AI and Ethics
https://doi.org/10.1007/s43681-022-00209-w

ORIGINAL RESEARCH

The uselessness of AI ethics

Luke Munn[1]

Received: 6 June 2022 / Accepted: 2 August 2022

Abstract
As the awareness of AI's power and danger has risen, the dominant response has been a turn to ethical principles. A flood of AI guidelines and codes of ethics have been released in both the public and private sector in the last several years. However, these are *meaningless principles* which are contested or incoherent, making them difficult to apply; they are *isolated principles* situated in an industry and education system which largely ignores ethics; and they are *toothless principles* which lack consequences and adhere to corporate agendas. For these reasons, I argue that AI ethical principles are useless, failing to mitigate the racial, social, and environmental damages of AI technologies in any meaningful sense. The result is a gap between high-minded principles and technological practice. Even when this gap is acknowledged and principles seek to be "operationalized," the translation from complex social concepts to technical rulesets is non-trivial. In a zero-sum world, the dominant turn to AI principles is not just fruitless but a dangerous distraction, diverting immense financial and human resources away from potentially more effective activity. I conclude by highlighting alternative approaches to AI justice that go beyond ethical principles: thinking more broadly about systems of oppression and more narrowly about accuracy and auditing.

Keywords AI · Artificial intelligence · Ethics · Ethical principles · Morality · Social ills · Research

1 Introduction

Artificial intelligence technologies are increasingly being deployed in a range of sectors, from healthcare to human resources, education, agriculture, manufacturing, and law enforcement. However, as the pervasiveness of AI grows, so does its capacity to damage lives and livelihoods. Within welfare and social support systems, automated decision making systems can exacerbate inequality and punish the poor [18]. Racialized assumptions can be embedded in information infrastructures, perpetuating stereotypes and prejudice [55]. Data-driven models can be opaque and biased, making detrimental choices in high stakes areas and undermining democratic and egalitarian conditions [60]. And all of these technologies operate on people and spaces that are already economically and socially stratified [51], elevating the importance and the difficulty of operating in ways that contribute to human rights and dignity. The promises of AI have been tempered by its potential for harm [64].

As the awareness of AI's power and danger has risen, the dominant response has been a turn to AI ethics—ethics being understood here in the narrow but well-established sense as "a set of moral principles" according to both the OED and Merriam-Webster dictionaries. The public and private sectors have released guidelines, frameworks, and principles that are meant to apply when creating new AI technology. Over 50 of these have been issued by government agencies, including national frameworks produced by the UK, the USA, Japan, China, India, Mexico, Australia, and New Zealand, amongst others [69]. There are the Beijing AI Principles, DeepMind's Ethics, and Society Principles [15], IEEE's Ethically Aligned Design [34], the Guidelines for Artificial Intelligence by Deutsche Telekom [17], and the Vatican AI Principles known as the Rome Call for AI Ethics [65]. Indeed, the list of AI Principles at AI Ethicist now stretches to over 80 entries, with more being constantly added [2].

This article argues that this deluge of AI ethical principles is largely useless. While this view is provocative, it is hardly alone: a growing sea of voices have begun critiquing the de-facto turn to AI ethical principles as ineffective [39, 48, 69, 78]. In the first three sections, I lay out three causes for this failure: *meaningless principles*, *isolated principles*,

✉ Luke Munn
luke.munn@gmail.com

[1] Institute for Culture and Society, Western Sydney University, Sydney, NSW, Australia

Published online: 23 August 2022

and *toothless principles*. The result of this failure is a gulf between high-minded ideals and technological development on the ground—a gap between principles and practice. While recent work has aimed to address this gap by operationalizing principles [12, 49], this work is fraught in attempting to translate contested social concepts to technical rules and featuresets. The final section argues that, in a zero-sum world, the obsession with AI principles is not just useless but dangerous in funneling human and financial resources away from more productive approaches. The article thus concludes by highlighting alternatives: the first thinks more broadly about AI justice, considering sociopolitical dynamics and systems of oppression [14, 46]; the second thinks more narrowly, focusing on concrete issues like accuracy, auditing, and governance [25, 67].

2 Meaningless principles

The deluge of AI codes of ethics, frameworks, and guidelines in recent years has produced a corresponding raft of principles. Indeed, there are now regular meta-surveys which attempt to collate and summarize these principles [35]. However, these principles are highly abstract and ambiguous, becoming incoherent. Mittelstadt [45, p 501] suggests that work on AI ethics has largely produced "vague, high-level principles, and value statements which promise to be action-guiding, but in practice provide few specific recommendations and fail to address fundamental normative and political tensions embedded in key concepts." The point here is not to debate the merits of any one value over another, but to highlight the fundamental lack of consensus around key terms. Commendable values like "fairness" and "privacy" break down when subjected to scrutiny, leading to disparate visions and deeply incompatible goals.

What are some common AI principles? Despite the mushrooming of ethical statements, Floridi and Cowls [21] suggest many values recur frequently and can be condensed into five core principles: beneficence, non-maleficence, autonomy, justice, and explicability. These ideals sound wonderful. After all, who could be against beneficence? However, problems immediately arise when we start to define what beneficence means. In the Montreal principles [77, p 545] for instance, "well-being" is the term used, suggesting that AI development should promote the "well-being of all sentient creatures." While laudable, clearly there are tensions to consider here. We might think, for instance, of how information technologies support certain conceptions of human flourishing by enabling communication and business transactions—while simultaneously contributing to carbon emissions, environmental degradation, and the climate crisis [33, 41, 52]. In other words, AI promotes the well-being of some creatures (humans) while actively undermining the well-being of others.

The same issue occurs with the Statement on Artificial Intelligence, Robotics, and Autonomous Systems [19]. In this Statement, beneficence is gestured to through the concept of "sustainability," asserting that AI must promote the basic preconditions for life on the planet. Few would argue directly against such a commendable aim. However, there are clearly wildly divergent views on how this goal should be achieved. Proponents of neoliberal interventions (free trade, globalization, deregulation) would argue that these interventions contribute to economic prosperity and in that sense sustain life on the planet. In fact, even the oil and gas industry champions the use of AI under the auspices of promoting sustainability [16]. Sustainability, then, is a highly ambiguous or even intellectually empty term [3, 40] that is wrapped around disparate activities and ideologies. In a sense, sustainability can mean whatever you need it to mean. Indeed, even one of the members of the European group denounced the guidelines as "lukewarm" and "deliberately vague," stating they "glossed over difficult problems" like explainability with rhetoric [43].

If sustainability is ambiguous, so are many of the key terms used in AI ethics frameworks. Safety, well-being, autonomy, and justice are contested concepts and often shift in significant ways depending on the context. Privacy, for example, has long overflowed with competing and contradictory definitions, with scholarship noting the lack of clarity and accepted consensus around their term [5]. Even the most influential conceptions of privacy characterize it as a big tent, housing a diverse group of interests and a diverse array of meanings [75]. Many key concepts in AI frameworks, then, are overburdened, brimming with contradictory meanings. Floridi [20] has suggested that developers of AI may conduct ethics shopping, borrowing liberally from different frameworks to arrive at a set of easy-to-implement norms. However, the fuzziness of AI principles outlined above suggests that this cynical mix-and-match approach may not even be necessary. Instead, terms like "beneficence" and "justice" can simply be defined in ways that suit, conforming to product features and business goals that have already been decided. Such ambiguity facilitates ethical "box ticking," allowing a company to claim adherence to a set of principles or ideals without engaging in any meaningful degree of reflection or reconfiguration.

3 Isolated principles

AI development does not take place in a vacuum. The development and adoption of technology is always highly social and cultural [27], embedded within a rich network of human and non-human actors [38]. This means that technology is

influenced by existing practices and structures, whether that is company cultures or organizational norms [61]. To suggest that an AI model is "biased" and only needs to be tweaked is to adopt a far too narrow scope, missing out on broader or more systemic issues. As Lauer [39] suggests, "the failure to build ethical AI can be traced to an organization-wide failure of ethics." In this sense, the lack of meaningful engagement with ethical issues from engineers is a symptom of a deeper problem. Unethical AI is the logical byproduct of an unethical industry.

The toxicity of tech culture and its propagation of sexism and misogyny is well documented [81]. This is a culture known for the hypermasculine coder or "brogrammer," for using "booth babes" to attract attention at conventions, and for its celebrated company founders who regularly drop porn references [10]. One global ride-share company, renowned for its innovation and financial success, was long helmed by a man who penned a sex memo for a company celebration and who described his ability to pick up women as "boober" [50]. This type of activity, openly flaunted by some of the most worshiped companies and founders, has contributed to a highly misogynistic environment. In a survey of over 200 female tech workers with over 10 year experiences in Silicon Valley, 60% of women reported unwanted sexual advances [80].

The same toxic conditions can be seen when it comes to race. A recent class action lawsuit has accused a widely celebrated tech company of fostering racist conditions for years, including daily subjection to racial slurs, being assigned menial jobs in a segregated area of the factory, and being passed over in promotions for management [63]. Or we might think of the ten page "anti-diversity" memo that circulated at another major tech company renowned for its work in artificial intelligence, a screed suggesting white men were being marginalized and oppressed [13]. Despite claims of being a postracial meritocracy, tech culture is still one marked by white, male, heteronormative values—and those who fail to conform to this identity are discriminated against in subtle but material ways [56].

If the tech industry lacks ethics, so does the education of the software engineers and technologists who will soon join it. Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training [59]. Software engineering, computer science, and other degrees that lead into AI development are tightly focused on technical challenges and their solutions. But there is little to no consideration of ethical challenges—how technology intersects with race, class, and culture and how these might introduce new harms or exacerbate existing inequalities [68]. Despite the clear ethical dilemmas presented by emerging technologies, García-Holgado et al. [23] have observed a lack of integration of computer ethics in the computer science curriculum in Spanish universities. Similarly in Australia, Gorur et al. [26] surveyed 12 curricula in universities, finding that they focused on micro-ethical concepts like professionalism while lacking macro-ethical agendas such as betterment of society and the planet. Ethics units are rarely included in computer science courses, and several of these are even shunted into the last few sessions if time allows [24], demonstrating the lowly status of ethics in AI education.

The lack of ethical training in education, combined with the lack of ethical application in the industry, suggests that AI development takes place in an ethically empty milieu. AI technologists cannot be said to be "unethical" because that would imply an awareness of ethical norms and a decision to actively ignore or violate them. Instead, these technologies are conceptualized, developed, and brought to market in an "a-ethical" space, a realm where ethical dilemmas never even enter the frame. In this sense, the problem-space considered when developing a technology is far too narrow, failing to encompass the ethical, moral, and social impacts of designing a product in a particular way [62]. Given these conditions, the presence of an AI code of ethics which is tightly focused on a digital product or service appears entirely insufficient. Such an ethical framework, situated "downstream" from company culture, will fail to address the more fundamental inequalities and underlying social issues that shape technological development.

4 Toothless principles

Finally, AI ethical principles have failed due to the lack of consequences. Rességuier and Rodrigues [69] argue that currently AI ethics has no teeth, and this is because ethics is being used in place of regulation. Ethics is being asked to do something it was never designed to do. AI ethical frameworks can set normative ideals but lack the mechanisms to enforce compliance with these values and principles [69]. After surveying 22 sets of guidelines, Hagendorff [28] concludes that AI ethics is failing on many levels; they lack any enforcement mechanisms and their values are easily overwritten by economic incentives, often becoming little more than marketing devices.

Principles are not "self-enforcing," notes Calo [11], "and there are no tangible penalties to violating them." In 2019, for instance, Google announced the creation of a new independent body to review the company's AI practices. The Advanced Technology External Advisory Council, composed of philosophers, engineers, and policy experts, would review the company's projects and evaluate whether they contravened their AI principles. However, the group would have no actual power to veto projects or halt them in any meaningful way [36].

The dominant focus on (toothless) ethics is a boon to technology companies, who have long attempted to outrun or avoid legislation. Uber outpaced regulation by expanding rapidly into cities across the globe with a business model designed to bypass labor responsibilities and protections [50]. Similarly, Airbnb swiftly expanded worldwide, running for years in major centers before eventually confronting regulation around house rental and hotels. When legislation does catch up, companies attempt to impede, resist, or overturn regulations, as high profile legal cases involving Apple, Google, Facebook, and others demonstrate.

Legislation takes time to draft, pass, and enforce, and in this sense, Nemitz [54] describes the focus on AI ethics and the subsequent deferral of regulation as a genius move by corporations. Placing the production of ethical statements into the limelight allows tech operations to continue unchecked, unhindered by lawsuits, fines, or other penalties. Ochigame [57] concurs, asserting that ethical AI is "aligned strategically with a Silicon Valley effort seeking to avoid legally enforceable restrictions of controversial technologies." Nemitz [54] thus calls for ethics to be swiftly followed by legislation: the law has democratic legitimacy and can be enforced, producing a credible threat that AI powerhouses would need to take into account.

Toothlessness is not just about lack of penalties, but also about the lack of friction between ethical principles and existing business principles. Green [27, p 209] suggests that ethics is "vague and toothless" and is "subsumed into corporate logics and incentives." Values listed in AI ethics statements and proposed by AI ethics organizations adhere closely to corporate values (or as the first section demonstrated, can be interpreted in ways that align with them). Such principles slot neatly into existing corporate playbooks, rarely questioning "the business culture, revenue models, or incentive mechanisms that continuously push these products into the markets" [31, 43]. The Partnership on AI, for instance, touts itself as a non-profit community of diverse stakeholders ranging from academia and civil society to industry and media. The implicit claim of such an organization is to give a voice to the people and in this way counter corporate overreach or at least keep it in check. However, Ochigame [57] observes that the Partnership's recommendations "aligned consistently with the corporate agenda" and essentially served to legitimize the activity of AI powerhouses.

Toothlessness means that corporations can buffer their reputation by carrying out high profile work on ethical frameworks, confident in the fact that such ethics will not fundamentally alter their product affordances, organizational hierarchies, or quarterly earnings. In other words, companies can enjoy the appearance of ethics without the substance. Borrowing from the well-known concept of "green washing," this phenomenon of "ethics washing" as a means of dodging regulation has risen to prominence in debates on AI ethics [20, 30, 43, 82].

5 The principles/practice gap

The failure of AI ethical principles is not spectacular but silent, resulting in the desired outcome: business as usual. In his AI Debate statement, Calo [11] highlights this paradox. AI is hailed as revolutionary, a transformation that will disrupt work and life in myriad ways—and yet there has been significant resistance to updating legislation and regulation to manage this shift. The obsession with AI ethics perpetuates this paradox, upholding the rhetoric of AI innovation while never allowing AI's transformative potential to alter legal frameworks or impinge on technical operations in any meaningful way.

Business as usual suggests a gulf between ethical guidelines and practical implementation, a gap between principles and practice. This chasm becomes clear when we turn to the production environments where AI technologies are developed. Industry bodies such as the Association for Computing Machinery have adopted codes of ethics that are meant to guide and govern engineering practice. However, in a study of software engineering students and professional software developers, McNamara et al. [42] found that explicitly instructing developers to consider this ethical code had no discernible difference compared to a control group. Developers did not alter their established ways of working.

In another study, Vakkuri et al. [79] carried out interviews at five different companies which were involved in AI development. While all the participants acknowledged the importance of ethics, when asked whether their AI development practices took ethics into account, all respondents answered no [79]. Building on this empirically based research, the authors suggest that there is a significant gap between the research and practice of AI ethics [71]. In a later study, Vakkuri et al. [78, p 195] specifically examined the attitudes of developers in software startup environments, concluding that there is a "complete ignorance of ethical consideration in AI endeavors." Ethics, so lauded in the academy and the research institute, are shrugged off when entering the engineering labs and developer studios where technologies are actually constructed.

Recognizing the current gap between AI principles and AI practice, researchers and companies have aimed to make ethical values feasible and actionable in real-world settings. There is a drive to bridge this ethics/practice gap [73], to operationalize AI ethics principles [12, 47] and to translate principles into practices [49]. High-minded normative statements must be integrated in meaningful ways into datasets, production pipelines, and product features. Taking a cue from software-as-a-service, Morley et al. [48] suggest ethics

could function as a service composed of an independent multi-disciplinary ethics board, a collaboratively developed ethical code, and AI practitioners themselves.

However, operationalizing AI ethics promises to be difficult or even impossible, a daunting challenge underestimated by a technically focused industry and even by ethicists. Hagendorff [28, p 103] for instance, makes a number of salient points but also suggests that privacy and fairness, which occur frequently in ethical frameworks, are aspects for which "technical fixes can be or have already been developed." He goes on to suggest that "accountability, explainability, privacy, justice, but also other values such as robustness or safety are most easily operationalized mathematically and thus tend to be implemented in terms of technical solutions" [28, p 103]. The ease with which issues like fairness and privacy are waved off as being "resolved" is stunning. These are highly contested issues, with high stakes. What is fair and who gets to decide it? How might the notion of fairness respond to historical inequalities suffered by a particular people group? And how might fairness play out differently in different contexts and conditions? These are complex questions which have shifted substantially over time and which intersect with race, gender, and culture [29, 58].

Of course, this is not to suggest that there has been no work around these issues in computer science. When it comes to privacy, for instance, cloud-based technologies unlock new ways of grouping data entries or encrypting variables so that the ability to identify subjects or de-anonymize them is minimized [53]. But such work adopts one particular understanding of privacy and responds to it in one particular way. And even within this narrow scope, there are always trade-offs and workarounds that need to be considered [53]. The same point applies to related concepts such as fairness, safety, and justice, which can in no way be considered "resolved" by the limited technical responses to-date.

Operationalization is not simply a perfunctory matter of "translating" an ethical value into a technological outcome. There are tensions and trade-offs that must be worked through and worked out into the material form of a data model or a digital product. Krijger [37] suggests there are two key tensions that apply when attempting to operationalize AI ethics: an inter-principle tension, where competing ethical demands are placed on an AI design; and an intra-principle tension, which highlights the difficulty of materializing a principle into a technological form. Based on the insights above, then, we can suggest two hurdles to operationalization: (1) the challenge of wrestling with competing principles to arrive at meaningful demands and (2) the challenge of implementing those demands as concrete features, interfaces, and infrastructures. This is difficult work which requires engaging with social and political questions and prototyping, testing, and rejecting different designs: there are no shortcuts.

6 Alternatives to ethical principles

The dominant turn to AI principles is simultaneously a turn away from alternative approaches. In a zero-sum world, the immense human and financial resources poured into generating AI ethics frameworks funnels it away from other programs of action. It is not enough, then, to denounce AI ethics as fruitless or useless. Instead, a critical assessment of the impact of ethics work to-date must conclude that it is dangerous, hoarding expertise and funding that should be devoted to more effective work. The high stakes of AI—its ability to harm some of the most vulnerable communities and ecologies in material ways—only increases the urgency of recognizing this strategic misstep and its misallocation of resources.

What would be more productive approaches than the de-facto turn to ethical principles? One approach, in essence, is to think more broadly about AI justice. Zalnieriute [84, p 139] argues that the current focus on AI procedural issues like transparency is blinkered, acting as an "obfuscation and redirection from more substantive and fundamental questions about the concentration of power, substantial policies, and actions of technology behemoths." Similarly, Powles [66] suggests that concentrating tightly on bias distracts us from more fundamental and urgent questions about power and AI.

AI justice provides a useful term that productively expands the ethical scope of inquiry and intervention. As Gabriel [22, p 218] notes, AI justice "reframes much of the discussion around 'AI ethics' by drawing attention to the fact that the moral properties of algorithms are not internal to the models themselves but rather a product of the social systems within which they are deployed." If ethical principles are situated within company cultures and broader systems of power (as discussed in Sect. 2), then it makes sense to expand the scope of ethical engagement. Or, put differently, if machine learning reflects, reproduces, and amplifies structural inequalities, then any ethical program must operate intersectionally, considering a wide array of social and political dynamics [14].

What might this broader analysis entail? As a brief example, AI justice may allow us to reflect more critically upon the universal notion of the "human" in AI rhetoric and the often empty truism that we need to design AI to benefit "humanity." History has shown that some racial and ethnic groups were deemed more "human" and deserving than others, while others were considered less-than-human or even subhuman [51]. Similarly, AI justice may provide useful ways to problematize a taken-for-granted principle like "fairness" which appears across many ethical frameworks. Historically fairness has been defined by hegemonic groups in ways that perpetuate their advantage:

far from being "common sense," fairness is always historical and cultural with major racialized and gendered dimensions [83].

What might a commitment to AI justice look like in practice? At a concrete level, it may mean organizations engaging with groups that bear the brunt of AI impacts but are not typically consulted: children, people of color, LGBTQIA + communities, migrants, and other groups. Those who develop AI need to better understand the particular needs of these communities, and then work with them in meaningful ways to ensure that AI contributes to their well-being and does not exacerbate historical inequities. Large tech companies and "tech-forward" governments particularly have a role to play here in leading by example and thus establishing a blueprint for best-practice AI work moving forward.

How else might AI justice manifest? Considering justice in AI more broadly might mean confronting the longstanding relationship between capitalism and computation [4], recognizing the extent to which technologies have marginalized women [32], or considering the knowledge-systems that have been privileged and the indigenous epistemologies that have been ignored [74]. One specific strain of work has begun to think more concretely about ways to decolonise AI, unraveling histories of inequality and asymmetric systems of power [46]. However, this work is nascent and it remains unclear how AI technologies might be decolonised or even what that might entail [1]. This is difficult work that may entail acknowledging privilege, confronting corporate assumptions, or developing community consensus. In contrast to the prominent work on AI principles, Bender [6] suggests that this work of reversing power centralization and longstanding systems of oppression is harder and less trendy—but work on ethical AI is useless without it.

If AI justice and its invitation to broaden our ethical horizon is one approach, the other, in essence, is to think more narrowly. Such work does not invoke the grand scope of AI ethics, but often uses more mundane but better-understood terms: accuracy, alignment, mismatch, and impacts. The work of Timnit Gebru and her colleagues is exemplary in this regard. If facial analysis misclassifies subjects because the datasets are dominated by light-skinned subjects, then this problem might be partially diagnosed and addressed by introducing a new dataset balanced by gender and skin type [9]. If the provenance and origins or datasets used in AI productions are often obscured, then this problem could be mitigated through "datasheets," standard documents that lay out a datasets creation, collection method, limitations, recommended uses, and so on [25].

This is granular work or even gruntwork, the less spectacular labor that digs into the data infrastructures and digital substrates of machine learning and AI production. AI, after all, is material not magical, cobbled together from datasets, software libraries, engineering expertise and hardware-accelerated computation. As Joanna Bryson [7] reminds us, AI and machine learning occurs through design and produces a material artifact; auditing, governance, and legislation should be applied to correct sloppy or inadequate manufacturing, just as we do with other products. The basic idea across this strain of research is to make concrete progress in improving AI by breaking the often nebulous concept of "ethics" down into measurable metrics and discrete goals.

Two aims emerge when surveying this work. First, there is transparency. This is the ability to see how a system operates, to grasp what its assumptions are, and to understand how it responds to different contexts and situations. Oversight and auditing are key terms within this theme. As one example, Raji et al. [67] have proposed an end-to-end framework for AI production. The system would allow developers to audit their work at each stage and see how well it matches organizational principles. Such tools aim to provide better oversight about the kinds of decisions that are being made and the kinds of (potentially harmful) consequences that may result. In a similar vein, Mitchell et al. [44] propose model cards for model reporting. These short documents would accompany trained machine learning models and provide benchmarked evaluation. Such tools would allow developers to see how the model responds across a variety of different conditions, analyzing, for example, its performance across different demographic or phenotypic groups.

Once we can understand what is wrong with a model or system, we need an ability to act on this information. Transparency must then be accompanied by accountability. Recourse, responsibility, and governance are key terms here. There needs to be clearly defined lines of accountability and both producers and consumers of technology must have the ability to meaningfully address harmful AI technologies. Redressing these harms might entail redesigning a product, consulting members of a community, or halting an AI service altogether. And accountability must be backed up by enforcement: lawsuits, fines, or banning from a particular jurisdiction. Such aims suggest a place for conventional governance structures using managerial hierarchies and humans in the loop to identify responsibility within an organization [8]. Equally, however, they suggest grassroot efforts that aim to redress harms by reimagining data and models in ways that better suit the needs of a particular community [71].

Taken together, these alternative approaches of thinking more broadly and more narrowly suggest that many different stakeholders have a part to play in reshaping AI. Designers and developers are able to code up particular affordances and integrate them into digital products and platforms. Managers can take the lead in implementing testing and auditing libraries. Business and community leaders can establish cultures which are reflective and open to forms of critical questioning and exploration. Governments can create new policy

mechanisms and enforce compliance by properly resourcing the relevant agencies. And even more minor actors like professional societies and insurance companies can exert force through codes of conduct and defining certain practices as risky. Together, these twin approaches go beyond ethical principles, making progress in this critical area by reflecting deeply and radically about the potentials and pitfalls of AI.

Acknowledgements Thanks to both reviewers from AI and Ethics for their thoughtful and constructive feedback, which led me to extend the alternatives section and more sharply articulate some of the key points and claims.

Author contributions Not applicable (solo-authored article).

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Not applicable.

Availability of data and materials Not applicable.

Declarations

Conflict of interest There are no competing interests in regards to this article or the author.

References

1. Adams, R.: Can artificial intelligence be decolonized? Interdiscipl. Sci. Rev. **46**(1–2), 176–197 (2021)
2. AI Ethicist.: AI principles. AI Ethicist (2021). https://www.aiethicist.org/ai-principles
3. Amsler, S.S.: Embracing the politics of ambiguity: towards a normative theory of 'Sustainability.' Capital. Nat. Social. **20**(2), 111–125 (2009)
4. Beller, J.: The World Computer: Derivative Conditions of Racial Capitalism. Duke University Press, Durham (2021)
5. Bellin, J.: Pure privacy. Northwest. Univ. Law Rev. **116**(2), 463 (2021)
6. Bender, E.: Working out systems of governance, appropriate regulations & most importantly how to reverse modern power centralization & long-standing systems of oppression is both much harder and much less trendy. But work on 'Responsible' or 'Ethical' ML/'AI' is useless without it. Tweet. Twitter (2022). https://twitter.com/emilymbender/status/1529556392268468224
7. Bryson, J.J.: The artificial intelligence of the ethics of artificial intelligence: an introductory overview for law and regulation. In: Dubber, M.D., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, New York (2020)
8. Buckley, R.P., Zetzsche, D.A., Arner, D.W., Tang, B.W.: Regulating artificial intelligence in finance: putting the human in the loop. Sydney Law Rev. **43**(1), 43–81 (2021)
9. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91. PMLR (2018)
10. Burleigh, N.: What silicon valley thinks of women. Newsweek **28**, 2015 (2015)
11. Calo, R.: Remark at AI debate 2. (2021). https://www.youtube.com/watch?v=XoYYpLIoxf0
12. Canca, C.: Operationalizing AI ethics principles. Commun. ACM **63**(12), 18–21 (2020)
13. Conger, K.: Exclusive: Here's the full 10-page anti-diversity screed circulating internally at Google. Gizmodo (2017). August 5, 2017. https://gizmodo.com/exclusive-heres-the-full-10-page-anti-diversity-screed-1797564320
14. Davis, J.L., Williams, A., Yang, M.W.: Algorithmic reparation. Big Data Soc. **8**(2), 20539517211044810 (2021)
15. DeepMind.: Ethics & society. (2020). https://www.deepmind.com/about/ethics-and-society. Accessed 25 May 2022
16. Desai, J.N., Pandian, S., Vij, R.K.: Big data analytics in upstream oil and gas industries for sustainable exploration and development: a review. Environ. Technol. Innov. **21**, 101186 (2021)
17. Deutsche Telekom.: AI guidelines. (2018). April 24, 2018. https://www.telekom.com/resource/blob/532446/f32ea4f5726ff3ed3902e97dd945fa14/dl-180710-ki-leitlinien-en-data.pdf
18. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, New York (2018)
19. European Group on Ethics in Science and New Technologies.: Statement on artificial intelligence, robotics and 'autonomous' systems: Brussels, 9 March 2018. European Commission, Brussels (2018). https://data.europa.eu/doi/10.2777/531856
20. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**(2), 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x
21. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. Harv. Data Sci. Rev. (2019). https://doi.org/10.1162/99608f92.8cd550d1
22. Gabriel, I.: Toward a theory of justice for artificial intelligence. Daedalus **151**(2), 218–231 (2022). https://doi.org/10.1162/daed_a_01911
23. García-Holgado, A., García-Peñalvo, F.J., Therón, R., Vázquez-Ingelmo, A., Gamazo, A., González-González, C.S., Iranzo, R.M.G., Silveira, I.F., Forment, M.A.: Development of a SPOC of computer ethics for students of computer science degree. In: 2021 XI International Conference on Virtual Campus (JICV), pp. 1–3. IEEE (2021)
24. Garrett, N., Beard, N., Fiesler, C.: More than 'If Time Allows': the role of ethics in AI education. Proc. AAAI/ACM Conference on AI, Ethics, and Society, pp. 272–8. (2020). https://dl.acm.org/doi/abs/10.1145/3375627.3375868.
25. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Commun. ACM **64**(12), 86–92 (2021). https://doi.org/10.1145/3458723
26. Gorur, R., Hoon, L., Kowal, E.: Computer science ethics education in Australia—a work in progress. In: 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), pp. 945–947. (2020). https://doi.org/10.1109/TALE48869.2020.9368375.

27. Green, B.: The contestation of tech ethics: a sociotechnical approach to technology ethics in practice. J. Soc. Comput. **2**(3), 209–225 (2021). https://doi.org/10.23919/JSC.2021.0018
28. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
29. Hanna, A., Denton, E., Smart, A., Smith-Loud, J.: Towards a critical race methodology in algorithmic fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 501–512 (2020)
30. Hao, K.: In 2020, let's stop AI ethics-washing and actually do something. *MIT Technology Review* (2019). https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/
31. Hickok, M.: Lessons learned from AI ethics principles for future actions. AI Ethics **1**(1), 41–47 (2021)
32. Hicks, M.: Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing. MIT Press, Cambridge (2018). https://mitpress.mit.edu/books/programmed-inequality
33. Hogan, M.: Data flows and water woes: the Utah Data Center. Big Data Soc. **2**(2), 205395171559242 (2015). https://doi.org/10.1177/2053951715592429
34. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.: Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems. IEEE (2019). https://ethicsinaction.ieee.org/
35. Khan, A.A., Badshah, S., Liang, P., Khan, B., Waseem, M., Niazi, M., Akbar, M.A.: Ethics of AI: a systematic literature review of principles and challenges. (2021). http://arxiv.org/abs/2109.07906 [Cs]
36. Knight, W.: Google appoints an 'AI Council' to head off controversy, but it proves controversial. MIT Technol. Rev. (2019). March 26, 2019. https://www.technologyreview.com/2019/03/26/136376/google-appoints-an-ai-council-to-head-off-controversy-but-it-proves-controversial/
37. Krijger, J.: Enter the metrics: critical theory and organizational operationalization of AI ethics. AI Soc. (2021). https://doi.org/10.1007/s00146-021-01256-3
38. Latour, B.: Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford (2007)
39. Lauer, D.: You cannot have AI ethics without ethics. AI Ethics **1**(1), 21–25 (2021). https://doi.org/10.1007/s43681-020-00013-4
40. Luke, T.W.: Neither sustainable nor development: reconsidering sustainability in development. Sustain. Dev. **13**(4), 228–238 (2005)
41. Maxwell, R., Miller, T.: Greening the Media. Oxford University Press, Oxford (2012)
42. McNamara, A., Smith, J., Murphy-Hill, E.: 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 (2018)
43. Metzinger, T.: Ethics washing made in Europe. Der Tagesspiegel Online. (2019). April 8, 2019. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html
44. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Deborah Raji, I., Gebru, T.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–229 (2019)
45. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
46. Mohamed, S., Png, M.-T., Isaac, W.: Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. Philos. Technol. **33**(4), 659–684 (2020). https://doi.org/10.1007/s13347-020-00405-8
47. Mökander, J., Floridi, L.: Operationalising AI governance through ethics-based auditing: an industry case study. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00171-7
48. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., Floridi, L.: Ethics as a service: a pragmatic operationalisation of AI ethics. Mind. Mach. **31**(2), 239–256 (2021)
49. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In: Ethics, Governance, and Policies in Artificial Intelligence, pp. 153–183. Springer, Berlin (2021)
50. Munn, L.: Cash burning machine: uber's logic of planetary expansion. TripleC Commun. Capital. Crit. Open Access J. Glob. Sustain. Inf. Soc. **17**(2), 1–17 (2019). https://doi.org/10.31269/triplec.v17i2.1097
51. Munn, L.: Automation is a Myth. Stanford University Press, Stanford (2022)
52. Munn, L.: Data and the new oil: cloud computing's lubrication of the petrotechnical□. J. Environ. Media **2**(2), 211–227 (2022). https://doi.org/10.1386/jem_00063_1
53. Munn, L., Hristova, T., Magee, L.: Clouded data: privacy and the promise of encryption. Big Data Soc. **6**(1), 2053951719848781 (2019). https://doi.org/10.1177/2053951719848781
54. Nemitz, P.: Constitutional democracy and technology in the age of artificial intelligence. Philos. Trans. Roy. Soc. A Math. Phys. Eng. Sci. **376**(2133), 1–13 (2018). https://doi.org/10.1098/rsta.2018.0089
55. Noble, S.: Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press, New York (2018)
56. Noble, S., Roberts, S.: Technological Elites, the Meritocracy, and Postracial Myths in Silicon Valley. Duke University Press, Durham (2019)
57. Ochigame, R.: How big tech manipulates academia to avoid regulation. The Intercept. (2019). December 21, 2019. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/
58. Ochigame, R.: The long history of algorithmic fairness. Phenomenal World (blog). (2020). January 30, 2020. https://www.phenomenalworld.org/analysis/long-history-algorithmic-fairness/
59. Oliver, J.C., McNeil, T.: Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training and domain-specific context. PeerJ Comput. Sci. **7**, e441 (2021)
60. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Penguin Books, London (2018)
61. Orlikowski, W.J.: The duality of technology: rethinking the concept of technology in organizations. Organ. Sci. **3**(3), 398–427 (1992)
62. Oswald, D.: From ethics to politics: if design is problem solving, what then are the problems. In: Proceedings of the 18th International Conference on Engineering and Product Design Education, pp. 620–625. Aalborg (2016)
63. Paul, K.: Black workers accused tesla of racism for years. Now California Is Stepping In. The Guardian. (2022). February 19, 2022. https://www.theguardian.com/technology/2022/feb/18/tesla-california-racial-harassment-discrimination-lawsuit
64. Pazzanese, C.: Ethical concerns mount as AI takes bigger decision-making role. Harvard Gazette (blog). (2020). October 26, 2020. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/
65. Pontifical Academy for Life.: Rome Call for AI Ethics. The Vatican, Rome (2020). https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf
66. Powles, J.: The seductive diversion of 'Solving' bias in artificial intelligence. OneZero (blog). (2018). December 7, 2018. https://

onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53
67. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P.: 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, pp. 33–44. FAT* '20. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3351095.3372873.
68. Reidy, M.: Lack of ethics education for computer programmers shocks expert. Stuff. (2017). July 1, 2017. https://www.stuff.co.nz/business/innovation/93629356/minimal-ethics-education-for-computer-programmers
69. Rességuier, A., Rodrigues, R.: AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc. **7**(2), 2053951720942541 (2020)
70. Resseguier, A., Rodrigues, R.: Ethics as attention to context: recommendations for the ethics of artificial intelligence. Open Res. Europe **1**(27), 27 (2021)
71. Sambasivan, N., Arnesen, E., Hutchinson, B., Prabhakaran, V.: Non-portability of algorithmic fairness in India. (2020). http://arxiv.org/abs/2012.03659 [Cs]
72. Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. "What's next for Ai Ethics, Policy, and Governance? A Global Overview." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–58.
73. Shneiderman, B.: Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans. Interact. Intell. Syst. **10**(4), 1–31 (2020). https://doi.org/10.1145/3419764
74. Smith, L.T.: Decolonizing Methodologies: Research and Indigenous Peoples. Zed Books Ltd, London (2021)
75. Solove, D.J.: A taxonomy of privacy. Univ. PA Law Rev. **154**, 477 (2005)
76. UNESCO.: Recommendation on the ethics of artificial intelligence. UNESCO (2020). February 27, 2020. https://en.unesco.org/artificial-intelligence/ethics
77. Université de Montréal.: The declaration. Université de Montréal, Montreal (2018). https://www.montrealdeclaration-responsibleai.com/the-declaration
78. Vakkuri, V., Kemell, K.-K., Jantunen, M., Abrahamsson, P.: 'This Is Just a Prototype': how ethics are ignored in software startup-like environments. In: International Conference on Agile Software Development, pp. 195–210. Springer, Cham (2020)
79. Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., Abrahamsson, P.: Ethically aligned design of autonomous systems: industry viewpoint and an empirical study. https://doi.org/10.48550/arXiv.1906.07946 (2019). http://arxiv.org/abs/1906.07946
80. Vassallo, T., Levy, E., Madansky, M., Mickell, H., Porter, B., Leas, M., Oberweis, J.: Elephant in the valley. The Elephant in the Valley. (2016). 2016. https://www.elephantinthevalley.com/.
81. Wachter-Boettcher, S.: Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech. WW Norton & Company, London (2017)
82. Wagner, B.: Ethics as an escape from regulation: from 'Ethics-Washing' to ethics-shopping? In: Bayamlioğlu, E., Baraliuc, I., Janssens, L., Hildebrandt, M. (eds.) Being profiled, pp. 84–89. Amsterdam University Press, Amsterdam (2018). https://doi.org/10.2307/j.ctvhrd092.18
83. Weinberg, L.: Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ML fairness approaches. J. Artif. Intell. Res. **74**, 1–35 (2022). https://doi.org/10.1613/jair.1.13196
84. Zalnieriute, M.: 'Transparency-Washing' in the digital age: a corporate agenda of procedural fetishism. In: The Digital Age: A Corporate Agenda of Procedural Fetishism, pp. 21–33. (2021)