Bianca PRIETL
DOI: 10.3217/978-3-85125-855-4-18

Why Ethics Norms are Not Enough, or: How Current Critique of Digital Data Technologies Preserves Power

B. Prietl

Institute of Sociology, Johannes Kepler University (JKU) Linz, Austria

Abstract. This paper contributes to the long-standing tradition within STS to analyze the intertwining of technoscientific discourses, practices and artefacts with power asymmetries and social inequalities. It does so by scrutinizing AI ethics guidelines and standards as the currently dominant form in which society articulates critique of digital data technologies and seeks to cope with this critique. Based on a discourse-analytical reflection of selected AI ethics norms, it is argued that the key premises underpinning this form of dealing with the challenges and risks accompanying the digital transformations of society come with some severe limitations, especially when it comes to understanding and questioning social relations of power and the role that digital data technologies play in upholding and/or producing them. Against this backdrop, the turn to ethical responses as a 'panacea' is described as conserving rather than reducing existing power relations. Therefore, the article pleads for strengthening critical/feminist STS-perspectives within the ongoing negotiations of how to understand and handle the risks and challenges that accompany the digital transformations of society.

1 *AI ethics* and the Manifold Contradictions of Digitalization, Datafication and AI

In recent years, a rising number of research has problematized the social, political, and economic contradictions of digitalization, datafication and AI, thus, challenging the promises made in the name of digital data technologies [1], namely: to foster emancipation, decentralization, democratization and objectivity. Racist risk assessment tools employed in the US criminal justice system, sexist recruiting tools deployed by private companies or highly classist algorithmic decision-making systems used in credit granting procedures – to name just a few – testify to the fact that digital data technologies are neither neutral nor objective, but prone to bias and discriminatory results (O'Neil 2016; Noble 2018; Prietl 2019a; Gebru 2020). Public uproar due to privacy breaches, filter bubbles and social bots influencing democratic processes (Pariser 2011; Wooley 2016; Pörsksen 2018; Dutton et al. 2019) have further fueled debates on how to regulate private as well as state organizations involved in the development and use of digital data technologies, especially as monopolization tendencies increase (Lyon 2004; Leighton et al. 2017; Srnicek 2018; Véliz 2021).

Whereas computer science researchers pursue technology-focused ways of addressing the aforementioned problems such as discrimination aware data mining (DADM) or fairness, accountability and transparency in machine learning (FAccT), the prime (public) reaction to be observed is a ‘call for ethics’ resounding within politics, academia as well as industry. Around the globe, responses to the challenges and risks posed by digital data technologies either take the form of ethics boards, audits, frameworks and guidelines, or of efforts to educate AI developers and data scientists in ethical awareness (McKay/Yallaly 2017): In 2016 the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems was launched as an industry connection program of the IEEE Standards Association, aspiring for “ethics to become the new green.” In 2019 the US National Science Foundation (NSF) has partnered with Amazon to jointly support computational research focused on fairness in AI, while the European Commission presented its *Ethics Guidelines for Trustworthy Artificial Intelligence.* In academia, research centers and chairs for AI ethics are being institutionalized such as the heavily disputed *Institute for Ethics in Artificial Intelligence* at Technical University of Munich that was founded with the financial support of Facebook. Additionally, the big global tech companies, such as Google or Microsoft, have released ethical principles as a sign of self-regulation.

As these examples show, industry is heavily involved in discussing and shaping the science, morality and laws of AI and digital data technologies more in general. Critics, thus, called the hype around AI ethics a smokescreen for carrying on with business as usual. Instead of initiating a genuine push towards social justice and equality, ethics were largely employed to convince politicians as well as the general public that ethics guidelines were enough or even better than regulations when it comes to handling the manifold contradictions accompanying the digital transformations of society. Considering furthermore that most of these self-imposed standards are hardly binding, even less enforceable by law, they might remain a mere gesture of goodwill (Sloane 2019; Benkler 2019; Daly et al. 2019; Nosthoff/Maschewski 2019). [2]

While I am sharing the skepticism vis à vis a so-called ‘ethics washing’, this paper wants to take the critique of an ‘ethical’ approach towards solving the challenges related to digital data technologies a step further: By taking a closer look at the key premises of what becomes currently implemented as AI ethics, it sets out to scrutinize this dominant form of criticizing digital data technologies and coping with this critique. Though critical towards the all-encompassing recourse to ethics, it also distances itself from what has been called ‘ethics bashing’, namely the over-simplified consideration of ethics as shallow communication strategy and instrumentalized cover-up, with (moral) philosophy being reduced to a mere academic endeavor that stands in opposition to political discussion and social organizing (Bietti 2020). However, this paper contents that the sole focus on ethical norms has its limitations when it comes to understanding, reflecting and challenging power asymmetries and social inequalities and the role that digital data technologies play in upholding them.

2 Approaching Ethics Norms as an Object of Study

The argument presented in this paper is informed by critical and feminist perspectives in STS. Digital data technologies are understood as sociotechnical phenomena that are the result of an inextricably interweaving of technology and society, of the material and the semiotic (e.g., Haraway 1985). As such they are considered to have politics (Winner 1980), materializing and at the same time (re)producing social structures (of power and inequality) as well as cultural orders (and symbolic hierarchies). Therefore, digital data technologies are neither neutral artefacts, nor are they isolated from society, its structures, institutions and norms. From this point of view, ethics guidelines and standards are an interesting object of study as they make the values and norms materialized in digital data technologies explicit—at least partly [3].

Although the question of how effective ethical guidelines are in changing the thinking and practices of technology designers is heavily disputed (Hagendorff 2020), ethics statements can be considered "powerful instruments for constructing and imposing a shared ethical frame on a contentious conversation" (Greene/Hoffmann/Stark 2019: 2129), thus, setting the table for further discussions on 'good' digital data technologies. They are constituted by and constituting the historical, local, political, economic, and cultural conditions of the society they shape and are shaped by. Conceptualized as *discursive elements*, I understand ethics guidelines and standards as taking part in (pre)structuring our perceptions of digital data technologies, the ways we think about their design, implementation and use. As such they are productive in the sense that they take part in enabling specific sociotechnical paths of developing and using digital data technologies, while impeding others. Put differently, ethical statements are powerful as well as political, being a means of power as well as instrumental to power relations (Paulitz 2005; Prietl 2019b).

Following this line of thinking, the core argument of this paper is that addressing the challenges and risks of digitalization, datafication and AI solely in terms of ethics is not well suited for challenging the existing power asymmetries and social structures of inequality in our society. Drawing on a discourse-analytical reflection of 16 ethics guidelines and standards that have been theoretically sampled using the *AI Ethics Guidelines Global Inventory* provided by the German watchdog-organization Algorithm Watch [4] as well as literature on AI ethics, I will sketch three limitations of this currently dominant form of criticizing digital data technologies and coping with this critique: (1)

an a-social understanding of action, (2) an individualistic conceptualization of problems as 'errors to be fixed' and (3) a focus on fairness of distribution as proposed solution.

3 State of the Debate on AI ethics Guidelines

The last few years have seen an explosion of literature on AI ethics, sometimes also called digital ethics, computer ethics, or information ethics. Following Dignum's (2018: 2) differentiation of three levels on which AI and ethics can be related, this vast and heterogenous field of research can be divided into three strands: first, *ethics by design* that is concerned with how to include non-human actors in moral thinking (e.g. Adam 2008) and how to program ethical reasoning capabilities within artificially intelligent artefacts allowing them to act 'as if' they were moral agents (Allen et al. 2006; Etzioni/Etzioni 2017; Cervantes et al. 2019); second, *ethics for design* that is concerned with developing specific rules and criteria for how to design 'good' AI, e.g. guidelines, frameworks and standards, including codes of conduct for designer and/or users of AI (Bostrom/Yudkowsky 2014; Filipovic et al. 2018); and third*, ethics in design* that is concerned with the regulatory and engineering methods that allow for analyzing the ethical implications of AI as they become implemented within society (e.g. Cath et al. 2018; Floridi et al. 2018). Additionally, there is a growing interest in how AI ethics are being developed and implemented across the globe and by different organizations (see below). The paper in hand contributes to the latter strand of research, thereby taking AI ethics as an object of STS-analysis.

Several comparative analyses of AI ethics guidelines and frameworks help to map the debate on how to design and implement 'good' AI, which to date centers around almost two hundred documents, of which the huge majority has been issued since 2018. While acknowledging that the design of AI is a legitimate site for ethical debate rather than a neutral domain, there are huge differences in relation to how the proposed principles are interpreted, how they are legitimized, whom they address and how they should be implemented and enforced (Jobin et al. 2019; Daly et al. 2019; Greene/Hoffmann/Stark 2019; Hagendorff 2020).

Looking at who issued these documents, it can be noted that private companies, amongst which are some of the global tech giants that lead research and development in AI, and governmental agencies play a major role in developing AI ethics, with actors from the so-called Global North being overrepresented (Jobin et al. 2019: 3-5; Schiff et al. 2000: 154). Whereas the UK and the USA together account for more than a third of all ethical AI-documents, followed by Japan, Germany, France, and Finland, voices from Africa, South and Central America, and Central Asia are underrepresented. The addressees of these guidelines and frameworks are for the most part either multiple stakeholder groups and/or AI practitioners, with a significant portion being self-directed (Jobin et al. 2019: 6). Albeit there existing no agreed upon set of ethical standards to govern the design, development and deployment of 'good' AI (Yeung/Howes/Pogrebna 2020), the documents seem to converge around a handful of ethical principles, namely: transparency and accountability, justice and fairness, non-maleficence, responsibility, and privacy.

Hagendorff (2020, 103) argues that all of the recurring elements of 'good' AI are requirements which are rather easily operationalizable mathematically and for which technical solutions are being developed, with some companies already offering specific technological fixes such as tools for bias mitigation or fairness in machine learning. On the other hand, there are a number of issues that are only rarely mentioned within the majority of these guidelines and frameworks. These blind spots contain questions of (ecological) sustainability and hidden costs of AI development, malevolent (mis)use of AI, democratic control and governance of AI, questions of human dignity as well as solidarity and social responsibility (Jobin et al. 2019, 7; Hagendorff 2020, 104-105). When it comes to how AI are understood in these documents, it can be noted that a rather deterministic vision of AI is dominant with AI artefacts being mostly understood as isolated entities that can be optimized by experts so as to find technical solutions for what are perceived of as social problems. Consequently, there seems to be little to no discussion on how AI could be constrained or limited. Instead, 'better building' is presented as the only ethical path forward (Greene/Hoffmann/Stark 2019, 2122-2128). As Greene, Hoffmann and Stark (2019) point out, ethical design is considered to be a project of expert oversight, whereas the experts in question are supposed to be primarily technical, and secondarily legal experts (2126). What is however lacking is a consideration of the wider social contexts and relationships within which AI is embedded (Hagendorf 2020, 104).

4 Limitations of the Currently Dominant Form of Criticizing Digital Data Technologies and Coping with this Critique

Taking the above literature review as a starting point and drawing on my own analysis of selected AI ethics guidelines and statements, I will now elaborate on how the key premises underpinning these discursive elements are related to questions of power and social inequality.

4.1 Understanding Action

In accordance with Western-Eurocentric philosophy, AI ethics guidelines and standards largely—and: unquestioningly—assume the existence of a rational and autonomous (human) being as the subject of any—be it ethical or unethical—action. Consequently, the diverse ethics norms continuously—albeit often implicitly—address the figure of the autonomous subject of action—be it designers, practitioners or users of AI and other digital data technologies—appealing to their understanding to modify their behavior and actions. At the same time, discussions about the sociocultural, political, economic or organizational context within which these actions take place are largely missing. Thus, the social embeddedness of all action that not only explains certain actions, but also pre-structures them and limits any 'simple' and willful alteration of so-far established modes of acting and behaving in certain circumstances, are neglected within the majority of documents analyzed. Neither is there a systematic mentioning, let alone discussion, of the fact the digital data technologies are developed by private companies whose main objective is economic profit, and not the dismantling of social inequalities; nor are software developers or AI practitioners addressed as employees, which they mostly are, and, thus, first and foremost obliged to comply with their employers' demands.

As a consequence, the social structures and symbolic hierarchies that are influential for how people—and machines—can act in certain situations and that—even more importantly—are out of their immediate reach, are hardly taken into account. Thus, one important factor for not only understanding, but also for changing existing sociotechnical relations and their consequences is consistently ignored in this currently dominant form of criticizing digital data technologies and their social effects.

What is needed instead, is a decidedly *social* understanding of action that allows for challenging the power asymmetries and social structures of inequality within which digital data technologies and the people developing and using them operate (see e.g.: Weber 1920; Bourdieu 1987; Emirbayer/Mische 1998). Acknowledging the social embeddedness of all action, namely: of embeddedness within hierarchical structures and symbolic orders, makes visible the limits of individual willful actions and the capability of changing them. It also draws our attention to these historically established, social phenomena that are in need of change, if existing power asymmetries shall not be reproduced.

Apart from the economic and organizational structuring of digital data technologies mentioned above, there are also symbolic asymmetries to be considered that manifest themselves in available data (sets). Take for example the case of automatic skin cancer identification (Zou/Schiebinger 2018). The fact that such technologies are much less effective in identifying skin cancer in people of color than in 'white' or lighter pigmented people is not so much the result of 'bad' actors—be it the developers of the respective technology or the medical applicants; instead, it is the historically established asymmetries of technoscientific in/visibilities that build the foundation for the overrepresentation of 'white people' within machine learning-training data sets that then results in the tool's increased ability to identify deviations in their skins. Only, if we keep these structural components of sociotechnical action in mind, can we effectively address the problems laying at the heart of the manifold contradictions accompanying the spread of digital data technologies.

4.2 Conceptualizing the Problem

Related to the aforementioned asocial understanding of action is a rather narrow causal thinking that focuses on—human or technical—'errors to be fixed' when conceptualizing the problems of AI and other digital data technologies. Consequently, the diverse ethics norms depict the risks and challenges related to digital data technologies largely as *individualistic* problems or errors that call for a singular solution. Initiatives such as discrimination aware data mining (DADM) or fairness, accountability and transparency in machine learning (FAccT) strive to rectifying bias in AI by developing better algorithms. Ethics by design initiatives on the other hand focus on human actors, such as AI researchers and developers, and propose codes of conduct that shall guarantee the development of 'good' digital data technologies.

Thus, it is a technosolutionist stance that underpins the majority of AI ethics guidelines and standards, according to which all problems—also genuinely social ones—can be solved by technical means. Again, neglecting the social structuring of technology, 'better building' is presented as the only way forward (Greene et al. 2019, 2122-2128). Alternative paths such as a fundamental socio-political debate about which technologies would be desirable for which situations, however, do not even seem to be an option in the respective documents.

Critical and feminist STS-perspectives instead remind us that "artefacts have politics" (Winner 1980). Acknowledging the political dimensions of AI and other digital data technologies raises questions about who is (not) involved in creating these technologies, whose wishes and needs are (not) accounted for and who profits from their implementation and use—and who does not (Weber/Prietl 2021). Additionally, such a sensitivity towards questions of power also draws attention to the inextricable interweaving of technology and society, the material and the symbolic, thus, drawing our attention to how these diverse human and non-human actants evolve together and produce certain effects in their intra-acting (Barad 2003).

Take for instance the AMAS-algorithm of the Austrian Public Employment Service, AMS, as an example. The tool has become heavily criticized for assessing the chances of women, migrant and elderly employment seekers for finding new employment systematically lower than those of younger, Austrian-born men, which has consequences for the job seekers' entitlement to receive specific benefits and services by the AMS. In order to understand and, thus, be in the position to criticize the introduction of this digital data technology it is not enough to point out these biases or—even worse—single out the *one* cause of the problem to be corrected. But, we need to account for the technology's connectedness to certain political aims (optimizing resource allocation within neoliberal welfare state reforms), a specific statistical model (having been set in a way to optimize its overall accuracy, i.e. its 'correct' prediction of future employment chances derived from patterns found in past employment chances), a labor market that is highly discriminatory (especially against women with children and people with a migration background), and a strong belief in technical efficiency and objectivity (Lopez 2019, Allhutter et al. 2020). Keeping all of these intra-acting elements in mind, of course, does not make it any easier to describe, understand and eventually solve the problem at hand, but it increases the chances of not remaining a mere gesture of goodwill.

4.3 Proposing a Solution

In the light of biases against minorities and vulnerable groups of people, it is of little surprise that fairness features prominently in many AI ethics guidelines and standards. However, in the documents analyzed fairness is—if at all—mostly defined as equal-treatment, with questions of power and social structures of inequality again being left out of the picture. Instead, transparency, accountability and trustworthiness are promoted as pathing the way to fair digital data technologies (see also: Daly et al. 2019; Greene et al. 2019; Hagendorff 2020). With equal treatment as core solution and normative goal, non-discrimination is declared to guarantee justice.

As Hoffmann (2019, 905ff.) has problematized with regards to US anti-discrimination politics, such a focus on equal treatment is not well suited to address the intersecting effects of different categories of discrimination and inequalities, such as face recognition software being least efficient in recognizing the faces of black women [5]. What is even more worrisome, is that equating fairness with equal treatment and the latter with non-discrimination does not account for the fact that people are positioned in a highly unequal and hierarchical way in our society, thus, essentially ignoring that there is no level playing field.

Consider for example the infamous COMPAS-algorithm employed in the US criminal justice system to assess the recidivism rates of defendants. Although the tool does not explicitly take into account the 'race' of the accused, African Americans are much more likely to be assigned a higher risk score than 'white' Americans (Angwin et al. 2016). Some computer scientist work on improving the algorithm's accuracy, hoping that the same accuracy for 'white' as well as African Americans will solve the problem of unequal treatment (Corbett-Davies/Goel 2018). What such an approach, however, does not account for is that people of color are much more likely to be targeted as high risk because of structural racism that makes it much more likely for them to have no higher education, to be unemployed or to be related to someone who has been charged with a criminal offense – all of which are factors accounted for in the risk model implemented in the COMPAS-algorithm. Ignoring these fundamentally unequal preconditions not only does not solve the problem at hand but threatens to mask the problem in a veil of equal treatment. Considering EU anti-discrimination law that aims for substantial, not only formal, equality, Wachter et al. (2021) therefore argue for fair machine learning techniques that 'transform bias'. Instead of 'preserving bias', such techniques explicitly account for historically established social inequalities and try to actively counteract them.

5 It's About Power, Stupid!

Based on a discourse-analytical reflection of AI ethics guidelines and standards, this article has outlined how this currently dominant form of criticizing digital data technologies and coping with this critique is strongly influenced by traditional Western-Eurocentric moral philosophy that is highly individualistic in its approach (also: Jaume-Palasi 2019: 483). Such an epistem-ontological underpinning entails considerable limitations when it comes to addressing the role that digital data technologies play for maintaining power asymmetries and social inequalities. Whereas the very social structures and symbolic orders within which digital data technologies are developed, produced and used are largely out of sight, attention is directed first and foremost to singular 'black sheep'—be they human or technical artefacts. Framing these 'bad' actors as responsible for the problems and challenges of digital data technologies, it is in their 'correction' that a 'solution' is sought for.

Following this line of reasoning, the recent turn to ethics norms seems to be neither a 'panacea' against the manifold contradictions accompanying digitalization processes, nor a neutral undertaking. On the contrary, AI ethics guidelines and standards can be described as preserving existing power relations and social inequalities as they leave the social relations within which digital data technologies and their developers, designers and users operate largely untouched. The monopoly-like structure established by a few, primarily private-sector but also state organizations, due to the extreme resource intensity and high economies of scale of data-based AI (Srnicek 2018), is not systematically addressed in these documents, let alone problematized. Thus, the few dominant actors can continue to develop and deploy digital data technologies primarily to pursue their own interests, specifically: "profit (for a few), surveillance (of the minoritized), and efficiency (amidst scarcity)" (D'Iganzio/Klein

2020, 41). Nor are the ‘conservative’ logics of algorithmic knowledge-production and decision-making considered as crucial to understanding—and, consequently, addressing—the challenges of digitalization processes, most importantly the idea that patterns found in data of the past, allow for predicting the future (Lopez 2019; Prietl 2019a, b).

Therefore, a strengthening of critical/feminist STS-perspectives is needed within the social negotiations of digitalization and AI. These perspectives would—amongst others—direct our attention to the following questions, and call for their public debate: Who is in charge of developing digital data technologies? Who benefits from their use? Which purposes are served by digital data technologies? Who and/or what aspects of (social) life are considered, in- and excluded? For which purposes do we want to use digital data technologies? What technologies do we want as a society? Starting from a debate on these and many more questions, a critical/feminist approach to challenging digital data technologies would also entail to give up on hopes of neutrality and objectivity but strive for digital data technologies that—albeit no longer being able to claim neutrality—are explicitly dedicated to reduce power asymmetries and social inequalities instead of upholding the status quo.

Endnotes

[1] Digital data technologies designate technical artifacts that operate with digital data, e.g. AI-technologies, mail programs or tracking-apps. It is digital data technologies that are at the heart of current sociotechnical transformation processes that are often discussed as ‚digitalization‘ (Houben/Prietl 2018).

[2] Metcal and colleagues (2019) draw a more nuanced picture in their study of people responsible for ethics in big tech companies, detailing the heterogeneous constraints within which their work is situated and that force them to fit ethical concerns within the organizational logics dominating the Silicon Valley, namely: meritocracy, technological solutionism and market fundamentalism.

[3] As has been pointed out in organization studies there might be a considerable mismatch between talk, action and decision (Brunsson 1993) as corporate actors may state things they don’t actually act upon, in order to manage conflicting expectations, such as profit maximization and social responsibility.

[4] See: https://inventory.algorithmwatch.org/ [4th of June 2021]. The selected AI ethics-statements allow for theoretically sampling (Strauss/Corbin 1990) with respect to (a) authors/issuing organization (private sector, governmental actors, academia, civil society), (b) geopolitical reach (national, international, global) and (c) degree of compliance (binding agreement, voluntary commitment, recommendation).

[5] See the work of Joy Buolamwini online under: https://www.media.mit.edu/people/joyab/updates/ [6th of Juni 2021].

References

Adam, Alison (2008): Ethics for things. In Ethics and Information Technology 10, pp. 149–154.

Allen, Colin; Wallach, Wendell.; Smit, Iva (2006): Why Machine Ethics? In IEEE Intelligent Systems 1541-1672 (06), pp. 12–17.

Allhutter, Doris; Cech, Florian; Fischer, Fabian; Grill, Gabriel; Mager, Astrid (2020): Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. In frontiers in Big Data. doi: 10.3389/fdata.2020.00005.

Angwin, Julian; Larson, Jeff; Surya, Mattu.; Kirchner, Lauren; Parris, Terry Jr. (2016): Machine Bias. There’s software used across the country to predict future criminals. And it’s biased against blacks. In ProPublica. https://www.propublica.org/article/machine-bias-riskassess-ments -in-criminal-sentencing (26/10/2020).

Barad, Karen (2003): Posthumanist Performativity. Toward an Understanding of How Matter Comes to Matter. In Signs 28 (3), pp. 801–831.

Benkler, Yochai (2019): Don’t let industry write the rules of AI. In Nature 569 (7754), p. 161.

Bietti, Elettra (2020): From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. In Proceedings to ACM FAT* Conference 2020.

Bostrom, Nick; Yudkowsky, Eliezer (2014): The ethics of artificial intelligence. In: Frankish, Keith; Ramsey, William M. (Eds.): The Cambridge Handbook of Artificial Intelligence. The Cambridge University Press, pp. 316–334.

Bourdieu, Pierre (1987) Sozialer Sinn. Frankfurt/M.: Suhrkamp.

Brunsson, Niels (1993): Ideas and actions: Justification and hypocrisy as alternatives to control. In Accounting, Organizations and Society 18 (6), pp. 489–506.

Cath, Corinna; Wachter, Sandra; Mittelstadt, Brent; Taddeo, Mariarosaria; Floridi, Luciano (2018): Artificial Intelligence and the ‘Good Society’: the US, EU, and UK approach. In Science and Engineering Ethics 24, pp. 505–528.

Cervantes, José-Antonio; López, Sonia; Rodriguez, Luis-Felipe; Cervantes, Salvador; Cervantes, Francisco; Ramos, Félix (2019): Artificial Moral Agents: A

Survey of the Current Status. In Science and Engineering Ethics. https://doi.org/10.1007/s11948-019-00151-x (26/10/2020).

Corbett-Davies, Sam; Goel, Sharad (2018): The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. https://arxiv.org/abs/1808.00023.

Daly, Angela; Hagendorff, Thilo; Hui, Li; Mann, Monique; Marda, Vidushi; Wagner, Ben; Wang, Wei; Witteborn, Saskia (2019): Artificial Intelligence Governance and Ethics: Global Perspectives. The Chinese University of Hong Kong, Faculty of Law: Research Paper No. 2019–15.

D'Ignazio, Catherine; Klein, Lauren F. (2020): Data Feminism. Cambridge/MA: The MIT Press.

Dignum, Virgina (2018): Ethics in artificial intelligence: introduction to the special issue. In Ethics and Information Technology 20, pp. 1–3. https://doi.org/10.1007/s10676-018-9450-z (26/10/2020).

Dutton, W.H.; Reisdorf, B.C.; Blank, G.; Dubois, E.; Fernandez, L. (2019): The Internet and Access to Information about Politics: Searching through Filter Bubbles, Echo Chambers, and Disinformation. In: Graham, M; Dutton, W.H. (Eds.): Society and the Internet. Oxford: Oxford University Press, pp. 228–247.

Emirbayer, Mustafa; Mische, Ann (1998): What is Agency? In American Journal of Sociology 103 (4), pp. 962–1023.

Etzioni, Amitai; Etzioni, Oren (2017): Incorporating Ethics into Artificial Intelligence. In Journal for Ethics 21, pp. 403–418.

Filipovic, Alexander; Koska, Christopher; Paganini, Claudia (2018): Developing a Professional Ethics for Algorithmists. Gütersloh: Berteslmann Stiftung.

Floridi, Luciano; Cowls, Josh; Beltrametti, Monica; Chatila, Raja; Chazerand, Patrice; Dignum, Virginia; Luetge, Christoph; Madelin, Robert; Pagallo, Ugo; Rossi, Francesca; Schafer, Burkhard; Valcke, Peggy; Vayena Effy (2018): AI4People – An Ethical Framework for a Good AI Society. In Minds and Machines 28, pp. 689–707.

Gebru, Timnit (2020) Race and Gender. In: Dubber, Markus; Pasquale, Frank; Das, Sunit (Eds.): The Oxford Handbook on AI Ethics. DOI: 10.1093/oxfordhb/9780190067397.013.16

Greene, Daniel; Hoffmann, Anna Lauren; Stark, Luke (2019): Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, pp. 2122–2131.

Hagendorff, Thilo (2020): The Ethics of AI Ethics: An Evaluation of Guidelines. In Minds & Machines 30, pp. 99–120.

Haraway, Donna (2004 [1985]): A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s. In Haraway, Donna (Ed.): The Haraway Reader. New York: Routledge, pp. 7–45.

Hoffmann, Anna Lauren (2019): Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. In Information, Communication & Society 22 (7), pp. 900–915.

Houben, Daniel; Prietl, Bianca (Eds.) (2018): Datengesellschaft. Einsichten in die Datafizierung des Sozialen. Bielefeld: transcript.

Jaume-Palasi, Lorena (2019): Why We Are Failing to Understand the Societal Impact of Artificial Intelligence. In Social Research: An International Quarterly 86 (2), pp. 477–498.

Jobin, Anna; Ienca, Marcello; Vayena, Effy (2019): Artificial Intelligence: the global landscape of ethics guidelines. In Nature Machine Intelligence 1, pp. 389–399.

Leighton, Andrews; Benbouzid, Bilel; Brice, Jeremy; Bygrave, Lee A.; Demortain, David; Griffiths, Alex; Lodge, Martin; Mennicken, Andrea; Yeung, Karen (2017): Algorithmic Regulation. London: LSE Discussion Paper 85.

Lopez, Paola (2019): Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. In: Proceedings of the 18th Annual IAS-STS Conference on Critical Issues in Science and Technology Studies, pp. 289-309 DOI: 10.3217/978-3-85125-668-0-16.

Lyon, David (2004): Globalizing Surveillance: Comparative and Sociological Perspectives. In International Sociology 19, pp. 135–149.

McKay, Tate; Yallaly, Patrick (2017): Algorithms, Organizations, and Ethics. Eastern Illinois University: MBA 5680 Organizational Behavior.

Metcal, Jacob; Moss, Emanuel; boyd, danah (2019): Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. In Social Research: An International Quaterly 86 (2), pp. 449–476.

Noble, Safiya U. (2018): Algorithms of Oppression: How Search Engines Reinforce Racism. New York: New York University Press.

Nosthoff, Anna-Verena; Maschewski, Felix (2019): Alles nur Fake Ethik. In: REPUBLIK, 22.05.2019. https://www.republik.ch/2019/05/22/alles-nur-fake-ethik.

O'Neil, Cathy (2016): Weapons of Math Destruction. Random House Publications.

Pariser, Eli (2011): The Filter Bubble. How the New Personalized Web is Changing What We Read and How We Think. London: Penguin Books.

Paulitz, Tanja (2005): Netzsubjektivität/en. Konstruktionen von Vernetzung als Technologien des sozialen Selbst. Münster: Dampfboot.

Pörksen, Bernhard (2018) Filter Clash. Die große Gereiztheit der vernetzten Welt. In: re:publica 18. https://www.youtube.com/watch?v=o3ei8qVgTtc.

Prietl, Bianca (2019a): Algorithmische Entscheidungssysteme revisited: Wie Maschinen gesellschaftliche Herrschaftsverhältnisse reproduzieren können. In feministische Studien 2, pp. 303–319.

Prietl, Bianca (2019b): Die Versprechen von Big Data im Spiegel feministischer Rationalitätskritik. In GENDER 3, pp. 11–25.

Schiff, Daniel; Borenstein, Jason; Biddle, Justin; Laas, Kelly (2000): What's Next for AI Ethics, Policy, and Governance? A Global Overview. In AIES '20. https://osf.io/preprints/socarxiv/8jaz4/.

Sloane, Mona (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. https://doi.org/10.34669/wi.cp/2.9.

Srnicek, Nick (2018): Platform Monopolies and the Political Economy of AI. In McDonnell, J. (Ed.): Economics for the Many. London: Verso, pp. 152–163.

Strauss, Anselm L.; Corbin, Juliet (1990): Grundlagen Qualitativer Sozialforschung. Weinheim: Beltz.

Véliz, Carissa (2021): Privacy is Power. Why and how you should take back control of your data. London: Penguin Books.

Wachter, Sandra; Mittelstadt, Brent; Russell, Chris (2021): Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. In West Virginia Law Review, forthcoming.

Weber, Max (2008 [1920]): Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie. Frankfurt/Main: Zweitausendeins.

Weber, Jutta; Prietl, Bianca (2021): AI in the Age of Technoscience. On the Rise of Data-Driven AI and its Epistem-Ontological Foundations. In Elliott, Anthony (Ed.): The Routledge Social Science Handbook of AI. New York: Routledge, forthcoming.

Winner, Landon (1980): Do Artifacts Have Politics? In Daedalus 109, pp. 121–136.

Woolley, Samuel C. (2016): Automating power: Social bot interference in global politics. In First Monday 21 (4). https://doi.org/10.5210/fm.v21i4.6161 (26/10/2020).

Yeung, Karen; Howes, Andrew; Pogrebna, Ganna (2020): AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing. In: In: Dubber, Markus; Pasquale, Frank; Das, Sunit (Eds.): The Oxford Handbook on AI Ethics. DOI: 10.1093/oxfordhb/9780190067397.013.5

Zou, James; Schiebinger, Londa (2018): Design AI so that it’s fair. In Nature 559, pp. 324–326.