetries are shaping the terms of debate is crucial [27]. As Sasha Costanza-Chock cautions "if you are not at the table, you are on the menu" [28].

Principles provide an informal means of holding people and organizations accountable and to reassure public concerns [29]. Although, in and of themselves, principles are not enough or as strong as regulations, widely accepted principles allow for citizens and consumers to hold the governments and private companies accountable for what they publicly announced as their guiding values. They provide the basis to challenge these entities to stay loyal to these values and demand action and change if they do not do as they preach. Different cultures will differ in what values they prioritize; however, understanding the same thing from a stated principle is crucial. Acceptance of the principles requires them to be less ambiguous in their definitions. Acceptance requires a consistency of understanding what they mean and how it should look like in practice. Finally, it also means that these principles are not just debated among experts but that they become a shared language across communities. What is currently taken for granted can be tested and recalibrated in the light of alternative and dissenting perspectives [30].

In expanding upon the lessons of the AI ethics principles and hence creating the roadmap for the next actions, the actors in the field, whether that is defined as NGOs, private companies, governments or research centers, must include the work from different regions and marginalized communities to be truly global. In doing so, these entities should also be careful about the possible dominance of technical experts. Even in public engagement settings, technical experts exert authority in framing the debate, knowledge production, and decision-making [31]. Special attention and intention need to be paid to listen to, amplify, cite, and collaborate with those who have lived experiences be truly inclusive [32].

A possible solution and approach to be adopted in future partnerships can be the outcome-oriented Design Justice method. This approach, for example, believes that everyone is an expert based on their own lived experiences, and tries to distribute the benefits and burdens of design fairly, work towards non-exploitative solutions, and ensure fair and meaningful participation in decisions [33]. Ethics enables organizations to identify and leverage new opportunities that are socially acceptable or preferable. It also enables them to anticipate and avoid or at least minimize costly mistakes [34]. Formal scientific training cannot be the only criterion by which to decide whose voices should be heard in an inclusive global forum [30].

To break some of the barriers and reach and engage additional representation, particularly from the Global South, UN Global Advisory Body on AI in the Secretary General's Roadmap on Digital Cooperation works to create an inclusive and informed foundation [35] for its future work. To that end the roundtable 3C tries to connect any interested party around the world under consultation calls and encourages as wide a response as possible. A similar approach can be used by a company to determine what AI products or services are adopted in different communities or regions and understand the priorities and values of these communities. Non-private entities can use this approach to determine what civil society, research and policy work is taking place by whom. The entities can then create research or advocacy partnerships with these groups and invest more in funding these partnerships. Through these interactions, certain barriers to entry into the global discussions might also be unearthed. Knowing what is happening in a region or community is just as important understanding what is not happening in that area. Together this knowledge helps actors in AI design, research or policy fields to remove barriers and create new spaces to raise awareness, advocate and include different stakeholders.

Just as crucial as a regional inclusivity, cross-disciplinary and cross-sectoral cooperation is a must to understand the perspectives and concerns of different actors. Therefore, it is important that policymakers, professional bodies, private companies, and universities incentivize financially cooperation and debate concerning the intersections between technology, social issues, legal studies, and ethics [36].

New academic fellowship opportunities or policy teams that are intentionally designed to bring multi-disciplinary researchers, and researchers from un(der)-represented regions together is crucial. A shift to any of these engagements will create a more inclusive approach for future policies, priorities, and governance methods.

Keeping technology companies and educational institutions accountable to reduce the gap in all its forms is important to improve workplace diversity. From an AI ethics perspective, this is also crucial to address the bias and discrimination concerns in future AI systems, research, policies, and standards. Policy interventions to challenge the toxic organizational cultures and to increase transparency in hiring, compensation, and performance are needed for course correction.

4 Call to action to move from principles to practice

AI ethics and implications of AI-powered products and services are coming under spotlight. 88% of the principles in Jobin et al. [10] meta-analysis was released in 2018 and 2019. AI is already ubiquitous in our lives. However, thanks to a number of AI ethics researchers, scholars [37–46] and investigative journalists who provided evidence of biased and/or unethical algorithms and their implications, we know now the importance of governance of AI and demand change and more transparency and accountability. Beginning of 2020, AI ethics field was challenged to "stop treating AI like magic, and take responsibility for creating, applying, and regulating it ethically". It was suggested that "for all the lip service paid to these issues, many organizations' AI ethics guidelines remain vague and hard to implement" [47]. Given all the high impact cases that surfaced in the news about facial recognition, AI in health decisions, disinformation on social media platforms or biased algorithms used in welfare eligibility or fraud detection to name a few, this call to move to practice is getting louder. Scholars, activists, ethicists, and journalists in the field agree that urgent action is necessary before more damage is done on an individual and societal level and before the damage erodes public trust of what is currently understood as AI or algorithms.

Along these lines, the future debate and the move to practices and governance should include a focus on the business and labor practices of big companies, the company culture and incentives, the social, racial, economic and environmental impacts of their technologies. The unique manner in which AI algorithms can quickly ingest, perpetuate, and legitimize forms of bias and harm represents a step change from previous technologies, warranting prompt reappraisal of these tools to ensure ethical and socially beneficial use [11]. When AI companies and institutions 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 [21]. Instead, we need genuine accountability mechanisms, external to companies and accessible to populations. Any AI system that is integrated into people's lives must be capable of contest, account, and redress to citizens and representatives of the public interest, asking which systems really deserve to be built? Which problems most need to be tackled? Who is best placed to build them? And who decides? [48, 49] Citizens and consumers all have a stake in the answers to these questions. Shanahan cautions for the "self-perpetuating tendency for power, wealth and resources to concentrate in the hands of few" [50]. Coeckelbergh advises that "if we endorse the ideal of democracy and if that concept includes inclusiveness and participation in decision-making about the future of our societies, then hearing the voice of stakeholders is not optional but ethically and politically required" [51].

Ethics principles and governance methods have a symbiotic relationship. The works on ethics rules for technology can be precursors of the law. They can give orientation on the possible content of legal rules but, they cannot replace the law [52]. Regulation itself, on the other hand, is not the only answer, and is only one of a range of tools that can be used to lend tangible shape to ethical principles. The lessons learned and synergistic use of various governance instruments at different levels (multi-level governance) are vital in view of the complexity and dynamism of data ecosystems [53].

5 Conclusion

Ethics and values are culture sensitive. Diversity of perspectives and experiences provide immense richness to any product and system. This new field of AI ethics still needs a lot of work to incorporate the wisdom and values of different cultures as it is going through its birthing pains and expand from its current heavily US-West-centric state. It needs to address the intersection of these principles with structures of power, with structural and historical inequities and ensure that the principles do not lead to the further power imbalances between private companies and states, or the disenfranchisement, oppression and exploitation of individuals and communities by either the private companies or states.

The development of future work needs to stay true to the principles' claims to be global and beneficial to all.

In reviewing the principles and frameworks, and in building a roadmap and choosing our next actions, we need to remember the questions Cath [54] raised, "who sets the agenda for AI governance? What cultural logic is represented by that agenda and who benefits from it?" It is crucial to remain critical of proposed generalizations and solutions when these eventually lead to development of norms, standards, practices, and regulation [36]. Private industry invests in the development of these products and benefits from them, so it is expected that there will be efforts to try to shape the conversation. Field of AI ethics needs to ensure that it is not blindly legitimizing these aims and efforts and hold these entities to higher standards and better practices. In short term, creating principles and business models that are responsible, ethical, and inclusive will cost money, time and resources. However, in the long term, it will help create better and sustainable products, services that respect society and work with the society's needs rather than governments' and private companies' interests.

Funding The author declares that no funding was received for this article.

Compliance with ethical standards

Conflict of interest The author declares that she has no conflict of interest.

References

1. Greene, D., Hoffmann, A., Stark, L.: Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In: HICSS (2019)
2. Benjamin, R.: Race After Technology: Abolitionist Tools for the New Jim Code. Polity, Cambridge (2019)
3. Mohamed, S., Png, M., Isaac, W.: Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. ArXiv, abs/2007.04068 (2020)
4. AlgorithmWatch. https://inventory.algorithmwatch.org/
5. AIethicist.org. https://www.aiethicist.org/ai-principles
6. AI Ethics Lab. https://aiethicslab.com/big-picture/
7. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
8. 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 No. 2020-1. https://doi.org/10.2139/ssrn.3518482 (2020)
9. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**, 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x
10. Jobin, A., Ienca, M., Vayena, E.: op. cit
11. Mohamed, Png & Isaac, op. cit
12. AI Ethics Lab: https://aiethicslab.com/big-picture/
13. Fjeld, Achten, Hilligoss, Nagy & Srikumar, op. cit
14. AlgoritmWatch. https://inventory.algorithmwatch.org/
15. Bietti, E.: From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. DRAFT—Final Paper Published in the Proceedings to ACM FAT* Conference (FAT* 2020). https://ssrn.com/abstract=3513182 (2019)
16. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J., Schwartz, O.: AI Now Report 2018, pp. 1–62 https://ainowinstitute.org/AI_Now_2018_Report.pdf (2018)
17. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
18. Vincent, J.: Tim Cook warns of 'data-industrial complex' in call for comprehensive US privacy laws https://www.theverge.com/2018/10/24/18017842/tim-cook-data-privacy-laws-us-speech-brussels (2018)
19. Greene, Hoffmann & Stark, op. cit
20. Whittaker, et al., op. cit
21. Hagendorff, op. cit
22. Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J.C., Lyons, T., Etchemendy, J., Grosz, B., Bauer, Z.: The AI Index 2018 Annual Report. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford (2018)
23. Simonite, T.: AI is the Future—But Where are the Women? WIRED. https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/ (2008)
24. World Economic Forum: The Global Gender Gap Index 2018. https://reports.weforum.org/global-gender-gap-report-2018/assessing-gender-gaps-in-artificial-intelligence/ (2019)
25. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. ArXiv, abs/1905.06876 (2019)
26. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., Cave, S.: Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. London: Nuffield Foundation. https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf (2019)
27. Jasanoff, S., Hurlbut, J.B.: A global observatory for gene editing. Nature **555**(7697), 435–437 (2018). https://doi.org/10.1038/d41586-018-03270-w
28. Costanza-Chock, S.: Design Justice: Community-Led Practices to Build the Worlds We Need. The MIT Press, Cambridge (2020)
29. Morley, Floridi, Kinsey & Elhalal, op. cit
30. Jasanoff, op. cit
31. Young, M., Magassa, L., Friedman, B.: Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. Ethics Inf. Technol. **21**, 89–103 (2019). https://doi.org/10.1007/s10676-019-09497-z
32. Kalluri, P.: Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature **583**(7815), 169 (2020). https://doi.org/10.1038/d41586-020-02003-2
33. Costanza-Chock, op. cit
34. Floridi, L., Cowls, J., Beltrametti, M., et al.: AI4People—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
35. UN Secretary-General's High-level Panel on Digital Cooperation. https://www.un.org/en/digital-cooperation-panel/ (2020)
36. Floridi, Cowls, Beltrametti, et al., op. cit
37. Pasquale, F.A.: The Black Box Society: The Secret Algorithms that Control Money and Information. Book Gallery (2015)

38. Zuboff, S.: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Public Affairs, New York (2019)
39. Barocas, S., Selbst, A.D.: Big Data's Disparate Impact. 104 California Law Review 671. https://doi.org/10.2139/ssrn.2477899 (2016)
40. Caliskan, A., Bryson, J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**, 183–186 (2017)
41. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. There software used across the country to predict future criminals. and its biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016)
42. Benjamin, op. cit
43. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91 (2018)
44. Eubanks, V.: Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press, New York (2018)
45. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5), 14–19 (2016)
46. Noble, S.U.: Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press, New York (2018)
47. Hao, K.: In 2020, let's stop AI ethics-washing and actually do something. MIT Technology Review. https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/ (2019)
48. Powles, J., Nissenbaum, H.: Not enough people are asking if artificial intelligence should be built in the first place. CNBC. https://www.cnbc.com/2018/12/14/the-seductive-diversion-of-solving-bias-in-artificial-intelligence.html (2018)
49. Zuboff, op. cit
50. Shanahan, M.: The Technological Singularity, p. 166. The MIT Press, Cambridge (2015)
51. Coeckelbergh, M.: AI Ethics, p. 170. The MIT Press, Cambridge (2020)
52. Nemitz, P.: Constitutional democracy and technology in the age of artificial intelligence. Philos. Trans. R. Soc. https://doi.org/10.1098/rsta.2018.0089 (2018)
53. Opinion of the Data Ethics Commission: Data Ethics Commission of the Federal Government. https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html (2019)
54. Cath, C.: Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos. Trans. R. Soc. https://doi.org/10.1098/rsta.2018.0080 (2018)