What is AI Ethics? Ethics as means of self-regulation and the need for critical reflection

Suzana Alpsancar, Paderborn University (Germany)

International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Abstract

In the wake of the recent digital transformation, AI ethics has been put into practice as a means of self-regulation. Current initiatives of ethical self-regulation can be distinguished into different ethical practices, namely ethics as rule setting (codes of conduct), ethics as rule following (value-oriented development), and ethics as rule compliance checking (boards and audits). Drawing from the literature, I demonstrate that these forms of AI ethics are in constant need of normative reflection and deliberation albeit the structural conditions under which they are enacted give very little room to do so. Accordingly, the AI community should think more about how to establish institutional frameworks that can be conducive for cultivating ethics as critical reflection and deliberation.

Keywords: Ethical guidelines, ethical algorithms, ethical reflection and deliberation, abstraction error, procedural ethics, designing for sociotechnical systems, uncertainty, social coping

1 Introduction

Ethical debates around AI development have long mostly been hypothetical (Floridi 2021). Today, still hypothetical in parts, ethical reflection faces practical questions regarding the actual "design, use, and longer-term impacts" of AI systems (Prem 2023, p. 1). The EU's High Level Expert Group (2019a, p. 6) defines AI as "software (and possibly also hardware) systems designed by humans that, given a specific goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected [...] data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal". AI is increasingly deployed in various societal fields, usually not as a stand-alone technique but as a component within larger systems. A series of reports by the White House and the EU has framed AI as both a key future technology as well as an ethical concern (European Commission 2014; Fox-Skelly et al. 2020; Miguel Beriain et al. 2022; Muñoz, Smith, and Patil 2016; Podesta et al. 2014). Particularly, the transformation of the Internet from a largely distributed communication structure to a commercialized space controlled in large parts by platform gatekeepers (Pasquale 2015; Schiller 1999; Zuboff 2019) as well as the expanding employment of algorithmic decision-making systems (ADM) in various societal fields has drawn much public attention. A plethora of scandals including debates on the recidivism prediction software COMPAS (Angwin et al. 2016) or Amazon's hiring algorithm (Dastin 2018) have shown that, albeit meant to balance insufficient or unjust decisions due to human epistemic limits and biases, ADM is not free from biases and discrimination per se (Creel and Hellman 2022; Lepri et al. 2018). The *Cambridge Analytica* scandal has led to question the harm political microtargeting in digital networks might bring to democratic societies (Dowling 2022; Hu 2020).

Against this background, the call for ethical orientation for AI development became vocal. Today, AI Ethics seem to be everywhere. It is practiced in various forms, by means of various tools and strategies, under differing institutional conditions, and by various experts coming from different professional or academic backgrounds. Most often, ethical initiatives are installed as a means of self-regulation by and for the AI community. In this paper, I argue that given the conditions and forms of ethics as self-regulation is currently enacted, it must fail in what it is meant to do – providing substantive orientation for AI development. To do so, I first take a step back and reflect why we, as a political community, should care about understanding, discussing and deliberating about AI

development (section 2). This reflection serves to put my systematic distinction of three practices of ethical self regulation and their corresponding understanding of ethics into perspective, namely ethics as 'rule setting' (section 3), ethics as 'rule enforcing' (section 4), and ethics as 'rule compliance checking' (section 5). Drawing from critical investigations of these ethical practices, I demonstrate that these initiatives must fail in providing a substantive ethical orientation not only due to a lacking regulatory framework but also due to a limited understanding of what ethics can and cannot do. Finally, I systematize current suggestions on how to cultivate a more substantial practice and corresponding understanding of AI Ethics as critical reflection and deliberation (section 6).

2 The power of technology – why we should care

The current AI ethics initiatives must largely be seen as corporate or political responses to the aforementioned scandals and their related public insecurity towards recent techno-industrial developments. Modern technology in general, and data-driven AI systems in particular, hold great potential to alter our lifeworlds by opening up new and different spaces of practices, while also closing alternative forms of doing things (or by making some more attractive than others). Power of technology generally manifests in different forms and is effective in various ways. Hildebrandt (2015, p. 10) describes it as a type of regulation of people's behavior. She invites us to contrast regulation by technologies with regulation by law in democratic constitutions, where the power of laws is legitimized by three basic normative principles: "Self-regulation, disobedience, and contestability". These three principles correspond to the separation of powers into law-making institutions (legislative), law-enforcing bodies (executive), and the judicial system (judiciary). The idea of self-regulation in the sense of political autonomy is guaranteed by forms of representation and participation in legislation. The possibility of disobedience exists as long as there is room to not follow the law (or social norms in general), including variations of what it means to obey and disobey a given rule. The ability to claim one's rights is ensured by the institutions of independent jurisdiction. Deliberating political issues in the public sphere is crucial for people's political and moral autonomy (Christman 2020). Indeed, liberal and critical social philosophers alike find disobedience and contestability as political practices mandatory for social change and open societies, i.e., the possibility for civil movements and actors to alter existing institutions and to stipulate normative discourses (Celikates 2016; Fiedler 2009; Sabl 2001).

Regulation by technology differs from regulation by law in all three accounts: First, technological regulation does not derive from a democratic authority. There is neither a genuine act of rule-making nor a legitimizing institution. The way technologies shape and change our practices depends on several interdependent factors of which the design decisions of developers (Verbeek 2006), the contextual conditions of developing and deploying (Mateescu and Elish 2019; Schneider 2020), as well as the actual, collective forms of adoption and appropriation (Moreno Gálvez and Sierra Caballero 2022; Rojas and Chalmers 2009) are crucial. Whether and how technologies are used can coincide with what the developers have intended or not.

Second, we usually have options to not follow social rules (including laws) for more or less good reasons. You may ignore a red light because you are in a hurry or because you want to help someone who is lying on the road. Whether we morally approve of breaking the rules or not depends on situational reasons. What is crucial here, is that we have in principle the possibility to break social rules, reinterpret them and to reason about their situational adequacy. When using technology, we are most often forced to follow the technical rules they bring with them, at times with a given set of options. We usually have no way of (re-)assessing the adequateness of the given technical rules, neither in the moment of usage nor in general. In theory, we are left with the choice of not using a certain technology at all, but in practice, opting-out becomes impossible as soon as their usage has become socially normal.

Third, the power of the state can be challenged in principle, both in the sense of a specific case and in the sense of political discourse about the goodness of given rules. In contrast, technological regulation is hard to contest, because "the technological defaults that regulate our lives [...] are often invisible and because most of the time there is no jurisdiction and no court" (Hildebrandt 2015, p. 12). Moreover, because there is no recognizable agent responsible to set the rules and because there is no strict causal institutionalized form of how rules are put into practice – as they are in the legal case – it is also unclear who to contest and how (Nersessian and Mancha 2020; Wachter, Mittelstadt, and Russell

2021). Because of this, we are facing an "accountability gap" (Santoni de Sio and Mecacci 2021).

Now, whether we find Hildebrandt (2015)'s comparison informative that should depend on if the comparison to *the legitimation* of power seems plausible or not. As technical products and technological feasibility belong to the societal fields of economics and research, it could seem counterintuitive to raise a comparison of such high standards for legitimizing power. However, it is precisely because of the weight of the power with which AI technologies alter and affect everyday lives, that Hildebrandt's suggestion seems appropriate. This is because in the case of the tech industry, we

are facing an economic power which seems as effective as state regulations. Namely, the market power of Alphabet (Google), Amazon, Microsoft, Apple, and Meta (Facebook) is discussed in terms of competition law, often drawing from the historical example of breaking "Big Oil" over hundred years ago (Akman 2019; Birch and Cochrane 2022; Moore and Tambini 2018). It is mostly these 'Big Five', among some other players whose services have transformed the Internet from a distributed, fractal, and open network in the 1990s towards a commercialized "hierarchical ecosystem ruled by a few gatekeepers" (Bietti 2023, p. 1). Considering these companies' immense power, that is why we should care about their influence on people's behavior and the practices and structures of our political community (Apostolicas 2019; Motupalli 2017).

In addition, there is a technological reason to call for democratic legitimization of 'AI's power', namely the fact that some AI applications can change peoples' lives decisively. In this respect, the European Commission has lately proposed a risk-based regulatory framework defining four risk categories: unacceptable risk, high risk, limited risk, and minimal or no risk. Application with unacceptable risk are to be banned within the EU, applications with minimal or no risk are allowed "free use", e.g., AI enabled video games or spam filters (European Commission 2022). Applications with limited risk should meet specific transparency requirements, e.g., users should be able to tell if they are chatting with an artificial chatbot or another person. Of particular interest are those applications falling under the category of high risk, including critical infrastructures (transport), educational or vocational training, safety components and products, but also AI used for employment, and management of workers among others. A lot of the debate around ADM relates to its application being regarded as high risk by the EU. Due to the potential great influence on people's lives, we are eager to question how that power can be legitimized.

3 Ethical Guidelines – setting the rules

When ethics is currently thought of as self-regulation, what is meant is not the democratic principle of self-regulation of a people but that of an industry. Accordingly, the idea that addressees and of the given rules correspond does not hold (Maas 2022). In the early 2000s, state and industry 'agreed' on the principle of self-regulation to govern the latest digital transformation and to set a value-oriented framework for the further development of the digital economy (Floridi 2021). It is the form of governance where all the power of the rules lies within and comes from within the industry itself (Tworek 2019, p. 100). Following Hildebrandt (2015)'s comparison, I describe, contextualize, and critically discuss how ethics is used to give value to this policy tool of academic-industrial self-regulation by setting rules (codes of conduct), by enforcing the declared rules (via design or corporate organization), and by checking compliance with these rules (ethical boards).

3.1 The normative assumptions of AI codes

In principle, ethical guidelines or codes of conduct can serve as a means of self-regulation and as a means for holding others (e.g., companies) not legally but ethically accountable for their actions. Compliance with these codes and guidelines cannot be sued in court but rests on self-commitment. They might promote acting morally beyond the legal minimum, be it in terms of 'good citizenship' or for 'competitive acceleration' (Floridi 2021). They can complement legal regulation by providing guidance for instances not covered by current law (Hildebrandt 2020) or serve as a means at hand where legal regulation is still missing, because advances in technical development have outpaced the

law (Lepri et al. 2018). In 2019, there were already more than 80 ethical guidelines or AI codes publicly available (Jobin, Ienca, and Vayena 2019; Morley et al. 2020), coming from industrial associations such as the IEEE or the ACM, from business such as IBM or Google, or from governmental institutions such as the EU's High Level Expert Group (2019b). In contrast to the sphere of law, there are no institutionalized meta-rules for industrial self-governance, leaving questions like the following unanswered: Who is authorized to declare such codes and based on what traditions? Who is meant to follow it and why? What does it practically mean to act in accordance with a code? What happens if someone does not follow the code?

In their study of prominent AI codes, Greene, Hoffmann, and Stark (2019, p. 2122) disclose a common "moral background" of these declarations. They analyzed the Partnership on AI to Benefit People and Society (a non-profit corporation consisting of the Big Five and some higher-education institutions, civil rights groups, and other industry partners), The Montreal Declaration, The Toronto Declaration, Open AI, as well as the FATML community and Axon's AI Ethics Board for Public Safety. They have found that their statements impose a universalist account of ethics as they all assume that "the positive and negative impacts of AI are a matter of universal concern" and are to be "addressed by objectively measuring those impacts" (ibid., p. 2126). This understanding tends to neglect differences in how different social groups benefit or experience harm due to AI applications and differences in how they (should) care. Next, while all of humanity is seen as the moral subject, moral agency is ascribed to experts alone. The tools of declaring codes, setting up ethical boards, or hiring ethical staff are strategies of expert oversights. The most noteworthy finding is what they call "values-driven determinism", which implies several deterministic assumptions: First, AI is simply presumed as 'coming' (over us), and society is seen as something that needs to react to this quasi-natural development. As AI is coming over us, we only have two options: a wild and unregulated spread of these technologies or the ethical path of taming the per-se occurring development, that is steering it to built better AI by ethical means (ibid., p. 2127). This assumption not only neglects human and social agency within the development process, it also homogenizes AI as if it were a straightforward thing. In reality, 'AI' is an umbrella term for heterogeneous categories, e.g., specific techniques, systems, software, knowledge, research fields, sociotechnical systems etc. Second, there is an implicit tension between this deterministic view and the confession to a value-oriented design, which actually implies that there are choices of how to design certain things including the choice to not purse a certain business model. Third, the commitment to a value-driven design still often leads to focus on the process alone as the place for ethical considerations. This, however, would only make sense if it was reasonable to imply a design-determinism for the usage of technology – what has been proven wrong many times (Ackerman 2000; Oudshoorn and Pinch 2003; Suchman 2007).

3.2 Practical and conceptual pitfalls

Given the different character of those who have declared ethical guidelines for AI, the convergence of core principles (i.e., autonomy/accountability, non-maleficence, fairness, explicability/transparency, and privacy; Jobin, Ienca, and Vayena 2019; Hagendorff 2020; Floridi et al. 2018) seems at first remarkable. However, focusing on core principles for ethical AI reflects the mainstream approach in medical and bioethics, so-called *principlism* (Beauchamp and Childress 2013), which emerged in the late 1970s in response to "ethically dubious medical research" (Prem 2023, p. 3; Shea 2020). Comparing principlism in medical and AI ethics, Mittelstadt (2019) argues that principlism can only be effective in health care because it rests on a traditional professional identity and is backed up by a rather strict regulatory framework. Medicine is guided by the common aim to prevent harm and do good, to promote health, albeit definitions of health and how to pursue it might differ among experts and professionals. AI development is not comparable in this regard. There are no well-established or widely shared "norms of good practice" (ibid., p. 503) within software engineering, which is not a formal profession but rather a heterogeneous field including various types of expertise, jobs, and professional practices. Whereas public as well as private medical institutions operate within a clear regulatory framework which protects patients and research participants from being merely subjected to corporate interests, the field of AI deployment is mostly characterized by the absence of such regulatory frameworks, despite some existing legal initiatives such as the GDPR (in force since 2016), followed by the Digital Markets Act and the Digital Service Act (in force since 2022), as well as the proposed AI Act within the EU.

Moreover, because of the hierarchical structure of medical decision making, it is usually clear who can be held accountable for what. The causal relation between, let's say a doctor's prescription of a certain medicine, and the effect on the patient is rather straight forward. There is no comparable clear picture in terms of agency, responsibility and causal effects in developing and operating AI software or robotics. This is to a large part due to international division of labor and multi-agent development paths (Mittelstadt 2019, p. 502; Sollie 2007; Gogoll et al. 2021), because the "profusion of agents obscures the location of agency" (Thompson 2017, p. 32) in absence of strict hierarchies. Another uncertainty factor is the open-endedness of AI enabled software: Many techniques, models or tools can be used in various contexts and for various purposes, e.g., a classifier or scoring algorithm can be used for personalized advertisement, credit-loaning, human-resource management etc. While many design decisions will most likely have some sort of impact, it is not easy to tell which and how. The picture complicates insofar AI systems are meant to continuously adapt their behavior in regard of their contextual input, such as environmental data or users' behavior. Consequently, it is a task of its own to determine, (a) if a design decision holds ethical impact or not, if yes then (b) how design choices might affect different groups, and finally (c) how to organize accountability for these choices given the network of 'multi-hands' involved (Nissenbaum 1994; Mittelstadt 2019, p. 503). While medical practitioners can stem from their professional identity and their regulatory frameworks to orient their actions and decisions, AI practitioners are left alone with little more than these declarations of abstract principles. AI codes of conduct overlook the specifics of contexts of use and manufacture (Prem 2023, p. 4) – like all codes do. But because AI codes are not better integrated, they must remain ineffective. This gap between principles and practice (Hallensleben et al. 2020; Shneiderman 2020) leaves too much room for arbitrary interpretations such as "cherry-picking Ethics" (randomly choosing which set of values fits the given case), "risk of indifference" (choosing the code which best justifies your own behavior), "ex-post orientation" (the focus on values only might draw attention away from broader normative issues of contexts and ways of life), as well as "the desire for gut feeling" (deciding on what your gut tells you in absent of any rationalizable decision making process, Gogoll et al. 2021, pp. 1097–1098).

In consequence, AI codes have turned out to be rather toothless (McNamara, Smith, and Murphy Hill 2018; Munn 2022; Popescu et al. 2016; Rességuier and Rodrigues 2020; Schwartz 2004). They have also been criticized as *Ethics-Washing* (Wagner 2018; Yeung, Howes, and Pogrebna 2020; Bietti 2020), not only because some companies exploit the lack of hard regulations in their favor (Floridi 2021, p. 620), but also because as industrial initiatives they have been interpreted as immunizing companies against public scrutiny. The superficial consensus on core principles overshadows political and normative conflicts instead of opening them up to public debate (Mittelstadt 2019, p. 501).

4 Ethical design – enforcing the rules

AI codes were meant to orient designing, applying, and managing digital technologies in an ethical way, thereby implying that dealing with AI-systems can be more or less ethical. Here, the word "ethics" is used as an attribute to qualify something as morally acceptable or 'good'. The key question is what we can reasonably attribute as good or ethical and how to achieve it. Traditionally, normative theories have presupposed different aspects of human agency (within a political community) as decisive, for example the virtuousness of a person, the goodness of the will or the consequences of an action. In Ethics of Technology, it is common ground to focus on the 'morality of the technology' either in terms of qualifying the performance or outcome of a technological system as (non-)ethical or in terms of how the performance or outcome of a technological system mediates human agency. This evaluation as ethical appeals directly to the call to consider human values (Bynum 2018), which conceptually links the declaration of core values in AI codes to the orientation towards these values in the development of AI. I critically discuss three common approaches to do so.

4.1 Designing artificial moral agents

One strand of designing ethical AI presumes the aforementioned 'value-determinism' as if the ethical quality could be *completely realized and fully ensured by the system's design*. In this view, AI agents can (to some respect) function as *artificial moral agents* (Dignum 2019, p. 81). Wallach and Allen (2008) classified three approaches to build such "Moral Machines": the top-down approach, the bottom-up

approach, and hybrid approaches. The basic idea in the first is to implement ethical behavior into the technical system by formalizing an ethical theory or principle. Bottom-up approaches strive to develop AI-systems that "can learn from the environment or from a set of examples what is ethically right and wrong" (Wong and Simon 2020, p. 3). Hybrid approaches combine strategies and techniques from both. However, these attempts face various theoretical and technical limitations: Top-down approaches presume that there could be one, and one only, adequate ethical theory for a given problem/application sector. In reality, there is no uncontroversial ethical principle and no universal ethical theory – what counts as ethical is always contestable and subject to the openness of societal developments (Aristotle 2013; Moore 1996). Furthermore, it is a challenge in itself to specify and formalize ethical theories so that they can be implemented as technical rules in AI-systems. In consequence, top-down approaches inherit the risk of being built on inadequate or false foundations (Wong and Simon 2020, p. 3). Bottom-up approaches "infer what is ethical from what is *popular* " (ibid., p. 3) hereby confusing what is right with what the majority thinks. It is unclear how hybrid approaches could overcome these challenges. Even more importantly, all three approaches lack ethical *justifications*, which provide good reasons why some choices were made over others (Dignum 2019). On top of these theoretical limitations, the technical limitation stems from the challenge to "effectively discern *ethical relevant* from *ethical irrelevant* information among a multitude of information available within a given context" (Wong and Simon 2020, p. 3), which again would then be in need of an ethical justification.

In addition, ethical-political concerns relate to protecting human autonomy and keeping responsibility manageable. In the case of artificial moral agents it would be unclear "who or what should be responsible for wrongful decisions of autonomous AI " (ibid., pp. 3–4). Additionally, artificial moral agents could significantly undermine human autonomy because the decisions made by them *for us* or *about us* will be beyond our control, thereby reducing our independence from external influences" (ibid., pp. 3–4). As an alternative strategy, efforts are being made to leave humans in control (Bryson and Theodorou 2019; Koulu 2020; Santoni de Sio and Hoven 2018). Here, AI is regarded either as a product whose properties may be ethically relevant and/or AI is seen as an artificial agent that mediates moral behavior of individuals and collectives (Verbeek 2005).

4.2 Designing mediating agents

A wide spread approach to respect values in design without presuming a straight value-determinism are so-called Value-Sensitive-Design (VSD) frameworks, which emerged in the 1980s from information system research (Friedman and Nissenbaum 1996; van der Hoven and Manders-Huits 2020). The basic idea is that “a given technology is more suitable for certain activities and more readily supports certain values while rendering other activities and values more difficult to realize" (Friedman, Kahn, and Borning 2002, p. 3). Because there is no deductive way to implement values in design requirements, respecting values in design is challenging. They must be conceptualized and specified for each case. This ’translation’ process is per se contestable as it involves multiple plausible interpretations (Hallensleben et al. 2020; van de Poel 2013). Different heuristics in the literature of how to systematically do so, agree that the best you can do is to explicate the value-relations and inferences and to open these up for ethical deliberation. While VSD has proven to be a most promising approach for the first aspect (explicating) it has been falling short in justifying the explicated interpretation choices (Simon 2017, p. 226; Winkler and Spiekermann 2021, p. 18) as well as in ensuring that the process of deliberation is legitimate (Dignum 2019, p. 85; Friedman, Harbers, et al. 2021). Critics of VSD have therefore called to enhance the approach. First, the value-choices need to be connected to normative theories to argue why certain values should be respected or not (Cenci and Cawthorne 2020; Manders-Huits 2011), how they can be meaningfully specified in a given case, and how they should be prioritized with regard to conflicting interests and value trade-offs (Hubig and Reidel 2003; Peylo et al. 2022). Second, VSD has traditionally been limited to an analysis of the designers involved in technology development and their guiding ideas and implementations. As a consequence, there was little reflection and justification about the selected stakeholders and respectively identified values (Winkler and Spiekermann 2021, p. 19). To prevent this, it is important to include all involved and all affected people, which means linking the task of value-oriented design to the task of deciding which values to respect; starting with the questions whose values and traditions are meant to be respected and why (Fox et al. 2017; Irani et al. 2010; Jacobs et al. 2021; Mainsah and Morrison 2014; van Norren 2023;

Wynsberghe 2013). Third, professional agency, hence responsibility, needs to be

organized and enacted as it does not evolve ’naturally’ out of what individual agents do in widely distributed actor-networks for software engineering. Here, it is crucial to sort out who should be responsible for what scope and decisions. Within a company, the strategic management most likely decides whether or not to pursue a certain business model and in which way while the development team will most likely have "some leeway in deciding how to exactly develop the product" (Gogoll et al. 2021, p. 1087). Fourth, the development of AI is particularly subject to a multitude of uncertainties, even beyond the user and multi-agent development trajectories, due to specific AI-features such as their context-sensitivity and continuous model adaption (Muschalik et al. 2022; Shaheen et al. 2022; van de Poel 2020). Because of that, we should prepare for the fallibility of our today’s decisions and regard our ethical assessment as provisional (Campbell 2006; Hubig 2007) that might become in need of a reassessment in the future (Rahwan 2018).

While the VSD-approach can be optimized in these three regards, it can only be as valuable as the circumstances allow. What we can learn from the Mittelstadt (2019)’s comparison with the field of health care is that without a professional self-image and a clear regulatory framework, ethical self-regulation remains arbitrary and subject to existing power relations.

4.3 Hiring ethics staff

The VSD approaches (or familiar heuristics such as participatory design, responsible research and innovation) are well-known in academic contexts, but less so beyond (ibid., p. 504). In commercial settings, ethics come at a cost (time, engagement, restrictions, etc.). Accordingly, "it cannot be assumed that value-conscious frameworks will be meaningfully implemented in commercial processes that value efficiency, speed and profit" (ibid., p. 504). Metcalf, Moss, and Boyd (2019) explore practices and structures of corporate ’doing ethics’ in Silicon Valley. In doing so, they draw attention to the fact that the tech industry itself plays a decisive role in shaping the notion of what ethics is, can, and should be by creating positions for ethicists. There is no explicit definition nor consensus of what ethics is supposed to be, how it works, and where it is located in these corporate organizations, but there is a common negative denominator: ethics workers are not hired out of a sense of duty or a calling to do good but to prevent worse: a bad reputation, tougher legal sanctions, or regulation (ibid., p. 459).

Metcalf, Moss, and Boyd (ibid.) link this ’negative’ understanding of ethics to Silicon Valley’s moral background, which is built from three fundamental norms: meritocracy, technological solutionism, and market fundamentalism. In a meritocracy power is earned and legitimized by individual achievements based on individual abilities and does not stem from social or other capital, such as traditions or collaborations. Silicon Valley’s expression of this cultural self-image is the announcement to hire and train only the best, resulting in an elitist community of high performers. Consequently, the tech industry wants to solve every problem by itself, neglecting outside criticism or help (ibid., p. 462). This self-image resonates with the neoliberal self (Bhatia and Priya 2018; Davidson 2011; Rose 1996) and makes it hard to locate responsibility other than as an individual task, as it tends to ignore structural conditions. It also fails to appreciate the diversity of different challenges, which is connected to the second norm, technological solutionism. Technological solutionism presumes that all thinkable problems can be solved by technology. This idea presupposes that ethical products are in general feasible, hence the possibility to engineer for the social good. It also implies that ethical challenges derive from imperfect technical problem solving, and can somehow be repaired or debugged, they can even "be ’solved’ once and for all" (Metcalf, Moss, and boyd 2019, p. 463).

The third identified norm, market fundamentalism, indicates that everything imaginable and doable is bound to the limits of the idea of a free market. Thus, ethical tools are not to interfere with companies' “bottom lines”: Employees cannot act against or outside the corporate rules of the game. Their efforts can only be as good as the respective political-economic and corporate-cultural framework allows. Given the competitive nature of most companies' behavior, there is no exchange of best practice or failures among the AI industry. As there is practically no option for smaller companies to set up their own ethical tools, they are going to adopt techniques developed by Big Tech. Caught in this framework, ethical reflection loses its critical sting. Instead of pointing to alternative possibilities, ethics turns into the opposite, it promotes the status quo of entrepreneurial action.

5 Ethical boards – rule compliance checking

Over the last two decades, many highly prolific companies have also set up ethical advisory boards for numerous concerns. Audits and boards are institutionalized ways of reviewing whether or not given rules are being followed. They can be set-up internally or brought in as external bodies (Raji et al. 2020, p. 35). Among the many prominent ones, Meta's Oversight Board is a telling example, active since 2020 and set-up to review content moderation decisions. Many contentious cases are sparked by the conflict between the right to free speech and the protection of human rights or personal rights (e.g., hate speech). There is also the problem of political manipulation sparked by misinformation such as doctored videos of public figures or other fake news (Klonick 2020; Neuvonen and Sirkkunen 2022). Posts are generally regulated by Meta's community rules which are being enforced by both machines and people. The Board is a meta-institution to control the rule-following. It can be seen as a regulatory intermediary (Medzini and Levi-Faur 2023) as it is meant to include outside opinions but was still set-up by the company itself. After announcing its formation, the company consulted with numerous different stakeholders on how to put the board together. The result is a group of 40 external experts, mainly with scientific, legal or activist backgrounds.

Although the board is financially independent from Meta, critics still question examples of such enhanced governance regimes for structural reasons. When board members hold contractual relationships with the company, a power asymmetry is created that questions the boards structural independence, e.g., "it is unclear whether Facebook's competitors can sign contracts with the board" (ibid., p. 19). Some argue, the accountability of these boards refer back to the accountability of their initiator. Namely, Mark Zuckerberg had drawn scepticism in this regard in the wake of the Cambridge Analytica scandal because he had "declined to testify in the United Kingdom and Canada, despite subpoenas" (Tworek 2019, p. 98). Moreover, the Board's work added to draw all attention towards the issue of moderating online content, with this obscuring other concerns, e.g., a public debate on how Meta develops its Newsfeed algorithms or ethically assessing personalized content and advertisements (Bietti 2020, p. 215). Critics point out that with its existence, Meta has found an effective way of immunizing the company against other forms of inquiry and public or third party investigations, or regulations.

In the end, industrial initiatives cannot overcome their asymmetric power structure on their own. Ethics as self-governance can only add to a sophisticated ethical reflection on the benefits, harms, and power of new technologies if linked with other regulatory measures.

6 Ethics as deliberation – How to curate the digital transformation?

Establishing ethical boards, issuing codes of conduct, or hiring ethicists is a "double-edged sword" (Bietti 2023). On the one hand, it indicates (at least to a certain degree) awareness of ethical issues, acknowledges the importance of moral concerns for the company's goals, and shows a certain willingness to face societal challenges and public criticism. But insofar as these corporate attempts serve as a means to pursue the companies' very own interests – and not the common good – they hereby simplify "the value of ethical work" (Bietti 2020, p. 210). Bietti (ibid.) takes a plea for a more substantive view of moral philosophy as a distinctive mode of inquiry into given practices, norms, situations, intuitions, or individual behavior and calls for dialogues on its moral quality in light of the complex reality constituted by legal, political, and economic structures. Moral philosophy is sometimes criticized as being too abstract and far away from real-world problems. However, it can also be seen as an offer to think slowly(er) and step back from short-headed decisions and pragmatic pressures (ibid., p. 213). In this spirit, I conclude with a systematization of current suggestions of how to turn from checklist-ethics to more procedural, reflective, and substantive practices.

Ethics should be understood as **an art of deliberation and reflection**. Ethical challenges call for normative considerations and are not solvable by technical means only. They should neither be solely conceptualized as individual failures or as collective failures, but as something that can and needs to be addressed on an organizational (Mittelstadt 2019, p. 505) and structural as well as on an individual level (Gogoll et al. 2021).

On the individual level, ethical deliberation and reflection is not something you can execute at the push of a button. Rather, there is a need to **cultivate a rational habit** "at the base of the development" (Gogoll et al. 2021). For example, engineers need to feel entitled and obliged to "be value-sensitive" (ibid., p. 1102). But they also need to find their practices situated within proper structural conditions that allow and give space for valuable ethical practices, so they don't face a personal tragedy like many whistleblowers did (Davis 1996; Hunt and Ferrario 2022). Many design decisions will potentially be ethically relevant without the optimal solution being clear, meaning that ethically correct assessments must be made for each specific case. Only cultivating "a way of dealing with normative matters" (Gogoll et al. 2021, p. 1101) will empower individuals in uncertain situations.

There are many unresolved points in the heterogeneous field of AI today and a major task would be to **(re-)negotiate agency**: Who is effectively making the design decisions on what terms and grounds? Whose voices are included in the interpretation of ethical values? Who is authorized to declare what rules to follow, for whom, and with what force? Who should be held accountable for the software-design? How can users or other affected people hold the responsible actors accountable? Depending on the potential impacts, design decisions should be opened up for all relevant stakeholders, e.g., via participatory design methods. At least, they must be justifiable (Dignum 2019, p. 82). For high risk applications they should be subject to public scrutiny and they should include **the possibility to re-assess** them later on.

Enhancing VSD approaches, values could be understood as hypotheses, which highlights that their interpretation has to be tested within a given situation. Respectively, design situations can be better understood against different 'storied' understandings of values: These cannot only be translated into statements but should be illustrated in stories to make sense of problematic, uncertain design situations (JafariNaimi, Nathan, and Hargraves 2015). To do so, interdisciplinary research teams (Lepri et al. 2018) seem helpful to bring in different perspectives. Last but not least, there is a call to engage in **public discourse**, because this is where the ends, benefits, harms, and risk of AI should be evaluated in a more general sense: What do we wish to use some concrete possibility x for? Do we need these new technologies at all? Which problems can they help to solve, how and at what cost?

As the discussion above has shown, ethical practices can only be as good as their circumstances allow. Much of today's concerns are caused by the state-like power of US Big Tech companies or China's corporate-state AI alliance. Many voices are calling for **clearer and stricter regulation** (Bietti 2023; Farthing and Sooriyakumaran 2021; Hildebrandt 2015; Ma 2021) not only to balance the economic dominance of a few players but also to provide further orientation on how to categorize and understand what we are dealing with in the field of AI. The EU's risk **categorization of applications** points in the right direction, but there are still many unanswered questions to be resolved resulting from AI's open-endedness, multi-agent development paths, and other sources of uncertainty. The comparison to medical ethics can serve as a blueprint for integrating some of the ethical initiatives. In analogy to the risk-categorization of AI, there could be a risk-categorization of developing different AI-components, e.g., partly professionalizing software engineering equivalent to other "high-risk professions" (Mittelstadt 2019, p. 505). Some argue for licensing the work of building software for the public sector, e.g., face recognition systems for policing and alike.

A key advantage of different types of industrial self-regulation is time. This governance form allows it to react more timely and flexible, thereby granting (some) users' appeals (Floridi 2021) without putting too much restriction on still emerging markets. However, we might consider other types of leveling the time asymmetry between technological innovation and societal understanding. For instance, van de Poel (2016, p. 684) suggests to frame the introduction of new technologies as **social experiments**, hereby acknowledging that we can "only experimentally and gradually find out some of the social consequences of these technologies". This would mean intentionally setting up provisional rules, space and time to try-out *if* and *how* different people and groups can, want, or should adapt to new technical possibilities on what costs. Within such a space for testing and understanding new potentials of AI, society could profit from ethical reflection and deliberation as an art to think slower than companies run for innovation. In this form, ethics could add to a real democratic governance of AI.

References

Ackerman, Mark S. (2000). "The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility". In: *Human–Computer Interaction* 15.2-3, pp. 179–203. doi: 10.1207/S15327051HCI1523_5.

Akman, Pınar (2019). "An agenda for competition law and policy in the digital economy". In: *Journal of European Competition Law & Practice* 10.10, pp. 589–590.

Angwin, Julia et al. (2016). "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks." In: *ProPublica*. url: https://www.propublica.org/article/machine- bias- risk- assessments- in- criminal- sentencing (visited on 03/27/2022).

Apostolicas, Paul (2019). "Silicon States: How Tech Titans are Acquiring State-like Powers". In: *Harvard International Review* 40.4, pp. 18–21.

Aristotle (2013). *Nicomachean Ethics*. Ed. by WD Ross. Cambridge: Cambridge Univ. Press.

Beauchamp, Tom L and James F Childress (2013). *Principles of biomedical ethics*. 7th ed. New York: Oxford Univ. Press.

Bhatia, Sunil and Kumar Ravi Priya (2018). "Decolonizing culture: Euro-American psychology and the shaping of neoliberal selves in India". In: *Theory & Psychology* 28.5, pp. 645–668. doi: 10.1177/0959354318791315.

Bietti, Elettra (2020). "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, pp. 210–219. doi: 10.1145/3351095.3372860.

— (2023). "A Genealogy of Digital Platform Regulation". In: *7 Geo. L. Tech. Rev. 1*, pp. 1–68. Birch, Kean and D. T. Cochrane (Jan. 2022). "Big Tech: Four Emerging Forms of Digital Rentiership". In: *null* 31.1, pp. 44–58. doi: 10.1080/09505431.2021.1932794.

Bryson, Joanna J and Andreas Theodorou (2019). "How society can maintain human-centric artificial intelligence". In: *Human-centered digitalization and services*. Ed. by Marja Toivonen and Eveliina Saari. Singapore: Springer, pp. 305–323. doi: /10.1007/978-981-13-7725-9_16.

Bynum, Terrell (2018). "Computer and Information Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. url: https://plato.stanford.edu/archives/sum2018/entries/ethics-computer/%3E (visited on 03/20/2023).

Campbell, Heather (2006). "Just planning: The art of situated ethical judgment". In: *Journal of Planning Education and Research* 26.1, pp. 92–106. doi: /10.1177/0739456X06288090.
Celikates, Robin (2016). "Democratizing civil disobedience". In: *Philosophy & Social Criticism* 42.10, pp. 982–994. doi: /10.1177/0191453716638562.

Cenci, Alessandra and Dylan Cawthorne (2020). "Refining Value Sensitive Design: A (Capability Based) Procedural Ethics Approach to Technological Design for Well-Being". In: *Science and Engineering Ethics* 26.5, pp. 2629–2662. doi: 10.1007/s11948-020-00223-3.

Christman, John (2020). "Autonomy in Moral and Political Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University. url: https : / / plato . stanford . edu / archives / fall2020 / entries / autonomy - moral/ (visited on 03/19/2023).

Creel, Kathleen and Deborah Hellman (2022). “The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems”. In: *Canadian Journal of Philosophy* 52.1, pp. 26–43. doi: 10.1017/can.2022.3.

Dastin, Jeffrey (2018). “Amazon scraps secret AI recruiting tool that showed bias against women”. In: *Reuters*. url: https://www.reuters.com/article/us-amazon-com-jobs-automation insight-idUSKCN1MK08G (visited on 04/05/2022).

Davidson, Elsa (2011). *The burdens of aspiration: Schools, youth, and success in the divided social worlds of Silicon Valley*. New York and London: NYU Press.

Davis, Michael (1996). “Some Paradoxes of Whistleblowing”. In: *Business and Professional Ethics Journal* 15.1, pp. 3–19. doi: 10.5840/bpej19961517.

Dignum, Virginia (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Cham: Springer. doi: 10.1007/978-3-030-30371-6_1.

Dowling, Melissa-Ellen (2022). “Cyber information operations: Cambridge Analytica’s challenge to democratic legitimacy”. In: *Journal of Cyber Policy* 7.2, pp. 230–248. doi: 10.1080/23738871. 2022.2081089.

European Commission (2014). *Digital agenda for Europe: rebooting Europe’s economy*. Publications Office. doi: doi/10.2775/41229.

— (2022). “Regulatory framework proposal on artificial intelligence”. In: url: https://digital strategy.ec.europa.eu/en/policies/regulatory-framework-ai (visited on 03/19/2023). Farthing, Rys and Dhakshayini Sooriyakumaran (2021). “Why the Era of Big Tech Self-Regulation Must End”. In: *AQ: Australian Quarterly* 92.4, pp. 3–10.

Fiedler, Sergio (2009). “The right to rebel: Social movements and civil disobedience”. In: *Cos mopolitan Civil Societies: An Interdisciplinary Journal* 1.2, pp. 42–51. doi: 10.3316/informit. 144528716085320.

Floridi, Luciano (2021). “The End of an Era: from Self-Regulation to Hard Law for the Digital Industry”. In: *Philosophy & Technology* 34.4, pp. 619–622. doi: 10.1007/s13347-021-00493-0.

Floridi, Luciano et al. (2018). “AI4People–An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations”. In: *Minds and Machines* 28.4, pp. 689–707. doi: 10.1007/s11023-018-9482-5.

Fox, Sarah et al. (2017). “Social Justice and Design: Power and oppression in collaborative systems”. In: *Companion of the 2017 acm conference on computer supported cooperative work and social computing*, pp. 117–122.

Fox-Skelly, J et al. (2020). *The ethics of artificial intelligence: issues and initiatives*. European Parliament. doi: doi/10.2861/6644.

Friedman, Batya, Maaike Harbers, et al. (2021). “Eight grand challenges for value sensitive design from the 2016 Lorentz workshop”. In: *Ethics and Information Technology* 23, pp. 5–16. doi: /10.1007/s10676-021-09586-y.

Friedman, Batya, Peter Kahn, and Alan Borning (2002). “Value sensitive design: Theory and methods”. In: *University of Washington technical report* 2-12.

Friedman, Batya and Helen Nissenbaum (1996). “Bias in Computer Systems”. In: *ACM Trans. Inf. Syst.* 14.3, pp. 330–347. doi: 10.1145/230538.230561.

Gogoll, Jan et al. (2021). “Ethics in the Software Development Process: from Codes of Conduct to Ethical Deliberation”. In: *Philosophy & Technology* 34.4, pp. 1085–1108. doi: 10.1007/s13347-021-00451-w.

Greene, Daniel, Anna Hoffmann, and Luke Stark (2019). “Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning”. In: *Pro ceedings of the 52nd Hawaii International Conference on System Sciences*, pp. 2122–2131. doi: 10.24251/HICSS.2019.258.

Hagendorff, Thilo (2020). “The Ethics of AI Ethics: An Evaluation of Guidelines”. In: *Minds and Machines* 30.1, pp. 99–120. doi: 10.1007/s11023-020-09517-8.

Hallensleben, Sebastian et al. (2020). *From Principles to Practice. An interdisciplinary framework to operationalise AI ethics*. Tech. rep. AI Ethics Impact Group. url: https://www.ai-ethics impact.org/en (visited on 02/14/2021).

High Level Expert Group (2019a). “A definition of AI: Main capabilities and disciplines”. In: *European Commission*. url: https://digital-strategy.ec.europa.eu/en/library/definition artificial-intelligence-main-capabilities-and-scientific-disciplines (visited on 02/24/2023).

— (2019b). “Ethics Guidelines for Trustworthy AI”. In: *European Commission*. url: https:// digital-strategy.ec.europa.eu/en/library/ethics- guidelines- trustworthy- ai (visited on 10/29/2021).

Hildebrandt, Mireille (2015). *Smart technologies and the end (s) of law: novel entanglements of law and technology*. Cheltenham and Northampton MA: Edward Elgar Publishing.

Hildebrandt, Mireille (2020). *Law for Computer Scientists and Other Folk*. Oxford: Oxford Univ. Press. doi: 10.1093/oso/9780198860877.001.0001.

Hu, Margaret (2020). “Cambridge Analytica’s black box”. In: *Big Data & Society* 7.2. doi: /10.1177/2053951720938091.

Hubig, Christoph (2007). *Die Kunst des Möglichen II. Grundlinien einer dialektischen Philosophie der Technik Band 2: Ethik der Technik als provisorische Moral*. Bielefeld: transcript. doi: 10.14361/9783839405314.

Hubig, Christoph and Johannes Reidel (2003). *Ethische Ingenieurverantwortung. Handlungsspielräume und Perspektiven der Kodifizierung*. Berlin: zigma.

Hunt, Lucy and Maria Angela Ferrario (2022). “A review of how whistleblowing is studied in software engineering, and the implications for research and practice”. In: *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society*, pp. 12–23. doi: 10.1145/3510458.3513013.

Irani, Lilly et al. (2010). “Postcolonial computing: a lens on design and development”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1311–1320. Jacobs, Mattis et al. (2021). “Value Sensitive Design and power in socio-technical ecosystems”. In: *Internet Policy Review* 10.3, pp. 1–26. doi: 10.14763/2021.3.1580.

JafariNaimi, Nassim, Lisa Nathan, and Ian Hargraves (2015). “Values as Hypotheses: Design, Inquiry, and the Service of Values”. In: *Design Issues* 31.4, pp. 91–104.

Jobin, Anna, Marcello Ienca, and Effy Vayena (2019). "Artificial Intelligence: the global landscape of ethics guidelines". In: *Nat. Mach. Intell.*, pp. 389–399. doi: 10.1038/s42256-019-0088-2. arXiv: 1906.11668 [cs.CY].

Klonick, Kate (2020). "The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression". In: *Yale Law Journal* 129.2418.

Koulu, Riikka (2020). "Human control over automation: EU policy and AI ethics". In: *Eur. J. Legal Stud.* 12, p. 9.

Lepri, Bruno et al. (2018). "Fair, Transparent, and Accountable Algorithmic Decision-making Pro cesses". In: *Philosophy & Technology* 31.4, pp. 611–627. doi: 10.1007/s13347-017-0279-x.

Ma, Winston (2021). "Breaking the Big Tech Monopoly: The Coming Decade of Big Tech Regulations". In: *Horizons: Journal of International Relations and Sustainable Development* 18, pp. 166–179.

Maas, Jonne (2022). "Machine learning and power relations". In: *AI & SOCIETY*, pp. 1–8. doi: /10.1007/s00146-022-01400-7.

Mainsah, Henry and Andrew Morrison (2014). "Participatory design through a cultural lens: insights from postcolonial theory". In: *Proceedings of the 13th Participatory Design Conference: Short Papers, Industry Cases, Workshop Descriptions, Doctoral Consortium papers, and Keynote abstracts-Volume 2*, pp. 83–86.

Manders-Huits, Noëmi (2011). "What Values in Design? The Challenge of Incorporating Moral Values into Design". In: *Science and Engineering Ethics* 17.2, pp. 271–287. doi: 10.1007/s11948-010-9198-2.

Mateescu, Alexandra and Madeleine Elish (2019). *AI in context: the labor of integrating new technologies*. Tech. rep. url: https://apo.org.au/node/217456 (visited on 03/30/2023).
McNamara, Andrew, Justin Smith, and Emerson Murphy-Hill (2018). "Does ACM's code of ethics change ethical decision making in software development?" In: *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pp. 729–733.

Medzini, Rotem and David Levi-Faur (2023). "Self-Governance via Intermediaries: Credibility in Three Different Modes of Governance". In: *Journal of Comparative Policy Analysis: Research and Practice*, pp. 1–23. doi: 10.1080/13876988.2022.2155516.

Metcalf, Jacob, Emanuel Moss, and danah boyd (2019). "Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics". In: *Social Research: An International Quarterly* 86.2, pp. 449–476. doi: 10.1353/sor.2019.0022.

Miguel Beriain, I et al. (2022). *Auditing the quality of datasets used in algorithmic decision-making systems*. Publications Office of the European Union. doi: doi/10.2861/98930.

Mittelstadt, Brent Daniel (2019). "Principles alone cannot guarantee ethical AI". In: *Nature Machine Intelligence*, pp. 1–7. doi: 10.1038/s42256-019-0114-4.

Moore, George E (1996). *Principia ethica*. Ed. by Burkhard Wisser and Martin Sandhop. Erw. Ausg. Stuttgart: Reclam.

Moore, Martin and Damian Tambini (2018). *Digital dominance: the power of Google, Amazon, Facebook, and Apple*. Oxford: Oxford Univ. Press.

Moreno Gálvez, Francisco Javier and Francisco Sierra Caballero (2022). "Social appropriation of new technologies". In: *Internet Policy Review* 11.1, pp. 1–11.

Morley, Jessica et al. (2020). "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices". In: *Science and Engineering Ethics* 26, pp. 2141–2168. doi: 10.2139/ssrn.3830348.

Motupalli, Venkat (2017). "How Big Data is changing democracy". In: *Journal of International Affairs* 71.1, pp. 71–80.

Munn, Luke (2022). "The uselessness of AI ethics". In: *AI and Ethics*. doi: 10.1007/s43681-022-00209-w.

Muñoz, Cecilia, Megan Smith, and DJ Patil (2016). "Big data: A report on algorithmic sys tems, opportunity, and civil rights". In: *Executive Office of the President* 1. url: https : / / obamawhitehouse . archives . gov / sites / default / files / microsites / ostp / 2016 _ 0504_data_discrimination.pdf (visited on 03/08/2023).

Muschalik, Maximilian et al. (2022). "Agnostic Explanation of Model Change based on Feature Importance". In: *KI - Künstliche Intelligenz*. doi: 10.1007/s13218-022-00766-6.

Nersessian, David and Ruben Mancha (2020). "From automation to autonomy: legal and ethical responsibility gaps in artificial intelligence innovation". In: *Mich. Tech. L. Rev.* 27, p. 55.

Neuvonen, Riku and Esa Sirkkunen (2022). "Outsourced justice: The case of the Facebook Oversight Board". In: *Journal of Digital Media & Policy*. doi: 10.1386/jdmp_00108_1.

Nissenbaum, Helen (1994). "Computing and Accountability". In: *Commun. ACM* 37.1, pp. 72–80. doi: 10.1145/175222.175228.

Oudshoorn, Nelly and Trevor Pinch, eds. (2003). *How Users Matter. The Co-Construction of Users and Technology. The Co-Construction of Users and Technology*. Cambridge: The MIT Press. doi: 10.7551/mitpress/3592.001.0001.

Pasquale, Frank (2015). *The Black Box Society. The Secret Algorithms That Control Money and Information*. doi: 10.4159/harvard.9780674736061.

Peylo, Christoph et al. (2022). "VCIO based description of systems for AI trustworthiness characterisa tion VDE SPEC 90012 V1.0 (en)". In: *VDE Association for Electrical, Electronic and Information Technologies*. url: https://www.vde.com/en/working-areas/standards/spec/vde-spec publications (visited on 03/24/2023).

Podesta, John et al. (2014). "Big data: seizing opportunities, preserving values (Executive Office of the President)". In: *The White House, Washington, DC*. url: https://obamawhitehouse. archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf (visited on 03/08/2023).

Popescu, Ada-Iuliana et al. (2016). "In brief: Pros and Cons of corporate codes of conduct". In: *Journal of Public Administration, Finance and Law* 09, pp. 125–130.

Prem, Erich (2023). "From ethical AI frameworks to tools: a review of approaches". In: *AI and Ethics*, pp. 1–18. doi: 10.1007/s43681-023-00258-9.

Rahwan, Iyad (2018). "Society-in-the-loop: programming the algorithmic social contract". In: *Ethics*

and Information Technology 20.1, pp. 5–14. doi: 10.1007/s10676-017-9430-8.

Raji, Inioluwa Deborah et al. (2020). "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, January 27–30, 2020, Barcelona, Spain*, pp. 33–44. doi: 10.1145/3351095.3372873.

Rességuier, Anaïs and Rowena Rodrigues (2020). "AI ethics should not remain toothless! A call to bring back the teeth of ethics". In: *Big Data & Society* 7.2, pp. 1–5. doi: 10.1177/2053951720942541.

Rojas, Jose and Matthew Chalmers (2009). "The Appropriation of Information and Communication Technology: A Cross-Cultural Perspective". In: *Human-Computer Interaction. New Trends*. Ed. by Julie A. Jacko. Berlin, Heidelberg: Springer, pp. 687–696.

Rose, Nikolas (1996). "The death of the social? Re-figuring the territory of government". In: *Economy and Society* 25.3, pp. 327–356. doi: 10.1080/03085149600000018.

Sabl, Andrew (2001). "Looking forward to justice: Rawlsian civil disobedience and its non-Rawlsian lessons". In: *Journal of Political Philosophy* 9.3, pp. 331–349.

Santoni de Sio, Filippo and Jeroen van den Hoven (2018). "Meaningful Human Control over Autonomous Systems: A Philosophical Account". In: *Frontiers in Robotics and AI* 5. doi: 10.3389/frobt.2018.00015.

Santoni de Sio, Filippo and Giulio Mecacci (2021). "Four responsibility gaps with artificial intelligence: Why they matter and how to address them". In: *Philosophy & Technology* 34, pp. 1057–1084. doi: /10.1007/s13347-021-00450-x.

Schiller, Dan (1999). *Digital capitalism: Networking the global market system*. Cambridge: MIT press.

Schneider, Ingrid (July 2020). "Democratic Governance of Digital Platforms and Artificial Intelligence?: Exploring Governance Models of China, the US, the EU and Mexico". In: *JeDEM - eJournal of eDemocracy and Open Government* 12, pp. 1–24. doi: 10.29379/jedem.v12i1.604.

Schwartz, Mark S (2004). "Effective corporate codes of ethics: Perceptions of code users". In: *Journal of business ethics* 55, pp. 321–341. doi: 10.1007/s10551-004-2169-2.

Shaheen, Khadija et al. (2022). "Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks". In: *Journal of Intelligent & Robotic Systems* 105.1, p. 9. doi: 10.1007/s10846-022-01603-6.

Shea, Matthew (2020). "Forty Years of the Four Principles: Enduring Themes from Beauchamp and Childress". In: *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 45.4-5, pp. 387–395. doi: 10.1093/jmp/jhaa020.

Shneiderman, Ben (2020). "Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems". In: *ACM Trans. Interact. Intell. Syst.* 10.4. doi: 10.1145/3419764.

Simon, Judith (2017). "Value Sensitive Design and Responsible Research and Innovation". In: *The Ethics of Technology: Methods and Approaches*. Ed. by Sven Ove Hansson. London: Rowman & Littlefield, pp. 219–236.

Sollie, Paul (2007). "Ethics, technology development and uncertainty: an outline for any future ethics of technology". In: *Journal of Information, Communication and Ethics in Society* 5.4, pp. 293–306. doi:

10.1108/14779960710846155.

Suchman, Lucy A. (2007). *Human-machine reconfigurations. Plans and situated actions*. 2. ed. Cambridge [u.a.]: Cambridge Univ. Press.

Thompson, Dennis F (2017). "Designing responsibility: the problem of many hands in complex organizations". In: *Designing in ethics*. Ed. by Jeroen van den Hoven, Seumas Miller, and Thomas Pogge. Cambridge: Cambridge Univ. Press, pp. 32–56.

Tworek, Heidi (2019). "Social Media Councils". In: *Models for Platform Governance*. Ed. by Taylor Owen et al., pp. 97–102. url: https://www.cigionline.org/articles/social-media councils/#article-body (visited on 09/16/2023).

van de Poel, Ibo (2013). "Translating Values into Design Requirements". In: *Philosophy and Engi neering: Reflections on Practice, Principles and Process*. Ed. by Diane P Michelfelder, Natasha McCarthy, and David E. Goldberg. Dordrecht: Springer, pp. 253–266. doi: 10.1007/978-94-007-7762-0_20.

— (2016). "An Ethical Framework for Evaluating Experimental Technology". In: *Science and Engi neering Ethics* 22.3, pp. 667–686. doi: 10.1007/s11948-015-9724-3.
— (2020). "Embedding Values in Artificial Intelligence (AI) Systems". In: *Minds and Machines* 30.3, pp. 385–409. doi: 10.1007/s11023-020-09537-4.

van der Hoven, Jeroen and Noemi Manders-Huits (2020). "Value-sensitive design". In: *The Ethics of Information Technologies*. Ed. by Jan Kyrre Berg Olsen, Stig Andur Pedersen, and Vincent F. Hendricks. Routledge, pp. 329–332.

van Norren, Dorine Eva (2023). "The ethics of artificial intelligence, UNESCO and the African Ubuntu perspective". In: *Journal of Information, Communication and Ethics in Society* 21.1, pp. 112–128. doi: 10.1108/JICES-04-2022-0037.

Verbeek, Peter-Paul (2005). *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Trans. from the Dutch by Robert P. Crease. Pennsylvania State University Press. — (2006). "Materializing Morality: Design Ethics and Technological Mediation". In: *Science, Tech nology, & Human Values* 31.3, pp. 361–380. doi: 10.1177/0162243905285847.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2021). "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI". In: *Computer Law & Security Review* 41, p. 105567. doi: https://doi.org/10.1016/j.clsr.2021.105567.

Wagner, Ben (2018). "Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics Shopping?" In: *Being Profiled: Cogitas Ergo Sum*. Ed. by Emre Bayamlioglu et al. Amsterdam: Amsterdam Univ. Press, pp. 84–89. doi: doi:10.1515/9789048550180-016.

Wallach, Wendell and Colin Allen (2008). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford Univ. Press.

Winkler, Till and Sarah Spiekermann (2021). "Twenty years of value sensitive design: a review of methodological practices in VSD projects". In: *Ethics and Information Technology* 23.1, pp. 17–21. doi: 10.1007/s10676-018-9476-2.

Wong, Pak-Hang and Judith Simon (2020). "Thinking about 'ethics' in the ethics of AI". In: *IDEES* 48. url: https://revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/ (visited on 02/24/2023).

Wynsberghe, Aimee van (2013). "Designing Robots for Care: Care Centered Value-Sensitive Design".

In: *Science and Engineering Ethics* 19.2, pp. 407–433. doi: 10.1007/s11948-011-9343-6.

Yeung, Karen, Andrew Howes, and Ganna Pogrebna (2020). "AI Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing". In: *The Oxford Handbook of Ethics of AI*. Oxford University Press. doi: 10.1093/oxfordhb/9780190067397.013.5.

Zuboff, Shoshana (2019). *The age of surveillance capitalism. The fight for a human future at the new frontier of power*. First edition. New York: PublicAffairs.,