Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice

Sloane, Mona

Erstveröffentlichung / Primary Publication
Konferenzbeitrag / conference paper

Empfohlene Zitierung / Suggested Citation:
Sloane, M. (2019). Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice. In *Proceedings of the Weizenbaum Conference 2019 "Challenges of Digital Inequality - Digital Education, Digital Work, Digital Life"* (pp. 1-9). Berlin https://doi.org/10.34669/wi.cp/2.9

PROCEEDINGS OF THE WEIZENBAUM CONFERENCE 2019
CHALLENGES OF DIGITAL INEQUALITY
DIGITAL EDUCATION | DIGITAL WORK | DIGITAL LIFE

INEQUALITY IS THE NAME OF THE GAME. THOUGHTS ON THE EMERGING FIELD OF TECHNOLOGY, ETHICS AND SOCIAL JUSTICE.

Mona Sloane
Institute for Public Knowledge
New York University
New York City, USA
mona.sloane@nyu.edu

ABSTRACT

This paper argues that the hype around 'ethics' as panacea for remedying algorithmic discrimination is a smokescreen for carrying on with business as usual. First, it analyses how the current discourses around digital innovation and algorithmic technologies (including artificial intelligence or AI), newly emerging technology policy and governmental funding patterns as well as global industry developments are currently re-configured around 'ethical' considerations. Here, the paper shows how this phenomenon can be broken down into policy approaches and technological approaches. Second, it sets out to provide three pillars for a sociological framework that can help reconceptualize the algorithmic harm and discrimination as an issue of *social inequality*, rather than ethics. Here, it builds on works on data classification, human agency in design and intersectional inequality. To conclude, the paper suggests three pragmatic steps that should be taken in order to center social justice in technology policy and computer science education.

KEYWORDS

Algorithms; Ethics; Inequality; Sociology; Design

1 INTRODUCTION

This paper provocatively argues that the hype around 'ethics' as panacea for offsetting discrimination in and through algorithmic technologies[1] is a smokescreen for carrying on with business as usual. It suggests that 'ethics' is largely deployed to gain competitive advantage (between firms, industries, nations) rather than initiating a genuine push towards social justice. The paper builds this argument in three steps: First, it analyses *how* the current discourse around algorithm innovation is re-configured around 'ethical' considerations. As part of that, it delves into current computer science scholarship on 'moral machines' and puts this into context with the latest works on technology and discrimination. Second, it provides a sociological framework for conceptualizing the harm and discrimination that can be caused by digital technologies as an issue of *social inequality*, rather than ethics. Here, it builds on sociological approaches to notions of 'the social' in data classification, human agency in networks of design and intersectional inequality. Based on that it, third, suggests three points that must inform new technology policy and computer science pedagogy in order to center digital innovation on social justice.

2 ALGORITHMIC HARM: ETHICS TO THE RESCUE?

Over the past years, we have seen more and more evidence that algorithmic technologies can disproportionately disadvantage and/or harm social groups that are already negatively affected by social segregation and oppression (Bolukbasi et al 2016; Buolamwini and Gebru 2018; Eubanks 2018; Noble 2018; O'Neil 2016). In response, many technologists and policy makers set out to remedy this situation by way of 'ethics'.

The issues of algorithmic discrimination and harm are increasingly addressed through emphasizing the need for 'ethics' in algorithmic technologies. While there is a substantial body of work that has long argued that technical artifacts *do* have politics (see famously Winner 1980) and thus *do* contain, in one form or another, values (see for example Nissenbaum 2010), the idea of ethics in algorithmic technology has recently taken particular shapes: 'ethical AI' has not only been announced as a 'top technology trend for 2019' (Lomas 2018), but is also being positioned as a key element in the global race for technology leadership, informing heavily funded university initiatives (such as MIT's $1 billion investment into the Stephen A. Schwarzman College of Computing which will have an explicit focus on 'ethical considerations relevant to computing and AI' [MIT News 2018]) as well as government and industry investments (Dutton 2018, Sloane 2018).

In practice, the recent rise of 'ethics' in the context of algorithmic technologies has informed two types of (often overlapping) approaches to mitigating algorithmic harm: ethics as a *policy* approach[2] and ethics as a *technology* approach. The *policy* approach often materializes as a form of self-regulation, for example through ethics codes, frameworks and principles that set out to define sets of rules and values to help guide a 'responsible' development of AI technology[3].

[1] In this paper, the term 'algorithmic technologies' refers broadly to any *digital* technology that is put to work based on an algorithm (including machine learning technologies), whereby an algorithm is a 'computational procedure for deriving a result, much like a recipe is a procedure for making a particular dish' (Broussard 2018, p. 20). Furthermore, the term 'algorithmic technologies', here, includes automated decision-making systems, commonly referred to as 'artificial intelligence' or 'AI'.

[2] Due to the limited scope of this paper, the policy approach, here, does exclude legislative frameworks that regulate issues adjacent to algorithmic technology, such as data (e.g. the EU's General Data Protection Regulation [GDPR]).

[3] Prominent examples include Google's 'Objectives for AI Applications' (Pichai 2018), the newly updated 'Code of Ethics and Professional Conduct' by the Association of Computing Machinery (ACM 2018) or the Institute of

But it may also take the form of external 'ethics boards'[4], fairness and ethics trainings for computer science students and professionals (Fiesler 2018; Vallor 2018) or the suggestion of algorithm designers and engineers swearing a 'Hippocratic Oath' (Etzioni 2018).

While the policy approach targets the human lead within algorithm design, the *technology* approach sets out to create what we may call 'moral machines' (Wallach and Allen 2009). The notion of 'ethics' that informs these efforts tends to be grounded in the tradition of moral philosophy. Without wanting to crudely simplify the vast scholarly tradition of moral philosophy dating back to Kant's categorical imperative, we may describe moral philosophy as a theory that is fundamentally concerned with what counts as a good life as basis for making a decision (Vallor 2016). The overarching goal of creating 'moral machines' is to work ethics, morality and values into the machines themselves (Anderson and Leigh Anderson 2011; Yu et al 2018). This consideration has become more urgent in the context of the increased complexity and computational capability of algorithmic technologies that are deployed as autonomous agents (or as 'AI'). The common rationale is that these agents now require a 'capacity for moral decision making' (Moniz Pereira and Saptawijaya 2016) when working towards achieving goals[5]. Related considerations and new strategies are emerging in the context of 'fairness' enhancement and 'bias' mitigation in algorithmic technologies[6].

3 INEQUALITY IS THE NAME OF THE GAME

Unsurprisingly, the way in which 'ethics' is currently enacted and deployed is increasingly criticized. A key critique is the fact that neither the policy approach, nor the technological approach to 'ethics' is grounded in a legal framework – 'ethics' is simply not enforceable by law (Chadwick 2018) and ultimately remains a gesture of goodwill of those who create algorithmic technologies. The overwhelming – and voluntary – commitment to 'ethics' by companies selling algorithmic technologies can, therefore, be seen as a form of 'whitewashing' (Wagner 2018). Additionally, there is new evidence indicating that ethical frameworks simply do not affect the decision making of technologists (McNamara, Smith and Murphy-Hill 2018). And as Greene, Hoffman and Stark (2018) show, 'ethics' tend to position algorithmic harm as a *social* problem that requires a *technical* solution. That the social problem is deeply entangled with the existing fault lines of social stratification falls somewhat outside of the ontology of 'ethical algorithms'. Therefore, relying on 'ethics', whether through policy or technical approaches, implies that the mechanics retaining the status quo remain untouched. In other words, the notion of 'ethics' does *not* require us to examine the historic,

Electrical and Electronics Engineers' (IEEE) framework for 'Ethically Aligned Design' (IEEE 2018).

[4] See, for example, the Axon AI and Policing Technology Ethics Board (Axon 2019), or Google's newly appointed Advanced Technology External Advisory Council (Walker 2019).

[5] See especially Noothigattu et al (2018) for insight into how autonomous agents may balance 'moral values' and 'game rewards' in the context of value-alignment.

[6] A substantial part of the discourse around 'fairness' in machine learning (and especially prediction-based decision-making systems) focuses on the question whether and how inequality patterns that inevitably emerge through data sets can be mitigated algorithmically (see especially the proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency [ACM FAT*], as well as Mitchell, Portash and Barocas 2018). Here, however, many fairness-in-machine learning scholars are careful not to suggest that the data that is used to train algorithmic systems and that describes the social world can be somewhat independent of the (unequal) structures that make up that social world (see especially Barocas, Hardt and Narayanan 2018). Sebastian Benthall and Bruce D. Haynes (2018) go further to argue that the social categories that are routinely used to classify social data are complicit in anchoring mechanisms of inequality and oppression. They take the case of racial categories in machine learning to illustrate this argument. Here, they suggest using *unsupervised* machine learning to dynamically detect patterns of segregation *prior* to group fairness interventions with the goal of preventing the perpetuation of racial categories as status categories of disadvantage.

systematic and complex inequalities that *cause* algorithmic bias and violence – even when we return to the fundamentals of moral philosophy: asking whether something is 'done ethically' does not question *who* defines and enforces what a good life is, and for whom, and from what position of power, or not, that decision is being made. That is to say that 'ethics' does not prompt us to reflect on and intervene in the social organization of algorithm design at large and the cultures and power relations that underpin it.

To illustrate this point, let us put it into the context of the current tech landscape: I would argue that there is a link between the fact that Alphabet Inc.'s search algorithm ('Google') tends to show white, male individuals in searches for the term 'CEO' (Sottek 2015), that Alphabet Inc. paid out a $90 Million exit package to a senior executive who had a track record in sexually harassing coworkers and that very recently 20,000 Google employees walked out of their offices in protest of this incident, and a misogynist work culture at large (Wakabayashi and Benner 2018). The link between these events is social, historical and cultural and it points to the ways in which different kinds of inequality manifest across all domains of social life, including technology design and the technology industry at large. And there is a growing body of work that supports this claim: Meredith Broussard (2018) has recently argued that the sexism and 'bro-culture' that is rampant in the tech industry is deeply entangled with the history of computing and mathematics in general while Marie Hicks (2017) has demonstrated how gendered inequalities in computation are not accidental, but derive from a particular cultural landscape and a series of policy decisions. Safiya Umoja Noble (2018) seminal study of search algorithms has revealed how capital, gender and race are central to the technological formation of social oppression. What cuts across these studies is one message: inequality, as a complex, historical and emergent phenomenon, is 'the name of the game'. And a narrow focus on 'ethics' through policy and technological approaches prevent us from examining the rules of this game from a *critical* point of view.

4 RECONFIGURING THE CONVERSATION

Clearly then, reclaiming this critical point of view requires a framework for conceptualizing the harm and discrimination that can be caused through algorithmic technologies as an issue of *social inequality*, rather than 'ethics'. This framework must narrow in on what we mean by 'inequality' in the context of algorithmic technology. It must also enable a critical observation of the contingencies of social life and the historical and cultural make-up of the contexts in which algorithmic technologies emerge. Here, it is useful to turn to the social sciences who have long dealt with these kinds of issues. I therefore propose to build on social and cultural theory and consider the following three aspects (which are not separate, but overlap) as part of pursuing more socially just technology design:

'The social' in data – As a basis for understanding inequality in algorithmic technology design, we must ask broader questions around *how* different notions of 'the social' get classified (see also Bowker and Leigh Star 1999) and embedded into the datasets that form the basis for algorithmic technology. Data selection and data classification are a way of world-making, they are based on humans making judgements about other things, social environments, other humans and so on. This world-making through the collection of data is not neutral, but steeped in history, culture, personal experience, social position and so on. This is where algorithmic inequality materializes as a continuation of existing social stratification and oppression. As judgements turn into data labels and data sets, they are decontextualized, so the backstory to their emergence is not carried over into the system. The issue of abstraction (see also Selbst et al 2018), of course, is one of the eternal tensions between quantitative and qualitative traditions of

knowing and describing the world. But as algorithmic technology takes on a constitutive role in the mediation of social life, the implications of abstractions scale up significantly. Ethics frameworks and moral machines do not put a question mark behind the way in which data becomes a social object and enshrines the status quo.

Human agency in technology design – Even though the current rhetoric cultivates the notion that algorithmic systems (especially 'AI') are capable of developing their own agency, it is clear that we are far away from systems that resemble a general artificial intelligence (Knight and Hao 2019). But it is a reality that algorithms are increasingly entrusted with making decisions about humans (Whittaker et al 2018). This means that we need a conceptual handle for assessing this new area of tension. That non-human actors play a constitutive role in society has long been established by schools of thought such as Science and Technology Studies (STS) and Actor-Network Theory (ANT). They are therefore useful for getting our heads around the relationship that evolves between humans and (computational) machines. Put broadly, STS and ANT promote an ontology whereby agency emerges in a network *between* human and non-human actors (Latour 2005). But the rise of algorithmic technology, paradoxically, makes a good case for *human* agency taking the lead in the formation of this assemblage: humans determine *who* becomes an algorithm designer, *how* the system is designed, how the *data* is selected and optimization targets (implicitly or explicitly) are set, and so on. And yet, the materiality of computational systems (from the increasingly powerful hardware to the 'neural network' structures enabling 'deep learning') plays a central role in the rise of algorithmic technologies. This means that in order to better understand the unfolding of agency (and politics) in algorithmic technology design, deployment and integration, we need a productive critique of STS and ANT. While 'ethics' are not a good vehicle for that, newer debates emerging adjacent to STS/ANT become central, particularly Antoine Hennion's (2016)[7] notion of pragmatism and Laura Forlano's (2017)[8] work on design in the context of nonhuman, the posthuman and the more than human.

Intersectional inequality – The use of 'ethics' in much of the current landscape of algorithmic technology does not only circumnavigate the concept of social inequality at large, but *intersectional* inequality specifically. Kimberlé Crenshaw's (1991) original notion of 'intersectionality' shows how categories of inequality, such as race, class and gender, intersect and are experienced. Crenshaw's study outlined how women of color were disproportionately affected by hiring discrimination and how neither the category of race, nor the category of gender fully captured their experience and could be leveraged in an anti-discrimination suit in court. Patricia Hill Collins (2000) developed the notion of intersectionality outside of the legal domain and proposed it as a general form of analysis 'claiming that systems of race, social class, gender, sexuality, ethnicity, nation, and age form mutually constructing features of social organization' (Collins 2000, p. 299). As a framework, intersectionality 'provides a complex understanding of inequality that takes multiple sources of disadvantage as the source and solution for inequality' (Hurtado 2018). Taking intersectional inequality seriously in the context of algorithmic technology means putting the lived

[7] Hennion (2016) positions pragmatism as a critique of ANT, starting from the issue that ANT's focus on object-people relation comes at the cost of diluting agency in a network between human and non-human actors. For him, pragmatism means "'socializing' objects, but not by emptying out their content" (Hennion, 2016, p. 299)

[8] Forlano (2017) critically analyses emergent design practices and perspectives against the backdrop of key works on the nonhuman, the posthuman and the more than human to suggest that it is important to acknowledge that posthumanism may not serve those communities who have traditionally been excluded from humanism in the first place, such as women, people of color, the LGTBQ community, and others.

experience of those affected by (algorithmic) discrimination front and centre in discussions around technology and social justice.

The notion of intersectional inequality is already informing new and important research in the context of algorithmic technology and design. Most notably, Sasha Costanza-Chock (2018) has built on the notion of intersectionality to show how *design* – as a socio-technical system at large – reproduces and is reproduced by a 'matrix of domination' in which gender, class and race serve as interlocking systems of oppression to formulate 'design justice principles' that can help break design's complicity with oppression. Relatedly, Joy Buolamwini and Timnit Gebru (2018) have taken intersectionality as a cue to use the Fitzpatrick skin type scale as a basis for a phenotypic evaluation of face-based gender classification accuracy in automated facial analysis. Schlesinger, Edwards and Grinter (2017) have built on the intersectionality lens in order to show how human-computer interaction (HCI) research can be comprised of clearer reporting of context to foster a deeper engagement with identity complexities.

These are all important advancements. But to help address the social problem of inequality *at large*, beyond the technological and ethical realm and as a broad research and policy goal, they need to by synthesized into a holistic framework that can help examine data categorization, materiality and agency, as suggested above.

5 MOVING FORWARD

On a pragmatic level, we must then take the following steps to foster a more focused consideration of inequality in technology practice and policy as well as computer science pedagogy:

(1) We must bring questions of data epistemology onto the top of the agenda, because knowing how data comes to describe and organize the social is key for understanding and mitigating algorithmic harm[9] and social inequality more broadly. Here, we may have to flip the script and focus on data classification as emerging from the lived experience of social actors, rather than as based on external evaluation and categorization. This acknowledges that data describing the social world can never be independent from the categories and hierarchies that organize that world. I can also put intersectional inequality at the heart of efforts to make algorithmic technologies socially just and prompt new political discussions about inequality *beyond* the technical realm.

(2) We need better collaborations between quantitative and qualitative scholarship, especially in the context of computer science pedagogy. Computer science students must be equipped with the conceptual tools they need to reflexively locate themselves, and their practice, in the social world. By the same token, we need social science and humanities scholars who are able to actively engage in data and computer science practice.

(3) We need a clearer picture of the terms that are at stake and currently do important political work, because the unclarity about key terms (such as 'algorithm', 'digitization', 'machine learning' and so on, but also 'fairness', 'bias', 'standardization', 'accountability') impacts our ability to have more productive conversations about the abilities and limits of new technologies, and explore regulatory possibilities.

These considerations, together with a framework that allows us to explore questions in the context of data classification, human agency and intersectional inequality in algorithm design, will allow us to reclaim digitization as a positive, rather than threatening, new way of knowing social life (see also Marres 2017). This will open up new possibilities for addressing the issue of social inequality at large, beyond the digital space.

[9] This is particularly salient in the context of the co-called 'black box problem', whereby it is unclear to the human actor how the algorithm reached its conclusion/prediction.

6 CONCLUSION

This paper has argued that the current focus on and enactment of 'ethics' will *not* facilitate social justice in algorithmic technology. To do so, it has mapped out how ethics – as policy approach and ethics as technological approach – fails to solve the root problem of algorithmic discrimination. To illustrate this point, the paper has built on recent developments in the tech industry and argued that the historic continuation of certain cultures, power structures and ways of socially organizing algorithm design require a conceptual handle that reconfigures algorithmic discrimination as an issue of *social inequality*, rather than ethics. Here, it has suggested to combine sociological approaches to notions of 'the social' in data classification, human agency in networks of design and intersectional inequality. To conclude, the paper has taken this framework as a cue to suggest three pragmatic steps that must be taken in order to move forward in technology policy and computer science education: (1) focusing on data epistemology as emergent from lived experience, (2) better dialogue between quantitative and qualitative scholarships, (especially in the context of computer science pedagogy), and (3) more clarity about key terms that are currently at stake in the discourse around digitization, algorithmic technology and inequality.

7 ACKNOWLEDGMENTS

I would like to thank the Institute of Public Knowledge at NYU, and particularly Prof Eric Klinenberg and Jessica Coffey, for their generous support of my work and for providing room to grow.

8 REFERENCES

1. Association for Computing Machinery (ACM) (2018). ACM Code of Ethics and Professional Conduct. https://www.acm.org/code-of-ethics.
2. Anderson, M., Leigh Anderson, S. eds. (2011). Machine Ethics. Cambridge, UK: Cambridge University Press.
3. Axon (2019). Axon AI and Policing Technology Ethics Board. https://www.axon.com/info/ai-ethics.
4. Barocas, S., Hardt, M. & Narayanan, A. (2018). Fairness and Machine Learning. fairmlbook. org, 2018 URL: http://www. fairmlbook.org.
5. Benthall, S., Haynes, B.D. (2019). Racial categories in machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency 2019. https://dl.acm.org/citation.cfm?id=3287575
6. Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems, 4349–4357, arXiv:1607.06520v1.
7. Bowker, G. C., Leigh Star, S. (1999). Sorting Things Out. Classification and Its Consequences. Cambrdige, MA: MIT Press.
8. Buolamwini, J., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research, 81:1–15, Conference on Fairness, Accountability, and Transparency 2018, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.
9. Broussard, M. (2018). Artificial Unintelligence: how computers misunderstand the world. Cambridge, MA: MIT Press.
10. Chadwick, P. (2018). To regulate AI we need new laws, not just a code of ethics. The Guardian, October 28, 2019, https://www.theguardian.com/commentisfree/2018/oct/28/regulate-ai-new-laws-code-of-ethics-technology-power.
11. Costanza-Chock, S. (2018). Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice. Proceedings of the Design Research Society 2018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3189696
12. Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6), 1241-1299, DOI: 10.2307/1229039.
13. Dutton, T. (2018). An Overview of National AI Strategies. https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.
14. Etzioni, O. (2018). A Hippocratic Oath for artificial intelligence practitioners. TechCrunch, https://techcrunch.com/2018/03/14/a-hippocratic-oath-for-artificial-intelligence-practitioners/.
15. Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St Martin's Press.

16. Fiesler, C. (2018). Tech Ethics Curricula: A Collection of Syllabi. Medium https://medium.com/@cfiesler/tech-ethics-curricula-a-collection-of-syllabi-3eedfb76be18.
17. Forlano, L. (2017). Posthumanism and Design. She Ji: The Journal of Design, Economics, and Innovation, 3(1), 16-29.
18. Greene, D., Hoffmann, A.L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. Hawaii International Conference on System Sciences (HICSS), Maui, HI.
19. Hennion, A. (2016): From ANT to Pragmatism: A Journey with Bruno Latour at the CSI. In New Literary History, 47(2&3), 289-308.
20. Hicks, M. (2017). Programmed Inequality. How Britain Discarded Women Technologists and Lost Its Edge in Computing. Cambridge, MA: MIT Press.
21. Hill Collins, P. (2000). Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment,1st edition. New York: Routledge.
22. Hurtado, A. (2018). Intersectional Understandings of Inequality. In Phillip L. Hammack Jr., (Ed.) The Oxford Handbook of Social Psychology and Social Justice (Oxford: Oxford University Press, 2018).
23. Institute of Electrical and Electronics Engineers' (IEEE) (2018). Ethically Aligned Design. https://ethicsinaction.ieee.org.
24. Knight W., Hao, K. (2019). Never mind killer robots—here are six real AI dangers to watch out for in 2019, MIT Technology Review, https://www.technologyreview.com/s/612689/never-mind-killer-robotshere-are-six-real-ai-dangers-to-watch-out-for-in-2019/.
25. Latour, B. (2005). Reassembling the Social. An Introduction to Actor-Network-Theory. Oxford: Oxford University Press.
26. Lomas, N. (2018). Gartner picks digital ethics and privacy as a strategic trend for 2019. In *TechCrunch,* https://techcrunch.com/2018/10/16/gartner-picks-digital-ethics-and-privacy-as-a-strategic-trend-for-2019/.
27. Marres, N. (2017). Digital Sociology: The Reinvention of Social Research. Malden, MA: Polity Press.
28. McNamara, A., Smith, J., Murphy-Hill, E. (2018). Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?. In: Proceedings of the 26th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '18), November 4–9, 2018, https://doi.org/10.1145/3236024.3264833.
29. Mitchell, S., Portash, E., & Barocas, S. (2018). Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. arXiv pre-print, arXiv: arXiv:1811.07867v1.
30. Moniz Pereira, L., Saptawijaya, A. (2016). Programming Machine Ethics. Springer International Publishing.
31. Nissenbaum, H. (2010). Privacy in Context. Technology, Policy, and the Integrity of Social Life. Stanford, CA: Stanford University Press.
32. Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press.
33. Noothigattu, R., Bouneffouf, D. Mattei, N., Chandra, R., Madan, P., Varshney, K., Campbell, M., Singh, M., Rossi, F. (2018). Interpretable Multi-Objective Reinforcement Learning through Policy Orchestration. arXiv pre-print, arXiv:1809.08343v1.
34. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. London: Penguin Books.
35. Pichai, S. (2018). AI at Google: our principles. https://www.blog.google/technology/ai/ai-principles/.
36. Schlesinger, A., Edwards W.K, & Grinter, R.E. (2017). Intersectional HCI: Engaging Identity through Gender, Race, and Class. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 5412-5427.
37. Selbst, A. D., boyd, d., Friedler, S., Venkatasubramanian S., Vertesi, J. (2018). Fairness and Abstraction in Sociotechnical Systems. In Proceedings ACM Conference on Fairness, Accountability, and Transparency (FAT*) 2019, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265913.
38. Sloane, M. (2018). Making artificial intelligence socially just: why the current focus on ethics is not enough. In LSE British Politics and Policy Blog, http://blogs.lse.ac.uk/politicsandpolicy/artificial-intelligence-and-society-ethics/.
39. Sottek, T. C. (2015). Google Search thinks the most important female CEO is Barbie. The Verge, April 9, 2015, https://www.theverge.com/tldr/2015/4/9/8378745/i-see-white-people.
40. Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford: Oxford University Press.

41. Vallor, S. (2018). An Introduction to Data Ethics. Markkula Center for Applied Ethics. https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/an-introduction-to-data-ethics/.
42. Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In M. Hildebrandt (Ed.), Being Profiling. Cogitas ergo sum. Amsterdam University Press.
43. Walker, K. (2019). An external advisory council to help advance the responsible development of AI. https://www.blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/.
44. Wakabayashi, D., Benner, K. (2018). How Google Protected Andy Rubin, the 'Father of Android'. The New York Times, Oct 25, 2018. https://www.nytimes.com/2018/10/25/technology/google-sexual-harassment-andy-rubin.html?partner=IFTTT
45. Wallach, W., Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. Oxford: Oxford University Press.
46. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J., Schwartz, O. (2018). AI Now Report 2018. AI Now Institute, New York University. https://ainowinstitute.org/AI_Now_2018_Report.pdf.
47. Winner, L. (1980). Do Artifacts Have Politics? In Daedalus, 109(1), 121-136.
48. Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., Yang, Q. (2018). Building Ethics into Artificial Intelligence. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 5527-5533.