Enter the metrics: critical theory and organizational operationalization of AI ethics

Joris Krijger¹

Received: 12 November 2020 / Accepted: 2 August 2021 / Published online: 6 September 2021
© The Author(s) 2021

Abstract

As artificial intelligence (AI) deployment is growing exponentially, questions have been raised whether the developed AI ethics discourse is apt to address the currently pressing questions in the field. Building on critical theory, this article aims to expand the scope of AI ethics by arguing that in addition to ethical principles and design, the organizational dimension (i.e. the background assumptions and values influencing design processes) plays a pivotal role in the operationalization of ethics in AI development and deployment contexts. Through the prism of critical theory, and the notions of underdetermination and technical code as developed by Feenberg in particular, the organizational dimension is related to two general challenges in operationalizing ethical principles in AI: (a) the challenge of ethical principles placing conflicting demands on an AI design that cannot be satisfied simultaneously, for which the term ‘inter-principle tension’ is coined, and (b) the challenge of translating an ethical principle to a technological form, constraint or demand, for which the term ‘intra-principle tension’ is coined. Rather than discussing principles, methods or metrics, the notion of technical code precipitates a discussion on the subsequent questions of value decisions, governance and procedural checks and balances. It is held that including and interrogating the organizational context in AI ethics approaches allows for a more in depth understanding of the current challenges concerning the formalization and implementation of ethical principles as well as of the ways in which these challenges could be met.

Keywords Artificial intelligence · Ethics · Critical theory · Feenberg · Philosophy · Organizational operationalization

1 Introduction

In the last few years, there has been a surge in artificial intelligence (AI) developments and accompanying AI ethical guidelines (e.g. Jobin et al. 2019). Although the importance of ethics to data science is widely recognized (e.g. Herschel and Miori 2017; Boddington 2017), questions have been raised whether the developed discourse on AI ethics is apt to address the currently pressing questions in the field of AI. Especially when it comes to implementing or operationalizing the existing ethical frameworks and methods in real-world contexts, significant challenges remain (e.g. Hagedorff 2020). As such, the general principle-based approach or principled approach, which seeks to form a code of ethics to inform organizations and data scientists on what ethical

principles and values should be taken into account to assure ethical development and ethical AI systems, has drawn increased criticism. Recognizing the constructive role of the principled approach in focusing the debate of AI ethics, its limitations have become increasingly apparent (see e.g. Mittelstadt 2019) and concerns are expressed that the approach at this point might even hinder ethical developments in the field of AI, as conflicting ideals and vague definitions can be barriers to the implementation of ethics and to meaningful accountability (Crawford et al. 2019).

This article sets out to answer the question of how the practical impact of AI ethics can be advanced by arguing that the organizational design context requires at least equal, if not more, consideration than the design principles and methods when it comes to understanding the practical embedding of ethics in AI development and deployment. More specifically, the article will build on critical theory, and Feenberg in particular, to argue that the organizational dimension, as the background of values and assumptions that shape the design process, plays a pivotal role in the

✉ Joris Krijger
krijger@esphil.eur.nl

¹ Erasmus School of Philosophy, Erasmus University
Rotterdam, Rotterdam, The Netherlands

operationalization of ethics in AI development and deployment contexts. For values and ethical principles to have a meaningful impact on the outcome of an AI system, it is held, a focus on the design context, the organizational systems and processes, and the interests that shape value decisions in the development process is indispensable. With its long tradition of reflection on the relation between human values and the structure of technology, critical theorists like Feenberg can make a valuable contribution to understanding and conceptualizing this dimension of background values and assumptions. In particular, it is held that the critical notions of underdetermination and technical code as defined by Feenberg (e.g. 1991; 2002) can be considered key concepts in expanding the scope of ethical inquiry in AI ethics. The article, therefore, aims to explore how the organizational dimension can be articulated through the work of critical theorists like Feenberg and how increased attention to these implicit contextual aspects of the AI development process can foster the meaningful operationalization of ethics in AI.

Section 2 will briefly discuss AI and algorithmic decision-making and will outline the general developments in the discourse of AI ethics along with its perceived shortcomings. Section 3 provides a brief introduction and background on critical theory, presenting some of its key claims regarding rationality and technology, after which the notions of ‘underdetermination’ and ‘technical code’, as they are developed in the work of Andrew Feenberg, will be discussed. Section 4 relates these notions to the central tensions and trade-offs in the operationalization of ethical principles in AI contexts and will indicate how the contextual and organizational dimension of operationalization plays a pivotal role in the impact of ethics on AI design and employment. Section 5 explores how this approach could contribute to the field of AI ethics and how it might be furthered as an additional way of understanding challenges in the field of AI ethics.

2 AI and AI ethics

AI, as it is understood today, can be traced back to the convergence of three historical trends (Miaillhe and Hodes 2017), (i) the availability of large quantities of data (Big Data), (ii) the possibility of processing this data with growing computer power at relatively constant costs (Moore’s law) and (iii) the achievement in computer science for algorithms to automatically sort out complex patterns out of very large data sets (machine learning). Although there is no generally agreed upon definition of AI, existing definitions are often based on some key functional characteristics of the software, focusing on the system’s interaction with its environment and the capacity to, in varying degrees of autonomy and by computational means, derive decision-relevant outputs such as predictions and recommendations

from large datasets of relevant past phenomena from the real world. In the broadest sense, this can include, as the European Commission stressed in her regulation proposal for AI (2021), software developed with machine learning approaches, logic- and knowledge-based approaches and statistical approaches such as Bayesian estimation. A definition based on functional characteristics avoids digressing on future AI developments such as general AI and brings into focus presently existing and ‘near-term artificial intelligence’ (Gunn and O’Neil 2019) that is already in use today.

Given these characteristics of AI, both the appeal and perils of these systems can be addressed. The possibility to train models on vast amounts of labeled and unlabeled data can reduce inefficiencies, improve human decision-making, and optimize organizational processes (Eitel-Porter 2021). Not surprisingly the range of AI applications is vast, going from relatively simple analytics tools to high stake decision-making systems informing or sometimes executing decisions in sensitive domains such as loan applications, child welfare, bail and parole decisions, education, police deployment and immigration (e.g. Whittaker et al. 2018). On the other hand, however, given their widespread use, the scale and depth in which these algorithms can impact individual lives seems unprecedented, making the ethical dimension of these applications immediately salient. The importance of ethics is further stipulated by notable incidents such as the US recidivism prediction algorithm that allegedly mislabeled African-American defendants as “high-risk” at nearly twice the rate as it mislabeled white defendants (see Angwin et al. 2016; or Chouldechova 2016 for discussion), hiring algorithms that, based on analyzing previous hiring decisions, penalized applicants from women’s colleges for technology related positions (Dastin 2018) or a healthcare entry selection program that exhibited racial bias against African-American patients (Obermeyer et al. 2019).

In a first attempt to address these ethical risks of AI “seemingly every organization with a connection to technology policy has authored or endorsed a set of principles for AI” (Fjeld et al. 2020, p. 4). In 2019, as Jobin et al. (2019) found, already over 84 of these frameworks for ethical AI had been published, with the majority published after 2016. This principle-based approach to the ethics of AI, focusing on ethical codes of ethical principles and values that should be taken into account in AI development, determined much of the scope and agenda of governmental policies, academic research on AI and Ethics and of regulatory bodies. In general, these frameworks consist of several key themes, exemplified by the HLEG Guidelines for trustworthy AI that lay out seven key requirements: (1) human agency, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, and (7) accountability. Although these frameworks and principles

have proven constructive in focussing the discussion and in imposing a shared ethical frame, their proliferation posed a new challenge, as Crawford et al. (2019) remark: “there are now so many ethics policy statements that it became necessary to aggregate them into standalone AI ethics surveys, which attempted to summarize and consolidate a representative sample of AI principle statements to identify themes and make normative assertions about the state of AI ethics” (p. 20). On a meta-analysis level the discourse became partly self-referential by highlighting commonalities and convergence between ethical frameworks for AI (e.g. Floridi et al. 2018).

The implementation and operationalization of these principles, or rather the lack of operationalization opportunities, fueled criticism of the principle-based approach as dominant discourse in the field. From a practical perspective, the principle-based approach seemed toothless, devoid of normative ends and it was remarked that many of the guidelines tilt towards determinism, aligning well with business interests (Greene et al. 2019). Furthermore Mittelstadt (2019) argues that we cannot expect the high-level ethical principles for AI development and deployment to enjoy the same success the principled approach has had in medical ethics as some elemental foundations are lacking (e.g. common aims, fiduciary duties and proven methods to translate principles into practice). Powles (2018) points out that, in the discussions on AI and ethics, there are many diversions (e.g. the existential threat of AI for the human race or the hypothetical possibility of general artificial intelligence) that deflect us from more urgent questions of power in AI. Or, as she states it, we seem to overlook what the state of AI and power is today, how it impacts the questions we ask, and how we might think about it in relation to the questions that we *should* be asking. Also from a more empirical stance, the approach has been criticized: a study done by McNamara et al. (2020) showed *near to zero* effect of ethical principles on designer decisions compared to a control group that didn’t read the principles before designing. The focus on principles could even interfere with meaningful change in the industry as for example Wagner (2018) notes that “much of the debate about ethics seems increasingly focused on private companies avoiding regulation where ethics gives meaning to self-regulatory initiatives or even devolves to ‘pre-empting and preventing legislation’” (p. 3). So despite the degree of coherence and overlap between the existing sets of principles, without an understanding of the meaningful operationalization and implementation of these principles in contexts of practice, it seems improbable that the principle-based approach as it evolved, is apt to guide ethical development when it comes to the pressing questions of operationalizing ethics in near term AI. A similar challenge can be envisaged for more established principled approaches such as the

human right principles that, according to some scholars, should have a more central role in AI strategies (e.g. Hidvegi and Leufer 2019; van Veen and Cath 2018). Even though human rights have a more institutionalized basis and as such have more substance, they remain principle based with similarly limited specifications when it comes to implementation in data science practices.

Additionally, although promising, the more design-oriented approaches to meaningfully operationalize ethics in AI currently seem to equally fall short when it comes to implementing ethical constraints in real life AI contexts. Methods, such as value sensitive design (Friedman 1999; Friedman et al. 2002), designing in ethics (Van den Hoven et al. 2017), ‘ethics in/for/by design’ (Dignum 2018) and the guidance ethics as ‘ethical accompaniment of technology’ (Verbeek 2010), so far still have to find their way to structural embedment outside of the academic sphere. These methods, part of the ‘design turn’ in the philosophy of technology, seek to proactively influence the design of technologies to account for important human values during the conception and design process, where the ethicist “brings in perspectives that might otherwise remain under-represented” (Verbeek 2013, p. 80). They have proven valuable in focussing the ethical discussion on contextualized applications but fall short when it comes to moral guidance on value conflicts and seemed insufficient for ethical system development in real world organizational contexts (e.g. Manders-Huits and Zimmer 2009; Bianchin and Heylighen 2018).

Based on the above, one could argue that for ethical principles to be meaningfully implemented in real-world contexts, understanding of ethical principles and design methods are in and of themselves insufficient for meaningful implementation of ethics in AI. This article, therefore, holds that the implementation of ethics in AI could benefit from a more in depth understanding of the organizational dimension of operationalization as a third element in the scope of AI ethics. The following sections will further develop this idea that to advance ethical principles and values into actual AI design contexts, along with the design and design principles, special emphasis should be given to the organizational *context*, as the organizational dimension of operationalization. Ethical decisions regarding model development and deployment are ultimately made within contexts of organizations that have to align the ethical principles with vested interests such as the organizational culture, mission and goals. Critical theory, it is argued, provides a valuable prism for the conceptualization of these organizational aspects of operationalization. In particular the notions of ‘underdetermination’ and ‘technical code’, as they have been developed by Feenberg, can help to bring the contextual background of assumptions, values, definitions and roles guiding design decisions to the fore. The following section will introduce Critical theory and Feenbergs notions of ‘underdetermination’ and ‘technical

code’ and outlines their merit in articulating this organizational dimension.

3 Critical theory

There is good reason to relate insights from critical theorists to the current discussion on AI and ethics as many of the technology related concerns stressed by critical theorists, such as the dominating and controlling role of technology, seem to prefigure current discussions on Big Data and AI [e.g. on ‘surveillance capitalism’ as introduced by Zuboff (2019)]. Moreover, as will be outlined below, their conception of the more implicit and socio-economic dimensions of technology as articulated in the work of authors that are generally related to the field of critical theory, can advance an understanding of the organizational dimension of AI ethics that can complement existing narratives in the field.

3.1 Critical theory

Critical theory originates in 1929–1930 from the social theorists in the Western European Marxist tradition known as the Frankfurt School. From the onset, it focuses on the pathologies in the social world of ‘late-capitalist’ modernity that emerge due to the growing domination of an economical form of reasoning propagated by a.o. technology. The movement examines the social preconditions that constitute our lived experience via critical inquiry, not just to articulate these structures but also to transform them, believing in a strict interconnection between critical understanding and transformative action (Corradetti 2018). An extensive part of these structures comprises the scientific and technological systems that have been developed and critical theorists call into question the “effects of technological and scientific progress inasmuch as such progress expands and enhances the various forms of functionalist reason (i.e. reason that aims for technical mastery)” (Celikates and Jaeggi 2018, p. 257). Rather than celebrating the increasingly rapid development of technology, the Frankfurt School and later critical theorists started from the counterintuitive idea that the human potential for self-determination and rational organization seems to diminish rather than increase as our scientific-technological progress advances. They locate this paradoxical finding in our use of reason and rationality.

Although reason and rationality, as traditionally recognized by German idealism, was conceived as the primary source of human emancipation and progress, it also has another significance that became gradually more dominant. Simply put this concerns rationalization as the ongoing expansion of calculating attitudes aiming for efficiency in all spheres of life. The Frankfurt School drew on Weber’s thesis that rationalization resulted in the differentiation

and autonomization of previously unified value spheres, where all value spheres are pervaded by the logic of instrumental or purposive rationality. This “constricts the range of values contained within each life-order (as ultimate values are reduced to mundane, materialistic means and ends), and leads in turn to the increasing sameness of modern culture” (Gane 2002, p. 43). The leading members of the Frankfurt School, Adorno and Horkheimer, argue in their *Dialectic of Enlightenment* (1944) that this instrumentality is in itself a form of domination. They warn that the socio-economic context of rationalization in modern societies demands the atomization and commodifying standardization of labor. As a result it will be “impossible for subjects to experience individuality or view themselves as agents”, resulting in “an alienated and objectifying relationship to self and world [...]” (Celikates and Jaeggi, p. 258). This relationship manifests itself also on a technological level as Horkheimer indicates in *Traditional and Critical Theory*, stating that: “[t]he proposition that tools are prolongations of human organs can [now] be inverted to state that the organs are also prolongations of the tools” (1972, p. 201).

It was Herbert Marcuse who became one of the most prominent members of the Frankfurt School to establish “a link between the Frankfurt School’s general social critique of rationality and a concrete analysis of the technology structured way of life of advanced societies” (Feenberg 1995, p. 32). Marcuse shares the concerns for the way in which the measureless expansion of instrumental reason threatens society and relates this threat to the technically rational means involved in the rationalization process. In fact, for Marcuse the sphere of technological rationality could not be separated from political rationality anymore, or as he states it: “in the medium of technology, culture, politics, and the economy merge into an omnipresent system [...]” (Marcuse 1964, p. xv–xvi). In line with the teachings of Heidegger on technology, Marcuse ascribes an inherently dominating tendency to technology that resides under the appearance of neutrality and instrumentality. This purpose of domination, to Marcuse, is “‘substantive’ and to this extent belongs to the very form of technical reason” (Marcuse 1964, p. 25). Despite the determinism, social change maintains a possibility for Marcuse. In his later work he stresses that “technology is always a historical-social project: in it is projected what a society and its ruling interest intend to do with means and things. The machine is not neutral; technical reason is the social reason ruling a given society and can be changed in its very structure” (Marcuse 1968, p. 224–225). For Marcuse, [normative] principles are insufficient by themselves to determine the contours of a specific technical form of life (Feenberg 1996).

3.2 Feenberg: underdetermination and technical code

Throughout the work of Andrew Feenberg, it seems to be the possibility for social change and democratization of technology discerned in Marcuse that propels his exploration of how, within the critical theory tradition, human values can be incorporated in the very structure of technology. Feenberg subscribes to the idea that modern societies are dominated by ever more powerful organizations who are legitimated by their technical effectiveness. Technology, according to Feenberg, should be considered a form of power that is skewed towards its own expansion and perpetuation. As Feenberg states, “where [...] society is organized around technology, technological power is the principle form of power in the society” (2005, p. 49). However, he does attempt to formulate a possible subversive rationalization adapted to a more humane and democratic society (e.g. Feenberg 1991, 2010). To do this he focuses on the specific social groups that gain control of society through their leading role in technical organization as well as by focusing on specific technological design contexts. Both of these aspects come together in the concept of ‘underdetermination’ as it was proposed by Honneth (1991). Honneth (1991) proposed the term ‘underdetermined’ to describe the fact that “technical rules incompletely prescribe the respective form of their transposition into concrete actions” (p. 254). Feenberg (1995), following up on this, notes that the reason that a certain design or application is selected, developed and successful “has more to do with social values and interests than with the intrinsic technical superiority of the final choice” (p. 35). Technology is “an ambivalent process of development suspended between different possibilities [...] distinguished from neutrality by the role it attributes to social values in the design, not merely the use of technical systems” (Feenberg 1991, p. 14). Technologies as underdetermined, leave “room for social choice between different designs that have overlapping functions but better serve one or another social interest” (Feenberg 2017, p. 46).

This role of social interests in the design processes of technologies, as constituted within a web of social, political, and economic relations, signifies the importance of contextuality and the organizational dimension in the operationalization of AI ethics. As Feenberg points out, it is in the site of operationalization where normative principles take on a meaningful form. Power is realized precisely in the context where designs are formalized “through designs which narrow the range of interests and concerns that can be represented by the normal functioning of the technology and the institutions which depend on it” (Feenberg 2005, p. 49). However, in these contexts, interests are never equally balanced. Feenberg (1995) proposes the term ‘technical code’ to describe “those features of technologies that reflect

the hegemonic values and beliefs that prevail in the design process” (p. 4). As such the apparent neutral technological rationality “is enlisted in support of a hegemony through the bias it acquires in the process of technical development” (1995, 87). It is against a background of values and assumptions that certain choices appear as rational in the technical decision-making process (Hamilton and Feenberg 2005) and this background can be referred to as the technical code. What normative terms come to mean in technical specification, therefore, depends on how the struggle over this code unfolds (Wolff 2019).

Relating the notions of ‘technical code’ and ‘underdetermination’ to the operationalizing of ethical principles in specific AI contexts one could say that they highlight an important yet understudied aspect of implementation: they articulate the dimension where the social and technical interact within the confines of the developing or deploying organization. As such it outlines the socio-economic context within which value decisions for specific designs are made, guiding and encouraging efficient design while simultaneously determining what actually counts as efficient, what we can expect from technologies and what metrics we use to evaluate these systems. Inevitably, in a system that favors the optimization of efficiency, defined as cost-reduction or profit maximization, value challenges and conflicts arise that can render the impact of imposed normative principles on the final design negligible. It could be argued that without insight into this dimension of development and without strategies to meaningfully address and alter the system’s structures and processes, the organizational implementation of ethics in AI will fall flat regardless of the attention for values or principles in either the policy making or design process. Advancing this line of thought, the following section will expand on the importance of this organizational dimension in the implementation of ethics in AI. It will discuss two general challenges in the formalization of ethical principles in data science applications to demonstrate how in both challenges the notions of underdetermination and technical code provide additional insight in the value conflict at hand and its call for a resolution on an organizational level.

4 Critical theory and AI ethics

Conflicting values or conflicting interpretations of values are far from a new problem in the field of applied ethics (e.g. Berlin 1969; Bianchin and Heylighen 2018). Outlining the central tensions in the operationalization of ethical principles in AI development, however, provides an opportunity to relate the general tensions of operationalizing ethics in AI to the discussion of the ‘technical code’ and the role of the organizational and socio-economic context. From a practical perspective, a crude distinction can be made between

two general forms of value conflicts or ethical tensions in the implementation of ethical values in AI: *inter-principle tension* (the challenge of implementing multiple values and principles in one design) and *intra-principle tension* (the challenge of translating a single normative principle (in)to a specific technological design). Inter-principle tension arises when, in a practical setting, ethical values place conflicting demands that can't be satisfied simultaneously [as discussed by e.g. Kleinberg et al. (2017)]. Social or financial services for example, may find themselves in a tension between the need to adhere to the data minimization principle and respect privacy, and their duty of care urging them to maximize data use for the best possible risk profiles and service performance. Additionally, intra-principle tension arises not *between* ethical demands on a design, but *within* the operationalization of a single value because there are mutually exclusive (technological) interpretations or because the notion is incommensurable with a technological form. This tension mainly applies to the normative principle of 'fairness', which has multiple quantitative definitions [discussed by a.o. Hutchinson and Mitchell (2019); Verma and Rubin (2018)] but is also relevant for operationalizing explainability, privacy and accountability measures. Without discussing these trade-offs and tensions in great length, it is explored how these tensions are related to the technical code to provide a first illustration of how research on design context can advance AI ethics towards the level of operationalization.

4.1 Inter-principle tension

Value pluralism is a challenge that is often discussed in the form of ethical dilemmas and is a common theme in the fields of business ethics (e.g. Lurie and Albin 2007) and technology ethics (e.g. Wright et al. 2014). Attempting to optimize for all relevant values results in what Bianchin and Heylichen (2017) call the 'inclusive design paradox' where positively improving a system to include as many values as possible might negatively influence the overall application leading them to conclude that "taking human differences seriously seems to imply that nothing can be designed that meets the needs of everyone" (p. 4). When it comes to implementing ethical principles in a data science context, this inevitably has consequences for model development and deployment. Multi-objective trade-offs where two or more desiderata in machine learning systems compete with each other are often inevitable in the operationalization of ethics in AI. For example, keeping models explainable might, on a model level, come at the cost of predictive performance (see e.g. Sokol and Flach 2020). Similarly, studying the implications of implementing fairness constraints in relation to a.o. profit maximization, Hardt et al. (2016) outlined and quantified how imposing non-discrimination conditions in credit scoring and lending has significant implications

on profitability. The cost of fairness constraints in relation to the value of safety was studied by Corbet-Davies et al. (2017), who examined an algorithm for pretrial decisions. They found a tension between the unconstrained and optimal functioning of the algorithm and the constrained algorithm that satisfied prevailing notions of algorithmic fairness. By analyzing data from Broward County they concluded that "optimizing for public safety yields stark racial disparities; conversely, satisfying past fairness definitions means releasing more high-risk defendants, adversely affecting public safety" (p. 8). What studies like these demonstrate is that, in implementing ethical principles, value trade-offs emerge that can and, one could argue, should not be resolved on the level of design and development. Rather they call for contextual and organizational resolutions that require an awareness of cultural and organizational values and how these relate to the ethical principles. This emphasizes the role of policymakers, boards and managers in the ethical development process, as they are the ones ultimately tasked with resolving these organizational challenges. For example, from a cost perspective, in addition to the additional monetary and nonmonetary costs that ethical considerations impose on organizations (Kretz 2016), implementing the ethical constraints in algorithms and data science practices might impose new structural costs on the organization. As of yet, for many organizations it is not clear how to address, assess and decide these sort of value questions and who should make these decisions.

Given the pace and scale with which AI models work, implementing constraints and solutions that are sub-optimal from a profit or efficiency maximizing perspective exposes the organization to performance impairment. Although the importance of these ethical constraints is indisputable, given the technical code and the existing hegemony of interests, one can expect that the background of reductive efficiency oriented assumptions in most organizations, will complicate the full implementation of these constraints. Since organizations, public and private, strive towards profits or efficiency and cost reduction, it is apparent that the operationalization of ethical principles might conflict with organizational interests. The operationalization of ethical principles hinges on the commitment from organizations to value these ethical principles over their primary organizational interests. It could even be argued that, given the background of assumptions and values against which most AI applications are developed in organizations, it is unsurprising that ethical principles and value sensitive design approaches have yielded limited results. Although public interests increasingly move organizations towards the incorporation of public values in the organization strategy (e.g. Frankel 2008), optimization of performance in terms of efficiency, growth or volume is still the dominant imperative, with profitability being the most widely accepted benchmark of overall

performance (Daft et al. 2010). Operationalizing ethical principles in AI contexts requires not only a revision of these background values and assumptions in organizations but might also entail a change in how performance of organizations is evaluated.

However, not just the tension *between* principles can be understood through the prism of the technical code. In a similar fashion, the tension that can occur in translating or interpreting a single ethical principle for AI, the ‘intra-principle tension’, can be shaped by the technical code.

4.2 Intra-principle tension

The underdetermination of technologies and the further elaboration on the technical code imply that the final form of a technological design is not necessarily decided by technical functions but rather by social values and the fit with context. The background values and assumptions are relevant when principles conflict but also in the formalization or implementation of a single ethical principle as for example Binns (2017, b) has noted. He points to the need for the formalization of fairness asking what it means “for a machine learning model to be ‘fair’ or ‘non-discriminatory’, in terms which can be operationalized?” (p. 1). As Binns points out, in the field of fair machine learning various different formal definitions of fairness have been proposed, all showing statistical limitations and each incommensurable with the others, indicating that satisfying a single fairness definition might make it impossible to satisfy the other forms of fairness. One could for example opt for equal treatment between groups but that could mean that the overall predictive accuracy for each group is impaired (as there might be legitimate grounds for certain forms of discrimination) (see e.g. Dwork et al. 2012; Angwin et al. 2016). As such, Corbett-Davis and Goel (2018) hold that it is unlikely that a universally applicable strategy will be found and claim that formal mathematical measures of fairness can inadvertently lead discussions astray. Technical and policy discussions on the interpretation of ethical values, they argue, should be grounded in terms of real-world quantities. The operationalization of fairness in credit risk models should be assessed on the basis of the risk assessment’s immediate and equilibrium effect on community development and the sustainability of a loan program. Ultimately this contextual dimension is what characterizes the operationalization of this value. Binns (2017) emphasizes this point when he states that “it is clear that the field of fair machine learning faces ‘an upfront set of conceptual ethical challenges; which measures of fairness are most appropriate in a given context?’ (Binns 2017, b, p. 2).

The tension in the interpretation and formalization of a single ethical principle in AI, necessary because most principles remain ambiguously defined in the principle-based

approach, brings the discussion again to the level of the technical code and organizational decision-making. In both forms of tension, inter-principle and intra-principle, the impact and extent to which ethical principles are operationalized are influenced not by specific design values but by the background assumptions and values of the design context. Both forms of ethical tension in the operationalization of ethical principles in AI point towards the importance of the fit between the AI applications in general and the context in which they operate.

5 Discussion

As it appears, the current principle-based debate on the ethics of AI and value sensitive methods can play an important role in the awareness of ethical risks but seem insufficient when it comes to fostering the implementation of ethical practices in the domains where AI is actually developed and deployed. Exploring how critical theory, and Feenberg in particular, might contribute to a more praxis oriented ethics in the field of AI, the notions of underdetermination and technical code were discussed to articulate aspects of the organizational dimension of operationalizing ethical principles in AI. This dimension, as was outlined, is an elemental aspect of the adequate assessment and understanding of two general types of ethical tension encountered in the operationalization of ethical principles: inter-principle tension (regarding the trade-off between multiple values or principles in a specific design) and intra-principle tension (the complexity of formalizing a single normative principle due its manifold and mutually exclusive forms of operationalization). Assessing these tensions through the prism of the ‘technical code’ allows for a first conceptualization of this organizational dimension in an AI ethics context. More specifically, Feenberg’s notions of ‘underdetermination’ and ‘technical code’ bring the relation between algorithmic design and the contextual background of values and assumptions shaping design decisions into focus. Ethical AI, it has been argued, requires in addition to sound guiding ethical principles and design methods, an improved understanding of the organizational or contextual background that shapes design decisions and the perceived rationality of design options at hand. Building on critical theory, the inter- and intra-principle tensions can be regarded as upfront ethical challenges that underscore the relevance of the context of operationalization and the hegemony of interest at play in the social dimension of design.

There are some important theoretical and practical insights afforded by the notions of underdetermination and technical code as outlined in this article that go beyond the, perhaps more obvious, point that guidelines do not guarantee ethical AI and that attempts to reshape the agendas

these technologies serve will need to address the institutional contexts. For one, addressing and interrogating the organizational context and values as part of the AI ethics operationalization introduces a novel aspect of assessment and implementation that has not received the proper attention in the discourse on ethical approaches so far. Moreover, drawing attention to the organizational context in a way that allows for an engagement with the current challenges in AI ethics concerning the formalization and implementation of ethical principles offers opportunities to develop a more in depth understanding of these challenges and the ways in which they could be met. As these challenges ultimately require awareness, deliberation and strategic decision-making on a socio-economic, political and organizational level, approaches directed towards the meaningful involvement of internal stakeholders over and above the departments and teams tasked with AI development seems necessary. Another way in which these insights surmount the mere underscoring of the role of material interests of people and corporations in AI development, is that, through the prism of the technical code, interests are specified and fleshed out by locating them in the organizational setting of technology development and deployment. Rather than dealing with abstract forces or interests, the value conflicts as addressed in relation to the technical code allow for a contextualization and articulation of specific value trade-offs. Moreover, making this dimension explicit and addressing the value conflicts as socio-economically embedded within an organizational context in turn makes it possible to evaluate the moral underpinnings of specific resolutions or decisions in the value conflicts at hand. Most importantly, however, this conception of the role of ethics in AI development and of what guarantees the best outcome for ethical values in this process, shifts the focus from specific techniques or individual developers towards organizational structures, processes, and business models. It opens a space in the AI ethics discussion to address the relation between the organizational goals, culture and values and the processes of technology development and AI design. Without disregarding the progress that has been made on the (open-sourced) tooling available to assess fairness or explainability of models, an organizational approach brings in scope a critical examination of the contextualized value decisions vis-a-vis AI systems as they are developed within organizations. As AI ethics metrics and assessments might provide guidance on signaling divergence from ethical principles or moral shortcomings in the design of an AI application, the notion of technical code precipitates a discussion on the subsequent questions of governance and procedural checks and balances (e.g. what to do next, how to mitigate the ethical risks, how to organize decision making around ethical risks, who should decide in these matters and how can decisions be morally motivated). Rather than discussing specific use cases or AI applications, the notion of

technical code extends the AI ethics discussion towards the ways in which ethical safeguards are embedded in the systems and processes of an organization. As particular critical tools one could conceptualize the instigation of new forms of ethical risk assessment for AI to complement existing technology-oriented assessments. Rather than focusing on compliance with guidelines through checklists, these supplementary assessment methods could focus on relational aspects of ethical risks such as the distribution of risks and benefits between the developing organization and the people subjected to the application. More generally speaking these notions, as developed from a critical theory perspective, advance a more holistic approach to the AI systems and the ethical values they embody, broadening the existing discourse on AI ethics.

From a more theoretical perspective the ideas of under-determination and technical code can expand and advance the discussion in AI ethics on moral responsibility, organizational incentives and what constitutes value in a more general sense. In questions on the ascription of moral responsibility in AI contexts, for example, the organizational dimension could be a valuable contribution. An oft-mentioned challenge in this field is the responsibility gap, which denotes the widening gap between the behavior of autonomous self-learning systems and their developers and programmers, and the subsequent impossibility to locate the responsibility for bad moral behavior from these systems (see Matthias 2004). Here the inclusion of the organizational dimension and the interaction between socio-economic interests, technology design and background values and assumptions could provide a valuable new approach inciting one to go beyond a reductive individual form of responsibility. As the article provides a first exploration of this theme of organizational operationalization from a critical theory perspective, various other threads for future research remain. The focus on organizational operationalization might provide common ground for other applied ethics disciplines such as business ethics to extend their insights to the AI ethics debate. Since AI is applied in many diverse fields and settings it is not surprising that systematic studies on design contexts and values from an organizational perspective have yet to be fully developed. However, making these interests and their influence on design outcomes in relation to the implementation of ethical principles explicit would significantly advance the field of AI ethics towards practical implementation. As such AI ethics could learn from business ethics and other fields of applied ethics that have already developed an approach to studying the effectiveness of ethical codes (e.g. Kaptein and Schwartz 2008). Second, more research on the relevance of critical theory for the current AI and ethics debate could be fruitful to further develop our understanding of the social impact of AI and the cultural background that guides design processes. The notion of organizational operationalization

could propel an inquiry into the broader political discussion of normative ends and contextual appropriateness of certain forms of value operationalization. For example, the technological mediation between AI systems as technical artifacts and human or organizational values, where AI might co-shape our social understanding of common ethical principles (see e.g. Verbeek 2007), is an interesting post-phenomenological thread that could contribute to the furthering of the organizational dimension. Additionally, the emphasis on the embeddedness of AI technologies and AI ethics within a broader political, economic and social context could help spark a more political debate on how interests should be balanced and how contextual appropriateness can be judged. Here a bridge can be made to some monumental works on institutionalized social justice from a.o. Rawls G.A. Cohen and Sen. A focus on the institutional dimension of operationalizing AI ethics allows for a convergence with social justice approaches that have a strong institutional focus and consider society's economic and political institutions and institutional arrangements as objects of justice.

In an idiosyncratic way, the maturation process the field of AI ethics has to follow could parallel the development of critical theory itself as described by Anderson (2011). Just as critical theory, through the course of its three generations, has been evolving from universalistic principles of morality, justice and truth towards issues of particularity, contextuality, and substantive, non-proceduralistic principles, so too could AI ethics progress from high-level principles towards dimensions of application, contextual justification and judgements of appropriateness.

6 Concluding thoughts

In addition to a reflection on the current state of AI ethics, this article is meant as a linking pin for the critical theory discourse and the discourse on AI ethics. AI ethics, as it stands, has developed some necessary but insufficient ethics mechanisms to adequately respond to the risks of increased development and deployment of AI. It is the aim of this article to, through the prism of critical theory, expand the scope of AI ethics to include the organizational dimension of background assumptions and values influencing design processes. As a first exploration, much remains to be further developed, nuanced and at times corrected in regard to the challenges and arguments outlined above. However, it should be apparent that critical theory has an important contribution to make when it comes to operationalizing ethical principles and expanding the scope of the discussion on AI ethics towards organizational operationalization. Entering the metrics means confronting the social reality in which AI systems are developed and within which ethical principles have to be formalized. Amidst the rapidly growing

deployment of AI systems, the ambition of technical democracy, where design is consciously oriented towards politically legitimated human values, seems more important than ever.

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

References

- Anderson J (2011) Situating Axel Honneth in the Frankfurt school tradition. In: Petherbridge D (ed) Axel Honneth: critical essays: with a reply by Axel Honneth. Brill Academic Publishers, Leiden, pp 207–232
- Angwin J, Larson J, Mattu S, Kirchner L (2016) How we analyzed the compas recidivism algorithm. <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>. Accessed 13 Sept 2020
- Berlin I (1969) Four essays on liberty. Oxford University Press, Oxford
- Bianchin M, Heylighen A (2017) Fair by design Addressing the paradox of inclusive design approaches. *Des J* 20(sup1):S3162–S3170
- Bianchin M, Heylighen A (2018) Ethics in Design: pluralism and the case for justice in inclusive design. *Des Res Soc Conf*. <https://doi.org/10.21606/drs.2018.221>
- Binns R (2017) Fairness in machine learning: lessons from political philosophy. *Proc Mach Learn Res* 81:1–11
- Binns R (2017b) Algorithmic accountability and public reason. *Philos Technol* 31:543–556
- Boddington P (2017) Towards a code of ethics for artificial intelligence. Springer international publishing, Berlin
- Celikates R, Jaeggi R (2018) Technology and reification: technology and science as 'Ideology.' In: Brunkhorst H, Kreide R, Lafont C (eds) The Habermas handbook. Colombia University Press, New York
- Corbett-Davies S, Goel S (2018) The Measure and mismeasure of fairness: a critical review of fair machine learning. *ArXiv*. <https://arxiv.org/abs/1808.00023>.
- Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of KDD'17, Halifax, NS, Canada, August 13–17.
- Corradetti C (2018) The Frankfurt school and critical theory. Accessed Sept 2019: <https://www.iep.utm.edu/frankfur/>. Accessed 21 Sept 2019
- Crawford K, Dobbe R, Dryer T, Fried G, Green B, Kaziunas E, Kak A, Mathur V, McElroy E, Nill Sánchez A, Raji D, Lisi Rankin J, Richardson R, Schultz J, Myers West S, Whittaker M (2019) AI now 2019 report. New York: AI Now Institute, 2019. https://ainowinstitute.org/AI_Now_2019_Report.html. Accessed 14 May 2020
- Daft RL, Murphy J, Willmot H (2010) Organization theory and design. Cengage Learning EMEA, Boston

- Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>. Accessed 20 Sept 2020
- Dignum V (2018) Ethics in artificial intelligence: introduction to the special issue. *Ethics Inf Technol* 20:1–3. <https://doi.org/10.1007/s10676-018-9450-z>
- Eitel-Porter R (2021) Beyond the promise: implementing ethical AI. *AI Ethics* 1:73–80
- European Commission (2021) Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. EUR-Lex - 52021PC0206. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>. Accessed 23 Apr 2021
- Feenberg A (1991) Critical theory of technology. Oxford University Press, Oxford
- Feenberg A (1995) Alternative modernity: the technical turn in philosophy and social theory. University of California Press, Berkeley
- Feenberg A (1996) Marcuse or Habermas: two critiques of technology. *Inquiry* 39:45–70
- Feenberg A (2002) Transforming technology - a critical theory revisited. Oxford University Press
- Feenberg A (2005) Critical theory of technology: an overview. *Taylor Biotechnol* 1(1):47–64
- Feenberg A (2010) Between reason and experience: essays in technology and modernity. Massachusetts Institute of Technology, Cambridge
- Feenberg A (2017) Technosystem: the social life of reason. Harvard University Press, Harvard
- Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center for Internet & Society, Cambridge
- Floridi L, Cows J, Beltrametti M et al (2018) AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mind Mach* 28:689–707
- Frankel EG (2008) Organizational effectiveness and performance. Quality decision management—the heart of effective futures-oriented management. Topics in safety, risk, reliability and quality, vol 14. Springer, Dordrecht
- Friedman B (1999) Value-Sensitive Design: A Research Agenda for Information Technology. Contract No: SBR-9729633, National Science Foundation, Arlington, VA
- Friedman B, Kahn P, Borning A (2002) Value sensitive design: Theory and methods (Technical Report No. 2–12). University of Washington
- Gane N (2002) Max weber and postmodern theory: rationalization versus re-enchantment. Palgrave MacMillan, London
- Greene D, Hoffman AL, Stark L (2019) Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Hawaii international conference on system sciences pp 1–10
- Gunn H, O'Neil C (2019) Near term AI. *Ethics of Artificial Intelligence*. In press
- Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. *Mind Mach* 30:99–120
- Hamilton E, Feenberg A (2005) The Technical Codes of Online Education. *E-Learn Dig Media* 2(2):104–121. <https://doi.org/10.2304/elea.2005.2.2.1>
- Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in neural information processing systems pp 3315–3323.
- Herschel R, Miori V (2017) Ethics & Big Data. *Technol Soc* 49:31–36
- Hidvegi F, Leufer D (2019) Laying down the law on AI: ethic done, now the EU must focus on human rights. Accessed 8 Apr
- Honneth A (1991) The critique of power: reflective stages in a critical social theory. The MIT Press, Cambridge
- Horkheimer M (1972) Traditional Theory and Critical Theory, in *Critical Theory: Selected Essays* trans. Matthew J. O'Connell et al.
- Hutchinson B, Mitchell M (2019) 50 Years of Test (Un)fairness: Lessons for Machine Learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency
- Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. *Nat Mach Intell* 1:389–399
- Kaptein M, Schwartz MS (2008) The Effectiveness of business codes: a critical examination of existing studies and the development of an integrated research model. *J Bus Ethics* 77:111–127
- Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent Trade-Offs in the Fair Determination of Risk Scores. [arXiv:1609.05807](https://arxiv.org/abs/1609.05807)
- Kretz L (2016) Costs of Ethics. In: Farazmand A (eds) *Global Encyclopedia of Public Administration, Public Policy, and Governance*. Springer, Cham. https://doi.org/10.1007/978-3-319-31816-5_2374-1
- Lurie Y, Albin R (2007) Moral dilemmas in business ethics: from decision procedures to edifying perspectives. *J Bus Ethics* 71:195–207
- Manders-Huys N, Zimmer M (2009) Values and pragmatic action: the challenges of introducing ethical intelligence in technical design communities. *Int Rev Inf Ethics* 10(2):37–45
- Marcuse H (1964) One dimensional man. Routledge & Keagan Paul, Milton Park
- Marcuse H (1968) Negations—essays in critical theory. Beacon Press, Boston
- Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf Technol* 6:175–183
- McNamara A, Smith J, Murphy-Hill (2020) Does ACM's code of ethics change ethical decision making in software development? In: ESEC/FSE'18, November 4–9, 2018. <https://doi.org/10.1145/3236024.3264833>
- Miahlhe N, Hodes C (2017) The third age of artificial intelligence. *J Field Action* 17:6–11
- Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 1:501–507
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453
- Powles J (2018) AI and power, with Julia Powles. In public: 2018 Uehiro-Carnegie-Oxford Conference: Ethics and the Future of Artificial Intelligence. May 11, 2018
- Sokol K, Flach PA (2020) Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM FAT* 2020
- Van den Hoven J, Miller S, Pogge T (eds) (2017) Designing in ethics. Cambridge University Press, Cambridge
- Van Veen C, Cath C (2018) Artificial Intelligence: what's human rights got to do with it? Data and society. Accessed 12 Mar 2020
- Verbeek PPCC (2007) The technological mediation of morality—a post-phenomenological approach to moral subjectivity and moral objectivity
- Verbeek PP (2010) Accompanying technology: philosophy of technology after the ethical turn. In: Techné: Research in Philosophy and Technology vol 14 No 1, pp 49–54
- Verbeek PPCC (2013) Technology design as experimental ethics. In: van der Burg S, Swierstra T (eds) *Ethics on the laboratory floor*. Palgrave Macmillan, London
- Verma S, Rubin J (2018) Fairness definitions explained. In: 2018 IEEE/ACM international workshop on software fairness (FairWare). IEEE, New York
- Wagner B (2018) Ethics as an escape from regulation: from ethics-washing to ethics-shopping? In: Hildebrandt M (ed) *Being profiling*. Cogitas ergo sum. Amsterdam University Press, Amsterdam

- Whittaker M, Crawford K, Dobbe R, Fried G, Kaziunas E, Mathur V, Richardson R, Schultz J, Myers West S, Schwartz O (2018) AI Now Report 2018. https://ainowinstitute.org/AI_Now_2018_Report.pdf. Accessed 16 May 2020
- Wolff R (2019) Towards a critical theory of the technosystem. *Jus Cogens* 1:173–185
- Wright D, Finn R, Gellert R, Gurwirth S, Schütz P, Friedewald M, Venier S, Mordini E (2014) Ethical dilemma scenarios and emerging technologies. *Technol Forecast Soc Chang* 87:325–336
- Zuboff S (2019) *The age of surveillance capitalism*. Profile Publishers, London
- Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.