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)

From Ethics Washing to Ethics Bashing

A View on Tech Ethics from Within Moral Philosophy

Elettra Bietti
Harvard Law School
Cambridge, MA, USA
ebietti@sjd.law.harvard.edu

ABSTRACT

The word 'ethics' is under siege in technology policy circles. Weaponized in support of deregulation, self-regulation or hands-off governance, "ethics" is increasingly identified with technology companies' self-regulatory efforts and with shallow appearances of ethical behavior. So-called "ethics washing" by tech companies is on the rise, prompting criticism and scrutiny from scholars and the tech community at large. In parallel to the growth of ethics washing, its condemnation has led to a tendency to engage in "ethics bashing." This consists in the trivialization of ethics and moral philosophy now understood as discrete tools or pre-formed social structures such as ethics boards, self-governance schemes or stakeholder groups.

The misunderstandings underlying ethics bashing are at least three-fold: (a) philosophy and "ethics" are seen as a communications strategy and as a form of instrumentalized cover-up or façade for unethical behavior, (b) philosophy is understood in opposition and as alternative to political representation and social organizing and (c) the role and importance of moral philosophy is downplayed and portrayed as mere "ivory tower" intellectualization of complex problems that need to be dealt with in practice.

This paper argues that the rhetoric of ethics and morality should not be reductively instrumentalized, either by the industry in the form of "ethics washing," or by scholars and policy-makers in the form of "ethics bashing." Grappling with the role of philosophy and ethics requires moving beyond both tendencies and seeing ethics as a mode of inquiry that facilitates the evaluation of competing tech policy strategies. In other words, we must resist narrow reductivism of moral philosophy as instrumentalized performance and renew our faith in its intrinsic moral value as a mode of knowledge-seeking and inquiry. Far from mandating a self-regulatory scheme or a given governance structure, moral philosophy in fact facilitates the questioning and reconsideration of any given practice, situating it within a complex web of legal, political and economic institutions. Moral philosophy indeed can shed new light on human practices by adding needed perspective, explaining the relationship between technology and other worthy goals, situating technology within the human, the social, the political. It has become urgent to start considering technology ethics also from within and not only from outside of ethics.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/01…$15.00
https://doi.org/10.1145/3351095.3372860

KEYWORDS

Moral Philosophy, Ethics, Technology Ethics, Regulation, Self-regulation, Technology Law, AI.

Elettra Bietti. 2019. From Ethics Washing to Ethics Bashing
A View on Tech Ethics from Within Moral Philosophy. In *Proceedings of ACM FAT* Conference (FAT* 2019). ACM, New York, NY, USA, 10 pages.* https://doi.org/10.1145/3351095.3372860

1 Introduction

On May 26th 2019, Google announced that it would put in place an external advisory council for the responsible development of AI [1], the Advanced Technology External Advisory Council (or ATEAC). Following a petition signed by 2,556 Google workers [2] demanding the removal of one of the body's board members, anti-LGBT advocate Kay Coles James [3], the advisory body was withdrawn approximately one week after its announcement. This episode and the backlash it produced provide a salient illustration of the tensions around the use of "ethics" language in technology policy. Instrumentalization and misuse of such language in technology policy has recently proliferated and taken two forms.

On the one hand, the term has been used by companies as an acceptable façade that justifies deregulation, self-regulation or market driven governance, and is increasingly identified with technology companies' self-interested adoption of appearances of ethical behavior. We call such growing instrumentalization of ethical language by tech companies "ethics washing."[4] Beyond AI ethics councils, ethics washing includes other attempts at simplifying the value of ethical work, which often form part of a corporate communications strategy: the hiring of in-house moral philosophers who have little power to shape internal company policies; the focus on humane design – e.g. nudging users to reduce time spent on apps – instead of tackling the risks inherent in the

existence of the products themselves [5]; the funding of work on "fair" machine learning systems which positively obscures deeper questioning around the broader impacts of those systems on society [6].

On the other hand, the technology community's criticism and scrutiny of instances of ethics washing often borders into the opposite fallacy, which we call "ethics bashing". This is a tendency, common amongst social scientists and non-philosophers, to trivialize "ethics" and "moral philosophy" by reducing more capacious forms of moral inquiry to the narrow conventional heuristics or misused corporate language they seek to criticize. Equating serious engagement in moral argument with the social and political dynamics within ethics boards, or understanding ethics as a political stance which is antithetic to – instead of complementary to – serious engagement in democratic decision-making, is a frequent and dangerous fallacy. The misunderstandings underlying ethics bashing are at least three-fold: (a) philosophy and "ethics" are seen as a communications strategy and as a form of cover-up or façade for unethical behavior, (b) the role and importance of moral philosophy is downplayed and portrayed as mere "ivory tower" intellectualization of complex problems that need to be dealt with in practice; and (c) philosophy is understood in opposition and as alternative to political representation and social organizing.

Grappling with the role of philosophy and ethics in tech policy requires moving beyond both ethics washing and ethics bashing and seeing ethics as a mode of inquiry. We do moral theorizing all of the time. When we ask whether a corporate ethics council can improve internal policy-making, whether a given machine learning system can lead to fairer criminal justice enforcement, or whether a given corporate decision is acceptable, we are asking moral questions that, properly framed, can lead to a better understanding of these phenomena and also to better policies. Becoming aware of this fact enables us to see things more subtly, at several levels of abstraction, and to more rigorously assess the legitimacy of corporate self-regulation and other ethics initiatives.

An important distinction must be made between the intrinsic and the instrumental value of ethics. The first is the value of ethics as a mode of inquiry which it is independently valuable when engaged in as an aspirational process and which takes moral principles seriously in achieving better knowledge and understanding of a state of affairs or phenomenon. Understood as ethical commitment, ethics here is about engaging in a justice-seeking process with or without others in the belief that the process itself has independent moral value. The instrumental perspective instead sees the value of ethics as lying in its results. These results can be good or bad. As employed in ethical codes of conduct, professional ethics, or 'ethics boards', ethics is a means to an end, it is instrumental to the achievement of certain more interesting or valuable outcomes such as reputation, innovation, profit, the integrity of a profession. Ethics understood in this instrumental way has no value independent of its end-results, it is not an internalized aspirational mode of inquiry that aims at a better world, but is rather valued for its causal role in bringing about other desired results.

Intrinsic and instrumental perspectives on ethics are not mutually exclusive and can exist side-by-side. We can value ethics as an intrinsically valuable process and also and at the same time appreciate the understanding and generous mindset that engaging in it enables. However, it will be argued that the more the process of engaging in ethics is motivated by outcomes independent of the process itself, the less it is taken as an aspirational and intrinsically valuable process, the more doubtful its moral value becomes for society. Ethics washing and ethics bashing are instrumental understandings of ethics, in that both positions or tendencies envision or experience ethics as a means to an end and nothing more.

Further, what is at stake in recent controversies around the weaponization of "ethics" rhetoric are competing thinner and thicker moral conceptions of technology companies' role, the former arguably being promoted through narrow instrumental understandings of the role of ethical work, the latter arguably being promoted through greater participatory democracy and activism. Yet this understanding obscures the potential role of ethics within a thicker conception of technology policy. The narrower the lens one uses to look at an ethical problem, the narrower and more limited the response one is willing to offer to address it. As will be argued in what follows, it is important to maintain a critical outlook on the instrumentalization of ethics in technology settings, while also recognizing and respecting its moral worth as an exercise and mode of inquiry capable of expanding our horizons and thickening our moral commitments.

This paper has three goals. First, it aims to articulate the weaknesses of both the ethics washing and ethics bashing fallacies, explaining why both are impoverished views of the relationship between technology and ethics. Second, it aims to clarify the role and importance of moral philosophy in debates about the impact of new technologies on society and to dissipate misunderstandings according to which moral philosophy is either too abstract to inform concrete policy or is a red herring that prevents proper focus on political and social action. Far from constituting a barrier to appropriate governance, moral philosophy enables us to seriously scrutinize the future of technology governance, law and policy, and to understand what humans need from new technologies and innovation from a unique vantage point. Third and finally, adopting a less instrumentalized view of moral philosophy from within allows us to be less deferential toward philosophical work in technology settings, enhancing our ability to scrutinize certain philosophical ideas or moral stances and the impacts they can have on technology and society without bashing an entire field of inquiry.

The paper is structured as follows. It first explains the function and meaning of ethics and moral philosophy, some common criticisms of moral philosophy and what it is for. It then clarifies what is

wrong with ethics washing, adopting a view from within moral philosophy. And finally, it clarifies what is wrong with ethics bashing concluding that we should move beyond both tech washing and tech bashing, adopt a less instrumentalist position on ethics and start taking moral philosophy seriously as a discipline and mode of inquiry.

2 A Word on Ethics and Moral Philosophy

The English word "ethics" is derived from the ancient Greek words *ēthikós* and *êthos* which refer to character and moral nature [7]. Morality comes from the Latin *moralis* which literally means manner, character, proper behavior. Both "ethics" and "morality" thus refer to the study of good and bad character, appropriate behavior and virtue. The two terms are often employed interchangeably but have slightly distinct uses and connotations. Morality is often associated with etiquette, and rules of appropriate social behavior, whereas ethics has instead a more personal connotation. Ethics pertains to the cultivation of individual virtue abstracted from society, and is sometimes used to refer to personal and professional standards of behavior embodied in "codes of ethics". In Confucian philosophy, morality is about respecting the family and pursuing social harmony and stability through virtues including altruism, loyalty, piety [8].

In the discussion to follow, the term "ethics" will refer to the rhetoric of morality employed in technology circles, and "moral philosophy" will instead refer to the philosophical discipline that investigates questions around human agency, freedom, responsibility, blame, and the relationships between individuals, amongst other questions. According to some accounts within what we are calling moral philosophy, the scope of the notion of moral philosophy is limited to relationships between humans whereas the notion of ethics extends beyond humans to animals and nature. Some would also distinguish moral from political philosophy while others such as Ronald Dworkin see them as interconnected [9]. Like Ronald Dworkin, I construe the "moral" widely as consisting of the domain of "value," i.e. an evaluative and interpretative mode of inquiry which one can distinguish from scientific or descriptive modes of inquiry, those that only (purportedly) pertain to facts [10]. The domain of "value" is the specific domain of inquiry of moral philosophers, and is sometimes considered to border into other domains of philosophy such as aesthetics, the study of beauty and aesthetic value, or epistemology, the study of knowledge and belief.

To better illustrate what moral philosophy is, let us use the example of surveillance. Let us ask: what is wrong or unethical about certain forms of surveillance? Disparate arguments can be offered to show that surveillance is wrong in some respects or worth carrying out in other respects. Different persons will likely have different views on which of these arguments are strongest. As philosophers might put it: the morality of surveillance is an evaluative matter, one that we might disagree on, a question of value, or otherwise put a moral question. Possible lines of reasoning supporting the wrongness of surveillance are as follows. Surveillance is objectionable on self-development and virtue ethics grounds because it incentivizes self-censorship, reduces human beings' ability to develop themselves or to engage in other valuable causes because of a fear that these actions will be held against them. Another argument focuses on harm: some forms of surveillance cause harm to individuals (e.g. they lead to unjustified and stereotype-enhancing discriminatory treatment, they create asymmetries of knowledge and power, they perpetuate pre-existing and unjustified inequalities). A third line of argument focuses on equal dignity and respect for persons: some forms of surveillance fail to treat individuals as equally worthy of respect because they are covert and because some people are surveilled more than others. There are many other possible lines of reasoning for why surveillance might be considered wrong in given circumstances.

Each line of reasoning points to different policy solutions. For instance, if we believe it is key to enable the pursuit of worthy behavior and individuality and that the core reason to resist surveillance is that it inhibits such behavior or individuality, we might be satisfied with aspects of surveillance that enhance the pursuit of certain worthy life goals, including targeted and personalized services. On the other hand, if we believe the core problem is that the information that is collected can cause harm to individuals, we might be prone to advocate for solutions that minimize discriminatory impacts and ensure that harms are reduced. Finally, if we believe surveillance leads to a degradation of human dignity and a failure to treat individuals with respect, we might be prone to ban surveillance completely, or to advocate for the leveling down of surveillance to a *de minimis* threshold. Which reasons we find most weighty is a matter of philosophical commitment and deliberation. The process of weighing our reasons against others' reasons allows us to overcome the intuitive and primitive belief that "surveillance is bad because I feel it is," to reject weak arguments and to ground or re-evaluate our position based on carefully weighed stronger reasons. Identifying the drawbacks of surveillance and its morally unacceptable core also allows us to devise nuanced concrete solutions for addressing it.

This process of revising and refining our moral beliefs through philosophical inquiry is what John Rawls has called *reflective equilibrium* [11]. This process is not, or not only, about choosing one theoretical approach to morality and applying it to all factual scenarios - be it consequentialism, deontology, or virtue ethics for example. Instead, it entails engaging with an issue of societal importance, locating it within existing debates, considering it from all relevant standpoints, and making an evaluative judgment as to which angle or way of approaching it is capable of shedding the most valuable light on it, and can best guide a strategy for addressing it. The broader the spectrum of considerations we take into account in our moral theorizing, the more interesting,

capacious and morally significant the result of an inquiry from within moral philosophy, the more inspiring and valuable its practical implications.

It is also important to emphasize that moral philosophy and ethics can mean different things as part of different fields of study and intellectual traditions. The above is intended to capture only a glimpse of a larger roadmap of possible uses of the terminology of ethics and moral philosophy in technology governance and policy. It is not intended to fix the meaning of these rich and complex modes of inquiry.

3 The Limits of Moral Philosophy

Work in moral philosophy and ethics has a number of limitations. Before turning to an analysis of how it can inform the debates on ethics washing and ethics bashing, we should recognize four common criticisms of the moral philosophy approach that is defended in this paper and that is relied on as a lens to develop objections to ethics washing, ethics bashing and, as part of a reflexive exercise, the instrumentalization of moral philosophy itself.

First, philosophy is sometimes criticized for being abstract and for not being accessible to large audiences. This makes philosophical work often unsuited to advocacy or activism or to making provocative contributions to time-sensitive issues. Philosophy is also rarely suited to op-eds, for example, or to those who aim at quick and easy policy fixes. Yet depth and abstraction are also one of the discipline's advantages: engaging with philosophical work prompts us to pause and think, to shield our thinking from pragmatic pressures, to enlarge the temporal and geographical scope of our research scope. As we engage in this process, our intuitions change, we extend our thoughts or revise them so that they can connect with and make sense of other problems, we learn how to think slower, to think with more depth and more systematically. We need more of this kind of slowness in technology scholarship.

Second, some work in moral philosophy, particularly in its connections with technology, is criticized for not going far enough prescriptively. Laying out general abstract principles without explaining how they apply to real life situations seems to falls short when it comes to making sense of urgent social problems, such as many of those that arise in relation to new technologies. Recent philosophical work, for example, has been focusing on how the trolley problem can guide the regulation of autonomous vehicles [12]. Far from telling us what ought to be done in different life scenarios, some of the best work in this field offers a higher level lens for understanding the role of trolley-based thinking in technological design. In the absence of a deep understanding of context, focusing on the trolley problem seems unlikely to lead to any workable and morally compelling regulatory solutions for autonomous vehicles. This and other similar examples leave many perplexed by the meager functional value of philosophical work: much of it seems irrelevant or unsuited to resolving pressing problems in real contexts. Greater emphasis on the special epistemic value philosophical work can add in given technological contexts could possibly address this limitation [13].

Third, in practice philosophical work can have effects in context that sometimes contradict the principles that motivated the work in the first place. Much has been said, for example, about the instrumentalization of Hegel and Nietzsche's philosophical ideas by the German nazi regime for their own inhumane ends, an instrumentalization that had little connection to what these philosophers were actually doing or thinking [14]. More concretely, the political candor often associated with philosophical work facilitates its frequent instrumentalization for unworthy ends. This happens in the technology space. The hiring of moral philosophers by technology companies is one example. Philosophers are hired, their skills are transformed into a service and subordinated to the commercial goals of their employers. In this way, work that might have seemed unproblematic, justified or even welcome in an academic setting can become positively harmful as a mode of reputational propaganda for corporates that reinforces stereotypes, serving some interests to the detriment of others. As important as it is, this criticism is not fatal to the kind of work philosophers do. The emergence of in-house philosophers means philosophical work must be scrutinized with even greater care, and philosophers must exercise an enhanced level of caution regarding the consequences of what they do. It may not be harmful, for instance, for some practically-oriented branches of philosophy to become more openly attuned to politics. And more importantly, it is time for the funding of philosophical work in the technology and governance field to become more openly disclosed and discussed.

Fourth, and importantly, philosophy is frequently criticized for creating an appearance of principled reasoning, neutrality and objectivity when much of what is at play are a philosopher's subjective views [15]. There is some validity to this criticism and many supporters of ethics bashing may have this intuitive criticism in mind. This critique of moral philosophy, however, is less powerful than it first appears. Good contemporary ethical work does not attempt to convey an appearance of absolute objectivity. Quite the contrary, such work is very clear regarding the uncertain bases on which it stands. As said, a large proportion of Anglo-American moral philosophy follows Rawls' reflective equilibrium methodology [16], whereby intuitions and beliefs are progressively made to match considered judgments. This iterative process is one many Anglo-American philosophers use to formulate conclusions. Although any philosophical conclusion necessarily originates in a thinker's subjective intuitions and beliefs, it is also the product of structured and iterative revisions that give such conclusion a solid and judgment-proof form that raw intuitions do not have. Far from presenting ultimate and final words on a subject, good philosophical work is rigorous yet porous and open to scrutiny, its aim is to broaden perspectives, allowing us to see the limits of the existing and to constantly revise our beliefs.

A common theme that straddles these criticisms is that moral philosophy can be a worthy enterprise but is too easily instrumentalized to serve unworthy goals. As philosophers and theorists, we should not only be aware of these vulnerabilities but must also combat them by embedding resistance to the exploitation and instrumentalization of moral inquiry into our very methodologies.

4 What Moral Philosophy Is For

Having considered the drawbacks of work in ethics and moral philosophy, we must now ask what the exercise of moral reasoning and inquiry can add to existing technology policy debates. Here the focus is not on how moral philosophy is instrumentalized, but on the worth of moral philosophy as a practice and an exercise that is taken seriously from within. Philosophical work is valuable and can add value in at least four ways.

First, philosophical reasoning and deliberation can act as a meta-level perspective from which to consider any disagreement relating to the governance of technology. Instead of taking arguments narrowly, intuitively or personally at face value, philosophical reasoning provides us with a framework for stepping back, situating any problem within its broader context and understanding it within or in relation to other relevant or analogous debates. As such, the practice or method of engaging in moral argument allows us to broaden our perspective on a debate, to look at it from a wider lens, overcoming confusions, filling in gaps, correcting inconsistencies and drawing clarifying distinctions. In the debate on surveillance examined above, for instance, a philosophical method can help us rethink our reasons for rejecting or promoting surveillance, it can help us clarify points of agreement with a variety of opponents and focus the discussion on where the real disagreement lies and what it entails in practice. Otherwise put, philosophy is a good antidote to knee-jerk reactions and ideological incompatibility, a method that can help us reduce unbridgeable value conflicts and make agreement possible by moving discussions to a different level of abstraction. This is not to say that ideology and value conflicts are unimportant, but merely to recognize the importance of philosophy as a method aimed at overcoming or clarifying those conflicts.

A second, related, contribution of moral philosophy to tech debates is that it can add a layer of rigorous principled thinking to value-laden discussions. Moral philosophy should be understood as an *explanatory* mode of inquiry which operates by requiring us to set out the justifications and reasons for advancing one view and not a different one. By centering attention on the explanation and the justification for a position rather than on defending the position itself, philosophy shifts and deepens our mindset. Winning the argument is no longer as important as placing all arguments on the table, as cards might be in a game. Evaluating these arguments' respective strength must precede the assessment of whether one of them is a winning one, and whether one position is philosophically sounder than another one. Such robust principled inquiry, which is what good philosophy is based on, is too frequently absent in technology policy and governance discussions, which are instead governed by instinctual reactions, topicality, dogmatism and reputational hubris. New technologies' fast-paced market-driven genealogy in other words is structurally inimical to principled and cautious reflection on desirable social and technical developments.

Third, a normative philosophical lens allows us to move beyond a narrow focus on procedural fairness, diversity and representation in technology governance. The problem is not just whether an AI ethics board's members come from a variety of backgrounds, but also whether the board's decisions actually constrain Google's actions for the benefit of the public or simply align with Google's incumbent interests. These substantive moral questions are questions that can be tackled from within moral philosophy. Whether to put a product on the market in spite of significant surveillance risks is one example, another one is whether to invest in building a product in the first place. A capacious understanding of moral philosophy allows us to move beyond checklists of procedural guarantees, and ask iteratively whether the outcomes of a given governance framework are morally acceptable and worth pursuing. In other words, it would be hardly morally justified to put in place a self-regulatory scheme that, although it operates independently and transparently, in fact leads to a consistent bias in favor of corporate interests.

Fourth, far from obscuring ideological conflicts and structural divisions [17], engaging in moral philosophy can facilitate dialogue, encourage the building of common ground, and provide a basis for collaborative and participatory approaches to policy-making capable of bridging divides in a polarized landscape. An important drawback of critical work that centers on power, value conflicts and unbridgeable ideological divides is that it renders dialogue between people holding different views or occupying different social positions more difficult. Pursuing such strategies has its advantages but it can also lead to fragmentation in an already polarized public sphere. Understanding philosophy as a dialectic discipline and grounding methodology in the aspirational possibilities of rationalization and conflict resolution can instead help us navigate fragmentation and polarization in the current climate [18]. Empirically, it has been shown that engaging in a discussion in the belief that agreement is a possible outcome can facilitate cooperation; this idea has found support in the literature on negotiation [19] and beyond academia [20].

Still, as we acknowledge the important contributions of Western philosophy to the promotion of an inclusive and discursive public sphere, we must also build within such discursive public sphere the awareness of power and inequality. Dialogue cannot always be premised on the idea that every human has the same voice and the same ability to be heard [21]. Equalizing a space in the face of structural inequality must thus be one of the first considerations when building of spaces for dialogue. It is indeed possible to devise a normative philosophical approach that embeds ideology and

structural asymmetries within normative philosophical inquiry, and that weaves those power and structural dimensions in the very articulation of what desirable dialogue or agreement means. Contemporary approaches that move in this direction [22] are able to maintain the benefits of a discursive methodology while expanding the horizon of philosophical inquiry to include issues of structural inequality, domination and ideological entrenchment.

All of these positive attributes of moral philosophy seem to hang on an understanding of it as a valuable pursuit independently of its effects. The following is an attempt to devise such a methodology and to apply it to issues of technology ethics, in particular to clarify the misconceptions that underlie both ethics washing and ethics bashing.

5 What's Wrong with Ethics *Washing*

Having examined some of the opportunities and limits of moral philosophy's role in informing technology policy, we must now ask what makes ethics-based practices as practiced in technology policy circles instead particularly problematic. Can moral philosophy help us evaluate the acceptability of these efforts? And should we call these efforts "ethics washing"?

As Google's ATEAC episode or the employment of ethicists by a number of companies show [23], companies such as Google, Apple, Microsoft, OpenAI, Palantir are increasingly making efforts from an ethical standpoint, and are particularly concerned about their ethical reputation in the face of new technological developments in AI and beyond. Putting in place boards of external experts and hiring moral philosophers who can engage in ethical thinking about the techniques and products being developed in-house indicates willingness on their part to add an internal layer of accountability and governance and to subject themselves to pre-emptive checks and internal constraints. The intentions behind these initiatives are often good, but they beg for further scrutiny. Notwithstanding good intentions, embedding philosophers or ethicists within technology companies is a double-edged sword and could shield these companies from capacious regulation more protective for consumers.

As we assess these initiatives, we are therefore pulled in two directions. On the one hand, we are tempted to welcome these developments as positive and as indicative of a willingness to embrace ethical issues. On the other hand, we are moved to criticize these company efforts for the potential harms they might bring about. Where we stand on this spectrum will often be tainted by our background, by the people who we trust or follow on social media, by who pays us, and by who we are. What moral philosophy as a method enables us to do is take a step back, to consider these two attitudes along a spectrum of more nuanced positions on companies' ethical behavior and to evaluate our reasons for supporting or resisting initiatives such as the Google ATEAC. It allows us to suspend our intuitive reactions and think them through by taking into account the reactions and reasons of others.

What is wrong with the instrumentalization of ethics language? And what is wrong with ethics boards or self-regulation even if they do not aim at reputational gains? An approach from within moral philosophy can guide us through these questions. The aim of the following analysis is not to attack ethics washing or what may look like ethics washing, but to guide us through a reality whose complexity neither the companies nor their critics are fairly portraying

Self-regulation and self-publicity at first both seem benign. Self-regulation in certain cases is not only tolerable but actually welcome, for instance where regulatory interference by a public agency is unlikely to be effective, and where a self-regulatory approach can lead to substantive policy improvements for individuals and society. Second, in principle, it does not seem morally objectionable to fund and develop initiatives that foster a positive image of one's business. It is not wrong for a business to engage in self-publicity and self-advocacy. Let us focus on a real case of self-regulation in relation to online content moderation.

In the United States governmental regulation of online speech is seen with suspicion. The solution to the regulation of online speech on Facebook has consequently materialized in the form of an internal Oversight Board (FOB), a quasi-judicial body set-up internally but composed of external experts to adjudicate on the acceptability of controversial user-content on the platform. The body has been praised as "*one of the most ambitious constitution-making projects of the modern era*," [24] and is seen as a workable and promising approach for taming Facebook's power over online content in the face of First Amendment restrictions on government regulation [25]. Nonetheless, while the Board may bring about needed transparency and an appearance that content moderation is being tackled fairly, we must look beyond Facebook's messaging to find its shortcomings. In spite of its carefully crafted set-up and the well-intentioned messaging around its existence, it is likely that the FOB will serve Facebook's interests more than users. First, it provides a clear and legitimate way for shielding Facebook from other forms of regulation and scrutiny on matters of content moderation and community guidelines, including the intervention of national or international courts but also the formulation and enforcement of legislative redlines and constraints. Second, by centering attention on content moderation and community guidelines, it allows Facebook to continue developing its Newsfeed algorithms as it pleases, and to continue showing individuals lucrative content, without interference from regulators or courts. Thus, far from addressing all questions of online speech harms, the FOB seems to divert attention toward some issues and away from the most pressing concerns around misinformation and political propaganda.

Self-regulatory initiatives such as the FOB should prompt us to look beyond appearances and ask whether their very existence, in spite of appearing useful and a step forward, might in fact performatively obscure more pressing problems and risk long-term

irreparable harm. The same might apply to AI ethics boards and in-house philosophers.

6 Three Critiques of Ethics Washing from Within Moral Philosophy

To explore the moral limits of these internal corporate efforts mostly aimed at developing more ethical artificial intelligence, we must again turn to moral philosophy. At least three possible arguments can be raised against initiatives that co-opt ethics language and self-regulation for internal purposes. First, the type of ethics work carried out within companies or ethics boards seems to lack instrumental value, it does not have beneficial effects on individuals and society, because it is undertaken under conditions that deny these beneficial effects. Second, these practices also seem to lack much of the intrinsic, or independent, value associated with philosophical inquiry and explored above insofar as they do not seem to be undertaken in good faith, or with the aim of achieving overall justice. Third, even if these ethics-based practices were carried out in absolute good faith and in the pursuit of justice, and thus maintained both their instrumental and intrinsic value, instrumentalizing ethics reasoning and language to pursue company goals generates a specific kind of epistemic concern. Indeed, it seems that the performative role of ethics language remains problematic even where, as the case of the Facebook Oversight Board has illustrated, these efforts are intended to address real issues and in fact have positive effects. This happens where in spite of having some instrumental value, these efforts instrumentalize ethics for the sake of other selfish or less valuable ends, yet are presented as if they exclusively served the public interest. The following explores these three arguments and their limits.

The first critique of self-regulation and company ethics is an argument grounded in the poor instrumental value, or narrow impact, of ethical work performed within a company. Another way of putting it is that as long as philosophical inquiry is carried out within the closed proprietary walls of tech companies, its contributions are likely to benefit companies more than society at large. The decisions of internal AI ethics committees are subjected to internal limits, subordinated to the endorsement of high management and dependent on company funding. This dependency on the company's benevolence makes such efforts inadequate for addressing serious cases of company misconduct and also importantly unfit for achieving desirable policy outcomes.

The narrow impact of ethics-based efforts carried out within tech companies is due in part to formal limitations on employee-philosophers' scope of work or on ethics boards' mandates. For example, Apple's philosopher in residence has been forbidden from making public appearances since he started working for the company and Microsoft's AI ethics board does not disclose the reasons for its decisions [26]. Formal limitations like these seriously curtail the value of the decisions taken by these individuals or bodies. Further, limited impacts are also due to more diffuse exercises of influence that shape the broader discourse around technological innovation and ethics. These more subtle forms of influence and constraint include companies' funding of research and policy initiatives that favor them, the careful selection of people to engage with and whose ideas are highlighted, including the people these companies choose to have as part of their ethics-based initiatives [27].

These formal and diffuse constraints on the work of in-house philosophers and ethics boards in turn affect the substance of the decisions they can issue. While some efforts can be made internally to make ethics boards more diverse and representative and in-house philosophers also more attuned to the politics of AI, it remains safe to say that these bodies or philosophers' decisions will remain conservative when it comes to questions that affect companies' shareholder profits, general strategies, and bottom line. These decisions more often than not tend to favor incumbents and the status-quo. Desirable shifts in policy can jeopardize the interests of companies: strong data protection guarantees, data minimization mandates, redlines on the use of AI in credit scoring, criminal procedure or content moderation could hardly be started from within a company's ethics board. This is especially the case in areas such as big data or AI where systemic fundamental rights concerns are in constant tension with company efforts at normalizing their practices and bottom lines. In this context, the role of in-house philosophers' decisions is likely to remain confined to steering, reviewing and advising on policies and product launches within the confines of existing business models. In extreme cases, they might contribute to advising against the launch of products on the pipeline. Yet it is likely that these cases will remain limited. As long as the ultimate decision-maker on any given AI policy is the company itself, internal ethics programs will keep benefiting incumbents more than users and society and will therefore lack instrumental value, all things considered.

The limits of a critique based on the instrumental value of ethics-based efforts are two-fold. First, it seems that the existence of such efforts and their actual positive impacts on society might outweigh any resistance to those efforts. Evidence of such positive impacts for society rather than for the companies themselves are yet to be seen, but our philosophical critique remains open to a consequentialist rebuttal in this sense. Second, our critique seems to neglect arguments based on procedural fairness. Provided real procedural and substantive teeth were placed on self-regulatory actions immunizing them from all forms of company pressures, perhaps this would make a substantive difference which would warrant a more deferential moral attitude toward the resulting decisions. In other words, it would be difficult to contest on instrumental grounds decisions that are taken under impeccable procedural conditions. This argument seems to hold in some cases but its merits are fact-specific. In many real-life settings, self-regulation carried out under impeccable procedural conditions and having real teeth seems an oxymoron.

Our second critique of so-called ethics washing looks at the act of engaging in these efforts by philosophers-in-residence, or members of ethics boards, and examines the intrinsic or independent value of the activity that these people engage in. Moral philosophy as a practice has value when engaged in in pursuit of independently valuable goals such as truth, justice or the well-being of society. To be valuable, engaging in moral argument must be done to a substantial extent out of commitment to moral principle, in the belief that it can lead to a better understanding of moral questions. If instead it is undertaken for the sake of earning money, pleasing employers or obtaining honors and recognitions, then the intrinsic value of the exercise becomes morally tainted, it loses that special mark of purity that makes us take the process and its outcomes seriously.

We might think that this critique is about the actual motivations of the philosophers and experts to engage in the exercise. When looking at cases of philosophers-in-residence, ethics boards, or academics who work closely with these companies, there are doubtless some individuals who do it to raise their profile or create connections that can lead to further work in the field, or even to obtain promotions, honors, or greater impact and salience for their work. Yet many also do it simply because they believe that their involvement might lead to a positive overall impact or in the hope of getting insights into how the company works. It is tempting to focus on these people's intentions and blame their shortsighted mindsets. Yet blame seems to fall short, not least because moral philosophy in academic settings is also characterized by similar tendencies to work for the sake of obtaining fame, money or academic reputability.

Instead, a better characterization of the independent value of ethics-based work is to say that it must be capacious, that actual commitment to moral principle requires going beyond and doing more than what philosophers are allowed to do within companies and corporate settings. It seems all right, for example, to say that a facial recognition algorithm should be reviewed because it systematically identifies white people more positively than black people. However rectifying such bias requires more than "fixing" the algorithm. It requires making sure that the algorithm is not deployed in settings where it might be used to cause irreparable harm to black people. It also possibly involves thinking about avoiding the use of such algorithms in the first place, and replacing them with human decision-making [6][28]. To the extent an ethics board or in-house philosopher engages in moral argument with a view to correcting the algorithm yet is prevented from considering or voluntarily ignores these other considerations, then that the exercise seems to lack substantial independent value. Such value requires full and unrestricted substantive commitment to moral principle and justice.

Third and finally, notwithstanding the intrinsic or instrumental value of these efforts or lack thereof, the expression "ethics washing" denotes a particular performative and epistemic function of the activities in question which requires distinct analysis. Ethics rhetoric may have the effect of freezing popular imagination and of preventing the emergence of valuable alternatives. It may promote and reinforce a narrow and confined vision of the possibilities of regulatory change, and inhibit dialogue.

It can, for example, mislead the public into believing that previously contested policies have now become acceptable, thus creating a legitimacy buffer for objectionable corporate action. Immunizing corporate action from public scrutiny is dangerous for more than one reason: apathy strengthens corporations and weakens activists, it shifts the burden of policing new technologies from deep-pocketed governments and private companies to poorly funded activist groups and other marginalized stakeholders. It can also discredit awareness enhancing efforts and narrow the spectrum of contestation and debate. Self-regulatory efforts, such as the example of the FOB provided above, tend to narrow the scope of a debate, marginalizing questions of structural injustice or disruptive change and instead centering attention around procedural fairness and fixable tweaks. This – predictably – ends up favoring the incumbents. Although the performative dimensions of ethics washing are easy to dismiss, they are in fact crucial to a comprehensive analysis of the moral value and acceptability of these efforts.

Overall, an analysis from within moral philosophy helps us see these in-house ethics washing efforts as lacking significant instrumental and intrinsic value and also as playing a performative function that can affect individuals and society. While there may be exceptions of companies really working to ensure the independence and actual valuable contribution of internal ethical work to a more just society, it is important for policy-makers not to overlook the salience and weight of these critiques of ethics washing in many of the existing cases of internal efforts at ethical work, particularly around AI.

6 What's Wrong with Ethics *Bashing*

If the reasons for criticizing and resisting ethics washing from within moral philosophy are likely to find support amongst moral philosophers and tech ethicists but also more widely, then why do we see so much ethics bashing in technology policy circles? This seems to happen for two main reasons: a linguistic misunderstanding, that is to say the conflation of instrumentalized ethics washing efforts with ethics as an aspirational exercise, and ignorance of or resistance to the possibilities and importance of moral philosophy.

The linguistic misunderstanding is due to what we have described in the last section as companies' cooptation of the language and performative function of "ethics" to pursue self-promotional goals. Instrumentalized and emptied of its instrumental and intrinsic value, what we might have wanted to optimistically call "ethics" now appears trapped between meanings: objectionable instances of self-regulation, static and incomplete lists of guiding principles and

other forms of narrow and conservative regulative "fixes". None of these embodied instances of the practice of ethics are actually likely to be fully morally defensible, but the word quickly becomes a buzz, it gets defended or criticized at face value by corporations and critics of corporate self-regulatory efforts at a similar level of reduction. These dynamics further entrench the misuse and instrumentalization of ethics language. In policy circles, the word becomes a red herring, a technique so flawed that it can hardly stand on its own legs let alone lead to better behavior in the industry. Much ink has been spilled in this way, explaining why "ethics" cannot fix technology companies or why adopting "ethical principles" does not address the risks attached to the development of AI systems.

Yet the misunderstanding at bottom is this: what might externally appear as an ethical practice, not necessarily is one. Indeed, the appearance of moral argument remains an appearance insofar as the instrumental and intrinsic value of the exercise remain secondary and are subordinated to other less morally significant and selfish goals. This is the case when the outcomes of such processes consistently favor the interests of industry stakeholders instead of the common good. A clear example would be an ethics board that consistently favored addictive products that lead to the lock-in of users in 'cool' new ways, yet mostly for the sake of enhanced profits.

Much of the ink used to bash "ethics" was perhaps justified but it could have been used more wisely by distinguishing ethics washing from the other meaning of ethics, the broad and capacious practice which requires us and moral philosophers alike to engage in principled thinking with a commitment to human improvement, freedom, equality and dignity. We too frequently neglect that "ethics" can and must encompass more than what companies make of it. Ethical practice, in fact, makes it possible to assess the competing or complementary merits of different kinds of regulation, including self-regulation and other forms of law and policy-making. Instrumentalizing and simplifying the meaning and use of ethics language is not only misleading, it is also counterproductive, and constitutes a missed opportunity to distinguish corporate washing from worthy scholarship and modes of thinking that can help us challenge it.

A richer critique of corporate self-regulatory efforts and of their cooptation of "ethics" rhetoric therefore demands that we operate at two levels: keep being critical of ethics washing, while also being aware that our very critique positions ourselves distinctly within moral philosophy. In other words, by criticizing or bashing certain practices we adopt a distinct moral stance that is within moral philosophy and not outside of it. We must thus be ready to engage more thoroughly with the flaws of narrow approaches to ethics and accept that defending more capacious ethical stances comes from a better understanding and awareness of moral philosophy's potential not a blank rejection of it as a language, practice, discipline and mode of inquiry. This requires a deep societal reckoning with moral philosophy; an understanding of what it is for and of its limits.

Criticized as complex, abstract, apolitical and misleadingly neutral or objective, philosophy gets a lot of bashing and is frequently dismissed in areas such a technology policy which are fast moving, full of ideological conflicts and in need of quick and effective responses. However today it is clear that the need for quick and effective fixes has been overplayed in technology policy and that ideological conflicts and the pace of innovation are not barriers to doing more impactful and valuable philosophical work in this sector. In fact, the current technological climate, the strong resistance to surveillance capitalism, the passing of new data privacy laws, the complicated relationship between big tech, big oil and climate justice, tech employee movements and whistleblowing, all suggest that something within technology is changing, and that it is time we adopt new tools and modes of thinking around technological justice. What the technology ecosystem is in greatest need of today, in fact, seems to be a slower, richer, more comprehensive investigation of what various technology companies and stakeholders owe to humans, to animals and to the planet. New technologies are also making us reinvestigate and question the commitments we humans owe to each other, as well as to other beings and to the global planet ecosystem. This is precisely what moral philosophy is for. We may want to stop bashing it.

7 Conclusion

This paper has argued that ethics washing and ethics bashing are both narrow approaches that rely on a limited understanding of what ethics actually entails. Ethical reasoning or moral inquiry can have intrinsic value as a process and instrumental value as a means to the achievement of other valuable outcomes. It has been argued that the more ethics is used in tech circles as a performative façade, the more it is instrumentalized and voided of its intrinsic value or role as a means to valuable ends such as enhanced understanding, the less value it can have overall as a practice and mode of inquiry. Adopting a perspective internal to moral philosophy helps us see the limits and actual similarities of what seem like polar opposites - ethics washing and ethics bashing – as two instances of instrumentalized ethics language.

The way to combat ethics washing, therefore, is not to instrumentalize ethical language, reduce and then dispose of it, but rather to distinguish performative and instrumentalized forms of ethics from valuable commitments to moral principle that promote advancements in self-knowledge and understanding. Although philosophers might have some work to do to attune their methods and discipline to fast-paced and politicized environments such as the tech space, we cannot disregard the immense depth and richness that philosophy can bring to any debate, not least technology governance ones.

It is hoped that technology scholars and policy-makers will embrace philosophy and be willing to dig deeper into its porous, principled and open-ended richness. It is also hoped that more moral philosophers will take on the difficult task of rethinking how new technologies interact with humans so as to provide answers to questions in urgent need of theorization. We all ask moral questions as part of our daily pursuits. To avoid falling into reductive epistemic and ideological traps, it is everyone's duty to nourish curiosity for ethics and moral philosophy's role in their personal and professional lives.

ACKNOWLEDGMENTS

I thank Ben Green, Lily Hu, Luke Stark, Reuben Binns, Yochai Benkler, Jonathan Zittrain, Lucas Stanczyk, Jeff Behrends, Brian Berkey, Mark Budolfson and three anonymous reviewers for their valuable input on this paper.

REFERENCES

[1] Kent Walker, *An external advisory council to help advance the responsible development of AI*, GOOGLE BLOG, May 26, 2019, https://blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/.
[2] Googlers Against Transphobia and Hate, *Google must remove Kay Coles James from its Advanced Technology External Advisory Council (ATEAC)*, MEDIUM, April 1st, 2019, https://medium.com/@against.transphobia/googlers-against-transphobia-and-hate-b1b0a5dbf76.
[3] Sam Levin, *Google scraps AI ethics council after backlash: "Back to the drawing board"*, THE GUARDIAN, April 4th, 2019, https://www.theguardian.com/technology/2019/apr/04/google-ai-ethics-council-backlash.
[4] Ben Wagner, *Ethics as an Escape from Regulation: From ethics-washing to ethics shopping?* in MIREILLE HILDEBRANDT (ED.), BEING PROFILING. COGITAS ERGO SUM (2018).
[5] See, e.g., the work of Tristan Harris at Google, his website is here: http://www.tristanharris.com/.
[6] Julia Powles and Helen Nissenbaum, *The Seductive Diversion of 'Solving' Bias in Artificial Intelligence*, MEDIUM, December 7, 2018, https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.
[7] See e.g. Gabriel Moran, *Ethics or Morality?* http://www.nyu.edu/classes/gmoran/Ethics%20or%20Morality.pdf.
[8] Dagobert D. Runes (ed.) Dictionary of Philosophy, (1983) at 338.
[9] *See* in particular RONALD DWORKIN, JUSTICE FOR HEDGEHOGS (2011).
[10] *See* RONALD DWORKIN, LAW'S EMPIRE (1986). More specifically, in JUSTICE FOR HEDGEHOGS (2011), Dworkin distinguishes two senses of ethics and two senses of morality. Morality can be divided into political and personal morality, ethics into narrower and wider ethics. Political morality applies to institutions and the state, and in some cases can be equated to political justice. (Note that Dworkin distinguishes political morality from the Rawlsian notion of justice, in JUSTICE FOR HEDGEHOGS (2011), Part V.) The other morality, personal morality, and ethics understood widely, are about how we ought to behave toward other people, what we owe to each other. Finally, the narrower understanding of ethics, purely personal ethics, is about how to live one's life well. For Dworkin all these notions are related, but these distinctions can clarify some of our uses of the language of ethics and morality.
[11] JOHN RAWLS, A THEORY OF JUSTICE (1971).
[12] Examples of such scholarship are cited in a paper by Jeffrey Behrends and John Basl on the relevance and flaws of "trolleology" for answering ethical questions about how to program autonomous vehicles.
[13] This is something that philosopher Helen Nissenbaum has done very well in her book PRIVACY IN CONTEXT (2009).
[14] See e.g. SUE PRIDEAUX, I AM DYNAMITE! A LIFE OF NIETZSCHE (2018); Charlotte Baumann, *Was Hegel an Authoritarian Thinker? Reading Hegel's Philosophy of History on the Basis of his Metaphysics*, ARCHIV FÜR GESCHICHTE DER PHILOSOPHIE (2019).
[15] *See* e.g. Juergen Habermas' famous critique of Rawls' original position as being far from uncontroversial in Jurgen Habermas, *Reconciliation Through the Public use of Reason: Remarks on John Rawls's Political Liberalism*, 92 THE JOURNAL OF PHILOSOPHY 109 (1995). In relation to technology *see* Anna Lauren Hoffman, *Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse*, 22 INFORMATION COMMUNICATION AND SOCIETY 900 (2019).
[16] KATRINA FORRESTER, IN THE SHADOW OF JUSTICE (2019). In this book, Forrester discusses the Anglo-American trend in moral and political philosophy to follow Rawls on questions of political philosophy but also ethics.
[17] *See* e.g. CAROLE PATEMAN AND CHARLES MILLS, CONTRACT AND DOMINATION (2007); Anna Lauren Hoffman, *Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse*, 22 INFORMATION COMMUNICATION AND SOCIETY 900 (2019).
[18] Examples of theorists who have developed understandings of philosophy based on dialectics and the aspirational faith in dialogue and rational argument include Frederik Hegel, Jurgen Habermas, John Rawls and Thomas Scanlon.
[19] *See* ROGER FISHER & WILLIAM URY, GETTING TO YES (1981).
[20] *See* e.g. Brian Resnick, *Most people are bad at arguing. These 2 techniques will make you better*, VOX, Nov. 26, 2019, https://perma.cc/SK98-FCH7.
[21] Nancy Fraser, *Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy*, 25 SOCIAL TEXT 56 (1990).
[22] A wide range of eclectic approaches to the relationship between moral philosophy and structural injustice can be found in the literature. *See*, e.g., CAROLE PATEMAN AND CHARLES MILLS, CONTRACT AND DOMINATION (2007), and in particular Chapter 3, Charles Mills, *The Domination Contract*. Also *see* STEPHEN LUKES, POWER: A RADICAL VIEW (1974); GERALD A. COHEN, WHY NOT SOCIALISM? (2009); IRIS MARION YOUNG, RESPONSIBILITY FOR JUSTICE (2011); GINA SCHOUTEN, LIBERALISM, NEUTRALITY, AND THE GENDERED DIVISION OF LABOR (2019); Lucas Stanczyk's work on productive justice. On the intersection and tensions between moral and political philosophy, structural injustice and technology, *see* Mireille Hildebrandt, *Closure: on ethics, code and law*, in MIREILLE HILDEBRANDT, LAW FOR COMPUTER SCIENTISTS (2019); and Anna Lauren Hoffman, *Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse*, 22 INFORMATION COMMUNICATION AND SOCIETY 900 (2019).
[23] Tobias Rees, *Why tech companies need philosophers—and how I convinced Google to hire them*, QUARTZ, November 22, 2019, https://perma.cc/2967-8H5R.
[24] Evelyn Douek, *Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility*, 21 NORTH CAROLINA JOURNAL OF LAW AND TECHNOLOGY (forthcoming, 2019).
[25] Thomas Kadri and Kate Klonick, *Facebook v. Sullivan: Building Constitutional Law for Online Speech*, SOUTHERN CALIFORNIA LAW REVIEW (forthcoming, 2019).
[26] An example is Apple's philosopher in residence Joshua Cohen who does not seem to be allowed to make public appearances, or Microsoft's AI ethics oversight committee which doesn't disclose the reasons behind its decisions. *See* Alexis Papazolgou, *Silicon Valley's Secret Philosophers Should Share Their Work*, August 28, 2019, WIRED, https://perma.cc/6KZR-ASJ9.
[27] Oscar Williams, *How Big Tech funds the debate on AI ethics*, June 6, 2019, NEWSTATESMAN, https://perma.cc/5999-57BW.
[28] Safyia Noble argues in favor of a slower approach to media which involves humans in decision-making rather than relying on machines, SAFIYA U. NOBLE, ALGORITHMS OF OPPRESSION (2018).