Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning

Daniel Greene
College of Information Studies
University of Maryland
dgreene1@umd.edu

Anna Lauren Hoffmann
The Information School
University of Washington
alho@uw.edu

Luke Stark
Microsoft Research Montreal
luke.stark@microsoft.com

Abstract

This paper uses frame analysis to examine recent high-profile values statements endorsing ethical design for artificial intelligence and machine learning (AI/ML). Guided by insights from values in design and the sociology of business ethics, we uncover the grounding assumptions and terms of debate that make some conversations about ethical design possible while forestalling alternative visions. Vision statements for ethical AI/ML co-opt the language of some critics, folding them into a limited, technologically deterministic, expert-driven view of what ethical AI/ML means and how it might work.

1. Introduction

Spurred in part by advances in machine learning, algorithmic processes and predictive analytics are being applied to domains from criminal justice [1] to consumer finance [2]. In response, computer scientists, engineers, and designers—as well as executives, philosophers, social scientists, regulators, lawyers, and activists—have proposed guidelines for the responsible development, deployment, and regulation of artificial intelligence and machine learning systems (AI/ML). All part of a broader debate over how, where, and why these technologies are integrated into political, economic, and social structures. Still 'in the making' [3], the ethics of these systems are 'up for grabs.'

These debates present an opportunity to assess emergent approaches to incorporating ethics and values into AI/ML. In this paper, we examine high-profile "values statements" or manifestos [4] that endorse principles of ethical design as a response to social anxieties surrounding AI/ML. Because widely applied AI/ML and their attendant ethical debates are relatively new, we are interested in how values statements work to construct a shared ethical frame—a seemingly common-sense yet hegemonic understanding of an 'ethics' of AI/ML, how those ethics should be adjudicated, and whose voices count in the process [5].

We proceed in four stages. First, we situate a number of high-profile values statements in the broader context of recent academic work on ethical AI/ML. Second, we review our theoretical background: joining the literature in values in design with the sociology of business ethics. Third, we describe our methods and our sample. Finally, we present our findings, identifying seven core themes: universal concerns, objectively measured; expert oversight; values-driven determinism; design as locus of ethical scrutiny; better building; stakeholder-driven legitimacy; and machine translation. Combined, these themes inform what Gabriel Abend [6] terms the 'moral background' of these values statements: the grounding assumptions and terms of debate that make conversations around ethics and AI/ML possible in the first place.

We draw two broad conclusions. First, these statements offer a deterministic vision of AI/ML, the ethics of which are best addressed through technical and design expertise. There is little sense from these documents that AI/ML can be limited or constrained (a feature perhaps stemming from the involvement of AI companies). Second, the ethical design parameters suggested by these statements share some of the processual elements and contextual framing of critical methodologies in science and technology studies (STS) and information science. However, this critical scholarship's explicit focus on normative ends devoted to social justice or equitable human flourishing is often missing from these vision statements. The "moral background" of ethical AI/ML discussions is closer to conventional business ethics than more radical traditions of social and political justice active today, such as prison abolitionism or workplace democracy.

2. Context

From 2014 to 2016, President Obama and the White House Office of Science and Technology Policy identified "big data" as both a strategic priority and an area of legal and ethical concern. In a series of reports [7]-[9], the administration made a commitment to support big data's "enormous potential for positive

impact" while also "ensuring that it does not create unintended discriminatory consequences" [9]. Building on important germinal scholarship [10]-[12], these reports framed concerns over discrimination explicitly within the context of United States civil rights legislation, focusing on algorithmic systems and automated processes that "inform decisions that affect our lives, such as whether or not we qualify for credit or employment opportunities, or which financial, employment and housing advertisements we see" [9].

Subsequent scholarly work in AI/ML ethics has been more or less aligned with the vision outlined in a series of reports by the Obama White House. Popular and academic texts [13], [14] and news outlets [15] have confronted issues of algorithmic bias. Computer scientists have sought to develop computational solutions for problems of discrimination [16], [17], justifying this work by reference to fairness and equality/equity [16]-[19]. Research communities have also begun to develop preliminary codes or principles to guide AI/ML development, while major Silicon Valley companies have committed time and attention to AI/ML's ethical challenges [20]-[22].

Though many of these efforts share common origins and aims, tensions have emerged. For example, debates centered on racial representation within AI/ML systems have surfaced unresolved issues around the meaning and scope of inclusion. Spurred in part by research demonstrating facial recognition technology's difficulty recognizing people of color (especially black people [23]), companies like Microsoft began touting "inclusivity" efforts aimed at improving facial recognition's performance across skin tones [24]. Others, however, were skeptical. Technologist and organizer Nabil Hassein questioned "whose interests would truly be served by the deployment of automated systems capable of reliably identifying Black people" [25]. Or, as sociologist Alondra Nelson noted, these efforts risked "confusing 'inclusion' in more 'diverse' surveillance systems with justice and equality." [26] Where companies like Microsoft understood 'inclusion' as a fix to faulty technical systems, others—like Hassein and Nelson—saw it as a threat to black communities historically targeted by surveillance technology.

Other tensions speak to the fractured relationships between different research communities. Computer scientist Cathy O'Neil, for instance, publicly bemoans a lack of academic attention to fairness, accountability and transparency in algorithmic systems—and especially a perceived lack of academic efforts to inform policymakers and regulators [27]. Against this framing, the seven members of the Pervasive Data Ethics for Computational Research (PERVADE) group argue O'Neil mischaracterized the problem: academics in information science, computer science, law, sociology, and STS have been doing this work for some time, but it does not easily translate into policy because it is underfunded, marginalized, and at odds with a US political apparatus generally favorable towards Silicon Valley [28]. Where O'Neil thinks more research needs to be brought to the table, the PERVADE team argued that the structure of the table is itself the problem.

Despite—or perhaps because of—these unresolved tensions, many high-profile companies, organizations, and communities have seized on public conversations as an opportunity to signal their commitment to ethics. Many of these efforts have involved the drafting of "values statements" articulating more or less robust visions for the ethical or responsible development of AI/ML. Such statements prompt more questions than answers. By setting the tone for conversations around ethics and AI/ML, these statements simultaneously erase existing tensions while producing new conflicts. They stake a claim for the territory of ethical AI/ML, define how we should debate them, and suggest how we should put them into practice—smoothing out otherwise fraught ethical terrain.

3. Theoretical background

Our critical evaluation of high-profile values statements for ethical AI/ML explores the assumptions and the terms of debate making such statements possible in the first place—what Abend [6] terms its "moral background". In order to surface these assumptions and terms, our analysis draws on two distinct areas of scholarship: 1) interdisciplinary research in values in technology design and 2) the sociology of business and professional ethics.

3.1 Values in technology design

Scholars have long paid attention to the relationship between human values and technology design (hereafter "VID" or "values in design"). These include Value-Sensitive Design (VSD) [29], Values @ Play [30], reflective design [31], adversarial design [32], and critical technical practice [33]. Such scholarship centers the political decisions within and political outcomes of technological design [34], [35]. In the 1990s, information scientists also began focusing on the values at work in the design of computer systems, building from and supplementing longstanding work in computing ethics [11], [36], [37]. These efforts have been supplemented by critical work examining how design supports racialized,

gendered, and colonialized hierarchies [38]-[40]. Ethnographic observations of the design processes have helped identify what Katie Shilton terms 'values levers': the collaborative and organizational processes that transform ethical values from an implicit element of engineering activity to an explicit matter of reflection and debate [41]. More recent work by Shilton [42], [43] has documented the methodological opportunities and challenges in applying "anticipatory ethics" to the design of emerging digital technologies, including AI/ML.

VSD, Values@Play, reflective design, and critical technical practice are particular mechanisms through which the analytical insights of VID can be realized in design. Informed by moral philosophy and provoked by encounters with biased or broken systems, Friedman [44] and Nissenbaum [45] pioneered the reflective, iterative process of building systems for human-computer interaction that prioritize trust and user welfare while endeavoring to reduce biased outcomes.

Identifying values is, of course, only the beginning of a principled design inquiry, a first step that opens the door to consideration of alternative courses of action and their potential outcomes [46]. In this spirit, efforts in feminist [47] and postcolonial HCI [48] not only identify patriarchal or Western values embedded in design processes, but also reroute those processes and upend our ideas of who and what counts in design.

3.2 Business and professional ethics

Recognizing values in design is one thing; tracing their origins is another. In addition to questions of values, there is a question of how specific modes of moral reasoning become embedded in specific ways of designing technologies. These are epistemological questions about different value systems and ontological questions of what ethics are and how they work. With assistance from the sociology of ethics, they can also become empirical questions for information science.

The study of practical ethics as applied in businesses and professions is a long one [49], [50]. Sociological work on these applied ethics emphasizes how ethics and ethical codes designate and defend social status and expertise more than enforce consistent moral or societal virtues [51], [52]. Abend [6] argues the history of business ethics is cyclical and not particularly interesting: Stock market crashes or management scandals occur to great public outrage, heads roll, companies fold, governments investigate, business ethicists build institutions to improve behavior and preach the gospel that good morals are good business, keeping society and markets intact until the next crisis.

Abend's insights resonate with the recent history of backlash against major Silicon Valley companies where cycles of malfeasance and apology from firms like Facebook are routine. The claims of business ethicists are familiar here—for example, that unethical behavior leads to business disasters. But popularizing and institutionalizing these claims, which are far from the only way of reasoning about business ethics, require what Abend calls a *moral background*—a specific arrangement of second-order social assumptions about what ethics mean and how they work, above first-order claims about ethical norms or behaviors. This is where the real social action is for Abend, and it is where our research focuses.

4. Method

The values attached to AI/ML work are still taking shape [3]. The sociology of business ethics helps us define that shape, where VID helps us see it applied by designers and in designs. High-profile values statements in AI/ML are particularly important to the formation of a "moral background," as they make the connection between values, ethics, and technologies explicit. Yet given their high visibility and influence, such statements also *shape* and *set* these connections. Adapting van Leeuwen's insights from discourse analysis [53], we suggest these statements represent the transformation of ethics and design into *discourses about* ethics and design. And as with discourse broadly, these statements legitimate (and delegitimize) certain practices, providing "answers to the spoken or unspoken questions 'Why should we do this' or 'Why should we do this in this way?'"

4.1 Frame analysis

Frame analysis is ideally suited to tracing the implicit terms of this debate. Developed as a method in communications research, frame analysis investigates messages not just for their denotative content but also for processes whereby certain elements are selected for salience or erased [54]. It is closely related to critical discourse analysis, which examines how social actors recontextualize practices through texts, giving them legitimacy and opening up or reinforcing power differentials [55]. Our research draws on the sociological branch of frame analysis, which focuses on how certain political issues become associated with a common-sense set of problems and solutions; they investigate the construction, maintenance, reception, and circulation of powerful

frames [56], [57]. Here, we focus on how AI/ML are framed as ethical problems and the reasoning that defines these problems and their solutions.

We analyzed seven significant public statements meant to guide the development, implementation, and regulation of AI and ML. Within them, we looked for:

- common themes about the construction of ethical claims and their grounding assumptions;
- the logic undergirding these themes; and
- divergences among the sample (e.g., places where first-order norms, or second-order assumptions differed).

Using inductive coding [53, 55], the lead author annotated the statements, progressively developing and refining a set of core themes, and coding the technologies, domains, and actors to which they applied. For instance, a passage highlighting academic credentials was tagged with the 'expertise' code. This code eventually encompassed decision-making or analytical processes that implicitly or explicitly required expert knowledge, and birthed several sub-codes used to distinguish sub-categories of expertise (e.g., 'technical' or 'legal'). Informed by prior discourse analyses of digital researchers [58] and technologists [59], the other authors collated and reviewed a set of shared themes and noted divergences, with reference to the sample and the prior stage's codes. Some divergences prompted a return to first-order coding, to ensure similar phenomena were coded similarly. Other divergences within a code set identified surprising patterns that informed our analysis of an emergent theme (e.g., noting that both legal and scientific knowledge were invoked as forms of expertise relevant to ethical AI/ML).

4.2 Data

Our sample focused on recent public statements of ethical principles issued by independent institutions. Most of these organizations are made up of technologists and firms active in the field of AI/ML. These "envisioning bodies" were usually convened for the express purpose of circulating their proposed principles and conducting further research, convenings, public education, and lobbying to support their application and dissemination.

The oldest vision statement we examined was from December 2015 (OpenAI); the newest was from May 2018 (The Toronto Declaration). Though, as a practical matter, our sample ends in May 2018, additional high-profile statements have already been posted (including, notably, Google's AI principles). That these statements continue to emerge is a testament to their central role in debates over ethical AI/ML.

Our sample includes the following institutions:

The Partnership on AI to Benefit People and Society
Membership: Nonprofit cooperative effort between Amazon, Apple, DeepMind, Google, Facebook, IBM and Microsoft, with second-tier partners from higher education, civil rights groups, and other industry partners.

The Montreal Declaration for a Responsible Development of Artificial Intelligence
Membership: Interdisciplinary team of academics and interested industry practitioners based at Montreal universities and the non-profit MILA (Montreal Institute for Learning Algorithms). Montreal is a noted hub for AI research [60].

The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems
Membership: Drafted by representatives of human rights nonprofits such as Amnesty International and Access Now, as well as machine learning ethicist Solon Barocas (also a convener of FATML).

OpenAI
Membership: Nonprofit AI research company working towards safe Artificial General Intelligence, i.e., an independently thinking machine 'mind'; sponsored by venture capitalists (e.g., Reid Hoffman, Elon Musk, Peter Thiel) and corporations (Microsoft, Amazon).

The Center for Humane Technology (CHT)
Membership: Founded by Tristan Harris (formerly Google) and Aza Ruskin (formerly Mozilla, Jawbone, Songza). Described as "concerned former tech insiders [who] understand the culture, business incentives, design techniques, and organizational structures driving how technology hijacks our minds."

Fairness, Accountability and Transparency in Machine Learning (FATML)
Membership: An interdisciplinary convening of computer scientists, statisticians, and ethicists prompted by President Obama's 2014 call for a 90-day review of the Big Data: Seizing Opportunities and Preserving Values report (now an annual conference attracting academic, industry, and government participants).

Axon's AI Ethics Board for Public Safety
Membership: Previously Taser, the largest US provider of non-lethal police weaponry. Their name

change reflects a pivot into police body cameras, data storage, and automated evaluation of video. After some criticism, Axon created an ethics board composed of police officials and AI researchers.

5. Findings

Digital technologies are legitimate objects of ethical concern—and human values are embedded in their design. This argument may seem obvious to VID scholars, but it is no small concession in the face of powerful discourses of technological neutrality pervasive in Silicon Valley and elsewhere [61]. When Karl Popper gave a lecture to the International Congress of Philosophy in Vienna in 1968, he noted everyone in the room was perfectly clear, before he even began, that his title, "The Moral Responsibility of the Scientist," was a "euphemism for the issue of biological and nuclear warfare" [62] The gravity of these technologies helped focus the minds of attendees. Such a shared consensus on the moral responsibility of computer engineers and data scientists towards their own inventions does not yet hold today; instead, a burgeoning movement we call 'ethical design' is attempting to foster such a consensus in AI/ML—or at least formalize the terms of debate.

However, the fact the Montreal Declaration can assume its interlocutors will accept the argument developers should be creating "moral machines" signals an emerging acceptance of design as a legitimate site for ethical debate, rather than something that can be delegated to other domains (e.g., law). This initial assumption grounds the seven other core themes that join to form the moral background of 'ethical design': universal concerns, objectively measured; expert oversight; values-driven determinism; design as locus of ethical scrutiny; better building; stakeholder-driven legitimacy; and machine translation.

5.1 Universal concerns, objectively measured

The precise reasons why AI/ML are matters of ethical concern differ from organization to organization. Some lean on the language of distributive justice, arguing AI/ML's benefits and penalties will be unevenly distributed. For example, the Toronto Declaration argues that 'marginalized groups' will feel the brunt of discriminatory ML and so should be explicitly included in the development process. However, all the statements agree a) the positive and negative impacts of AI are a matter of universal concern, b) there is a shared language of ethical concern across the species, and c) those concerns can be addressed by objectively measuring those impacts. This is a universalist project that brooks little relativist interpretation.

Often this ethical universalism is justified by reference to a hazy biological essentialism. CHT's core argument is that social media is hacking our attention, and that human brains cannot adequately cope with these addictive designs. Beneath a picture of a cartoon brain swarming with red notification icons on their website is the sentence "There's an invisible problem that's affecting all of society." The Montreal Declaration follows a similar line: The community of concern is "all human beings" or even "all sentient creatures." The Toronto Declaration departs from this conception of biological community, pursuing instead a legal universalism via human rights law as the grounds on which harms and remedies are understood.

In order to address these shared concerns, ethical design advances a program of objective measurement of harms. FATML, for example, aims to support research addressing bias and discrimination through "computationally rigorous methods." The third core value in its Principles for Accountable Algorithms is 'accuracy', which is meant to encourage the detailed logging of errors and uncertainty. Similarly, the Toronto Declaration endorses a program of impact assessment throughout the ML lifecycle.

5.2 Expert oversight

Despite assuming a universal community of ethical concern, these vision statements are not mass mobilization documents. Rather, they frame ethical design as a project of expert oversight, wherein primarily technical, and secondarily legal, experts come together to articulate concerns and implement primarily technical, and secondarily legal solutions. They draw a narrow circle of who can or should adjudicate ethical concerns around AI/ML.

This assumption of expertise is clear from both the voices within these documents (e.g., ranging from major AI corporations to leading academics and legal minds) and from the substance of their proposals. The Partnership demarcates "the public", a body to be educated and surveyed, from "stakeholders", scientists, engineers and businesspersons who will educate and survey. Elsewhere, in describing Our Work, The Partnership separates out "Engagement of Experts" from "Engagement of Other Stakeholders". The former are leaders of scientific disciplines addressing or building AI, the latter range from individual product users to large corporations purchasing AI solutions or disrupted by AI in their

sector. Experts make AI happen, Other Stakeholders have AI happen to them.

Less frequently, statements acknowledge the importance of non-technical expertise. The Toronto Declaration's Preamble positions "the universal, binding and actionable body of human rights law and standards" as an invaluable supplement to technical debates. It argues not just for regulators to become experts in ML, but for ML procurers and developers to become experts in human rights and international due process standards—or at least employ such experts. Similarly, Axon's Ethics Board includes not only roboticists and computer vision experts but privacy researchers, former and current police officers, and criminologists. Over time, FATML's calls-for-papers embrace a broader community of experts. In 2014 and 2015 the "machine learning community" was the explicit audience. Later, this assumption broadens somewhat: the 2016 call has "researchers and practitioners" responding to the concerns of "policymakers, regulators, and advocates", while 2017's encourages submissions from "practitioners in industry, government, and civil society."

5.3 Values-driven determinism

The envisioning documents all offer deterministic framings of AI/ML as world-historical forces of change—inevitable seismic shifts to which humans can only react. Paradoxically, AI/ML are also at the same time described as values-driven, insofar as human beings create them. They are forces to which we must adapt and for which we are also responsible.

The Montreal Declaration captures this tension well. While there is overriding hope that "AI will make our societies better", sections exploring individual values such as Justice veer between instrumental impact (e.g., "What types of discrimination could AI create or exacerbate?") and active human agency (e.g., "Should the development of AI be neutral or should it seek to reduce social and economic inequalities?"). Similarly, OpenAI's Charter is aimed at the medium-term impact of inevitable "highly autonomous systems that outperform humans at most economically valuable work", by collaborating on "value-aligned, safety-conscious project[s]" in the present and near-term. This tension between ethical conflict in design and instrumentalism in impact is perhaps resolved by reference to the expert oversight described above: Human agency is integral to ethical design, but it is largely a property of experts responsible for the design, implementation and, sometimes, oversight of AI/ML.

In other places, this determinism manifests as teleology. It is taken as given that AI/ML technologies a) are coming and b) they will replace a broad swathe of human jobs and decisions. In its thematic pillar "AI, Labor, and the Economy", The Partnership on AI assures readers that AI will "undoubtedly" disrupt the labor market, "as new kinds of work are created and other types of work become less needed due to automation." Consequently, ethical debate is largely limited to appropriate design and implementation—not whether these systems should be built in the first place.

Crucially, edicts to do something new are framed as moral imperatives, while the possibility of *not* doing something is only a suggestion, if mentioned at all. This is true even for the more critical statements like The Toronto Declaration, which stresses that some groups deserve extra care when collecting and processing their data and that such care extends throughout a product's lifecycle. Here, attention is paid to the risks of harm from design through execution—yet it is still taken as a given that these data will be collected.

5.4 Design as locus of ethical scrutiny

Following from the previous theme, business practices which might affect AI/ML design and use (and which tend to overpower individual ethical concerns) remain a lacuna. Business decisions are never positioned as needing the same level of scrutiny as design decisions. In this way, the vision statements are reminiscent of many professional codes of ethics, which often detail the responsibilities of individual professionals without actively scrutinizing the nature of the profession or business in question [50], [63]. This is particularly significant as the cutting edge of the field is (due to the enormous amount of data and fixed capital necessary to train AI, store data, and implement code) largely the province of those large corporations funding The Partnership—who also have a habit of acquiring smaller upstarts.

The Montreal Declaration's "Justice" plank nods towards the problem of "the concentration of power and wealth in the hands of a small number of AI companies" but the principle that follows from this question returns to a focus on developing AI that promotes justice and reduces discrimination. The CHT appears on the surface to take a strong ethical stance against the business model of attention hacking, but the proposals that flow from its ethical stance are largely limited to design considerations—many of which have already have already been embraced by the industry verbatim (e.g., Facebook's emergent focus on 'time well spent' rather than raw engagement time, and Google's release of "digital wellness" tools in its new version Android [64], [65]). While, empirically,

we do not have access to the convening decisions of these bodies, it is fair to speculate that embracing (largely technical and legal) expertise and rejecting critiques of business practice leads to the exclusion of critical researchers from these bodies, especially those who highlight how profit motives and institutional racism corrupt what could be considered public information resources [66], [67].

5.5 Better building

An important consequence of business practices being discursively "off the table" is the implication that "better building" is the only ethical path forward. The overarching focus of the Montreal Declaration is the creation of 'responsible' AI, and while its questions ask stakeholders to consider whether autonomous agents should be able to run an abattoir or kill an animal, the proposed principles reframe the debate on the affirmative grounds of designing AI to fulfill social goods (i.e., eliminating discrimination, protecting humans from propaganda, etc.). Rhetorically, the corporate members of the Partnership commit themselves to "better building" by seeking to maximize the benefits of AI, minimize their disruptions and negative impacts, and educate the broader public on the role of AI in their lives. The only red line drawn by the Partnership is in its Tenets, which commit to "opposing the development and use of AI technologies that would violate international conventions or human rights."

Axon's Ethics Board announcement similarly frames their initiative as the responsible shepherding of innovations destined to improve policing. CEO Rick Smith says the Ethics Board was created "to ensure any AI technology in public safety is developed responsibly." In a comment to the *Washington Post,* Smith echoed the values-driven determinism described earlier: "It would be both naive and counterproductive to say law enforcement shouldn't have these new technologies" [68]. He clarified that the Ethics Board has no veto power over Axon's plans.

It is clear that "better building" is the only way forward because no statement offers "not building" as an alternative. Across the statements, the Toronto Declaration contained the only gesture toward not building, but ultimately demurs: "Where the risk of discrimination or other rights violations has been assessed to be too high or impossible to mitigate the private sector should consider not deploying a machine learning application." There are no other red lines which should not be crossed, for state or corporate actors.

5.6 Stakeholder-driven legitimacy

Proponents of ethical design often articulate a desire to open or sustain conversations by engaging as many stakeholders—largely experts—as possible. This positions ethical design *as ethical*, in part, because it is given a thorough vetting. Vetting legitimates decisions through an appeal to transparency, but without specifying any subsequent substantive commitments.

This legitimacy appears to hold even if, as in Axon's case, consulted stakeholders are limited in their capacity to impact design. Indeed, it appears part of the mission of Axon's Ethics Board is to simply release reports that "demonstrate a commitment to public transparency." The Partnership's explicit mission is to bring disparate others into discussions by extending the discussion out those disparate others. They break down "How We're Doing It" into four prongs: engaging domain experts "to discuss and provide guidance"; hearing the concerns of non-expert stakeholders in industries affected by AI and bringing those concerns back into research and development; producing third-party studies, supporting moonshot ideas, and "the identification and celebration of important work"; and developing "informational materials" for the broader public.

FATML and the Toronto Declaration evidence a similar commitment. Both recognize that a conversation among technical experts is ongoing, and other (expert) voices need to be brought to the table. FATML seems to have come to this conclusion over time, as evidenced in its CFPs, while the Declaration's Preamble highlights the importance of bringing human rights literacy to a conversation heretofore dominated by engineers. Decisions made about AI/ML need to be made with a wide community of experts, and the wide-ranging impact of AI/ML demands a wide-ranging group of experts to research those impacts.

5.7 Machine Translation

The broad circle of (expert) consultation is also extended to AI and ML technologies themselves. Vision statements often position 'explicable' and 'transparent' (as opposed to "black-boxed") systems as both a foundation of moral AI/ML and a means by which moral questions are pursued. Under the Montreal Declaration's value of Knowledge, there are questions posed not just about what AI means for human knowledge (e.g., "Does the development of AI put critical thinking at risk?") but what knowledge humans should have about AI (e.g., "Is it acceptable not to be informed that medical or legal advice has been given by a chatbot?"). This dual emphasis makes clear that the proposed principle that follows—"The

development of AI should promote critical thinking and protect us from propaganda and manipulation"—is a two-way street: Moral machines not only shield us from fake news, they make their inner workings clear enough to ensure no fake news lies within.

FATML further grounds this background element in technical specifics with its Principles for Accountable Algorithms, originally developed at a Dagstuhl Seminar entitled "Data, Responsibly", with an explicit audience of "developers and product managers." The explicit goal is to make AI and ML "publicly accountable" via "an obligation to report, explain or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms." This is grounded in five core values, including explainability and auditability, each linked with steps to take and questions to take in building social impact statements for algorithms. Developers should have a plan to explain algorithmic decisions and should "consider whether a directly interpretable or explainable model can be used."

6. Summary

We have identified seven core elements of ethical design's moral background: Universal concerns, objectively measured; expert oversight; values-driven determinism; design as locus of ethical scrutiny; better building; stakeholder-driven legitimacy; and machine translation. What unites them? Two underlying themes stand out.

First, conversations around the design and deployment of ethical AI/ML are taking place among experts well aware they are under public scrutiny—the prospect of massive job losses and rigged elections has raised the public profile of their work. Building a moral background for ethical design is partly about shaping public perception, providing the concepts through which AI/ML can be understood. One goal for these envisioning statements is thus to generate the moral consensus Popper knew already existed within the scientific community on nuclear and biological weapons: acknowledgment of a specific set of threats, and a specific set of people, tools, and ideas ready to respond. Yet the problems remain, in this view, fundamentally technical, shielded from democratic intervention. Other forms of expertise appear in these statements, but the problems themselves are to be solved by experts in the technical features of AI/ML systems.

Second, and perhaps to the surprise of critical researchers engaged in this work for decades, ethical design seems to share many conceptual similarities with Values@Play, Values Sensitive Design, and neighboring fields. FATML's Principle for Accountable Algorithms in particular makes it clear that, "Algorithms and the data that drive them are designed and created by people—There is always a human ultimately responsible for decisions made or informed by an algorithm. 'The algorithm did it' is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes." This language, common in these values statements, paints a clear picture of moral causation: Poor ethics lead to bad designs, which produce harmful outcomes. This is far from an obvious, let alone the only possible, conclusion—but it is, nonetheless, the causation narrative on offer here.

7. Conclusion

Our overview and analysis is indicative of the broader debates surrounding the ethical development of AI/ML. High-profile values statements are powerful instruments for constructing and imposing a shared ethical frame on a contentious conversation. As our analysis shows, however, this frame is not an innocuous one; rather, it sets and shapes the 'moral background' that make conversations around ethics and technology possible in the first place. Specifically, it offers a deterministic vision of AI/ML, the ethics of which are best addressed through certain kinds of technical and design expertise.

They also take for granted the non-obvious assumption that poor ethics and bad designs produce harmful outcomes. Other causation narratives about the chaos of new, intelligent tools interacting in the wild, or large corporations dominating political processes with no democratic accountability, are equally plausible. There is a warning here for critical scholars: ethical design possesses some of the same processual elements as VID methodologies but lacks their often explicit focus on normative ends devoted to social justice or equitable human flourishing. This presents a new problem for sociotechnical scholars used to being ignored: What if, instead of being brushed aside, our critiques are being heard but transformed into something we might not recognize?

This warning, however, suggests a corresponding opportunity for the development of competing frames and alternative movements for progressive technological reform [69]. Our own scholarship seeks to advance these goals in two directions. First, with forthcoming historical work that compares current attempts to formulate professional codes of ethics with the successes and failures of other professions' codes [70]. And second, by building on the present research

to compare the movement for ethical AI/ML with parallel projects of technological reform: The workplace democracy of the #TechWontBuildIt campaign that targets companies building software for war or immigrant detention, and the abolitionist work of the Movement for Black Lives that seeks to correct the harms of police surveillance and return the resources spent on it to the community. This research is aimed at encouraging a broader ethical conversation focused explicitly on social justice and AI/ML – a conversation urgently needed as these technologies become increasingly ubiquitous in our everyday lives.

8. References

[1] E. Joh, "The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing," *Harvard Law and Policy Review* vol. 10, no. 15, pp. 15-42, 2016.

[2] D.K. Citron and F. Pasquale, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* vol 89, pp 1-33, 2014.

[3] D.G. Johnson, "Ethics and Technology 'in the Making': An Essay on the Challenge of Nanoethics." *Nanoethics,* vol. 1, no. 1, pp 21-30, 2007.

[4] C. Noessel, "Untold AI" https://scifiinterfaces.com/2018/06/12/untold-ai-2/ 12-June-2018.

[5] J. Matthes, "Framing Politics: An Integrative Approach," *Amer. Behav. Sci.*, vol. 56, no. 3, pp 247-259, 2012.

[6] G. Abend, *The Moral Background: An Inquiry Into the History of Business Ethics*. Princeton University Press, 2014.

[7] J. Podesta, P. Pritzker, E. J. Moniz, J. Holdren, and J. Zientz, "Big Data: Seizing Opportunities, Preserving Values," Executive Office of the President, Washington, D.C., May 2014.

[8] President's Council of Advisors on Science and Technology, "Big Data and Privacy: A Technological Perspective," Executive Office of the President, Washington, D.C., May 2014.

[9] C. Muñoz, M. Smith, and D. J. Patil. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights," Executive Office of the President, Washington, D.C., May 2016.

[10] A. Narayanan and S. Vallor, "Why Software Engineering Courses Should Include Ethics Coverage," *Communications of the ACM* vol. 57, no. 3, pp 23-25, 2014.

[11] B. Friedman and H. Nissenbaum, "Bias in Computer Systems," *ACM Transactions on Information Systems* vol. 14, no. 3, pp. 330-347, 1996

[12] L. Sweeney, "Discrimination in Online Ad Delivery," *Queue*, vol. 11, no. 3, pp. 10:10–10:29, Mar. 2013.

[13] C. O'Neil, *Weapons of Math Destruction*. Broadway Books, New York, 2017.

[14] F. Pasquale, *The Black Box Society*. Harvard University Press, Cambridge, MA, 2015.

[15] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, 23-May-2016.

[16] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A Comparative Study of Fairness-enhancing Interventions in Machine Learning," arXiv:1802.04422 [cs, stat], Feb. 2018.

[17] P. Adler et al., "Auditing Black-box Models for Indirect Influence," *Knowledge and Information Systems*, vol. 54, no. 1, pp. 95–122, Jan. 2018.

[18] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* vol. 81, 1-15, 2018.

[19] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," *Proceedings of Machine Learning Research* vol. 81, pp.1-11, 2018.

[20] M. Zook *et al.*, "Ten Simple Rules for Responsible Big Data Research," *PLOS Computational Biology*, vol. 13, no. 3, p. e1005399, Mar. 2017.

[21] T. Markoff, "How Tech Giants Are Designing Real Ethics for Artificial Intelligence." *The New York Times,* 1-Sept-16

[22] T. Simonite, "Artificial Intelligence Seeks an Ethical Conscience." *Wired,* 7-Dec-17

[23] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Proc. Of the Conf. on Fairness, Accountability and Transparency* (2018), pp. 77-91.

[24] J. Roach, "Microsoft Improves Facial Recognition Technology to Perform Well Across All Skin Tones, Genders." *Microsoft AI Blog*, 26-June-2018.

[25] N. Hassein "Against Black Inclusion in Facial Recognition." *The Digital Talking Drum,* 17-Aug-2017.

[26] A. Nelson (@alondra). "We must stop confusing 'inclusion' in more diverse surveillance systems with justice and equality." 26-June-2018, 10:41 PM. Tweet.

[27] C. O'Neill, "The Ivory Tower Can't Keep Ignoring Tech." *The New York Times,* 17-Nov-17

[28] K. Shilton, M. Zimmer, C. Fiesler, A. Narayanan, J. Metcalf, B. Mietz, J. Vitak, "We're Awake—But We're Not at The Wheel." *Medium,* 15-Nov-17.

[29] B. Friedman, P.H. Kahn, and A. Borning. "Value Sensitive Design and Information Systems," In B. Schneiderman, P. Zhang and D. Galletta, eds., *Human-Computer Interaction in Management Information Systems: Foundations*. M.E. Sharpe, Inc., New York, 2006, 348–372.

[30] M. Flanagan, D. C. Howe, and H. Nissenbaum, "Values at Play: Design Tradeoffs in Socially-oriented Game Design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2005, pp. 751–760.

[31] P. Sengers, K. Boehner, S. David, and J. Kaye, "Reflective Design," *Proc. of 4th Decennial Aarhus Conf. on Critical Computing* (2005), 49–58.

[32] C. DiSalvo, *Adversarial Design*. MIT Press, 2012.

[33] P.E. Agre, *Computation and Human Experience*. Cambridge University Press, Cambridge, UK, 1997.

[34] L. Winner, “Do Artifacts Have Politics?” *Daedalus* vol. 109, no. 1, pp. 121-136, 1980.
[35] W.E. Bijker, “Do Not Despair: There is Life After Constructivism.” *Science, Technology, and Human Values*, vol. 18, no. 1, pp. 113-138, 1993.
[36] L.D. Introna and H. Nissenbaum, “Shaping the Web: Why the Politics of Search Engines Matters.” *The information society*, vol. 16, no. 3, pp. 169-185, 2000.
[37] G. Bowker and S.L. Starr, *Sorting Things Out: Classification and its Consequences.* The MIT Press, 1999.
[38] R.N. Weber, “Manufacturing Gender in Military Cockpit Design.” In D. MacKenzie & J. Wajcman (Eds.), *The Social Shaping of Technology* (2 edition). McGraw Hill Education / Open University, 1999.
[39] R. Eglash, “Race, Sex and Nerds: From Black Geeks to Asian American Hipsters.” *Social Text* vol. 20, no. 2 pp. 49–64, 2002.
[40] L. Irani, J. Vertesi, P. Dourish, K. Philip, R.E. Grinter, “Postcolonial Computing: A Lens on Design and Development,” in *Proc. of the 28th int. conf. on Human factors in comp. sys.* (pp. 1311–1320). New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753522
[41] K. Shilton, “Values Levers: Building Ethics Into Design.” *Science, Technology & Human Values*, vol. 38., no. 3., pp. 374-397, 2013.
[42] K. Shilton, “‘That's Not an Architecture Problem!’: Techniques and Challenges for Practicing Anticipatory Technology Ethics*.”* In *iConference 2015 Proceedings*, 2015.
[43] K. Shilton, “Engaging Values Despite Neutrality.” *Science, Technology, & Human Values* 43, 2 (2018), 247–269.
[44] B. Friedman and P.H. Kahn, “Human Agency and Responsible Computing: Implications for Computer System Design*.” Journal of Systems and Software*, vol. 17, no. 1, pp. 7–14, 1992.
[45] H. Nissenbaum. “Accountability in a Computerized Society,” *Sci. and Eng. Ethics* 2, (1996), 25–42.
[46] N. JafariNaimi, L. Nathan, and I. Hargreaves, “Values as Hypotheses: Design, Inquiry, and the Service of Values.” *Design Issues,* vol. 31, no. 4, pp. 91-104, 2015.
[47] S. Bardzell, ”Feminist HCI: Taking Stock and Outlining an Agenda for Design.” *Proc. of the 28th int. conf. on Human factors in comp. syst*., pp. 1301-1310, 2010.
[48] H. Mainsah and A. Morrison, “Participatory Design Through a Cultural Lens: Insights from Postcolonial Theory.” *Proc. of the 13th Participatory Design Conf*, Vol. 2, pp. 83-86, 2014.
[49] L. Newton, “The Origin of Professionalism: Sociological Conclusions and Ethical Implications,” *Business Professional Ethics Journal* vol. 1, no. 4, pp 33-43, 1982.
[50] A. Abbott, “Professional Ethics,” *American Journal of Sociology* vol. 88, no. 5, pp 855-885, 1985.
[51] A. Gewirth, “Professional Ethics: The Separatist Thesis,” *Ethics* vol. 96, no. 2, pp 282-300, 1986.
[52] M.S. Frankel, “Professional Codes: Why, How, and with What Impact?” *Journal of Business Ethics* 8, 2/3 (1989), pp. 109–115.
[53] T. Van Leeuwen. *Discourse and Practice: New Tools for Critical Discourse Analysis*. Oxford University Press, 2008.
[54] R. Entman. “Framing: Toward Clarification of a Fractured Paradigm.” *Jour. of Comm*., vol 43, no. 4, pp. 51-58, 1993.
[55] N. Fairclough. *Critical Discourse Analysis: The Critical Study of Language*. Routledge, 2003.
[56] D. Greene. “Discovering the Divide: Technology and Poverty in the New Economy.” *Int. Jour. of Comm.*, vol. 10., pp. 1212-1231, 2016.
[57] K.M. Carragee and W. Roefs. (2004). “The Neglect of Power in Recent Framing Research.” *Jour. of Comm.,* vol. 54, no. 2, pp. 214-233, 2004.
[58] K. Shilton and S. Sayles, “‘We Aren't All Going to Be on the Same Page about Ethics’: Ethical Practices and Challenges in Research on Digital and Social Media.” *IEEE* pp 1909–1918, 2016.
[59] D. Greene and K. Shilton. “Platform Privacies: Governance, Collaboration, and the Different Meanings of “Privacy” in iOS and Android Development,” *New Media & Society* vol. 20, no. 4, pp 1640-1657, 2017.
[60] P. High. "Why Montreal has Emerged as an Artificial Intelligence Powerhouse." Forbes, 6-Nov-17
[61] M. Weigel, “Silicon Valley’s Sixty-Year Love Affair with the Word ‘Tool,’” *The New Yorker*, 12-Apr-2018.
[62] K. Popper. “The Moral Responsibility of the Scientist.” *Induction, Physics and Ethics*, pp. 329-336. Springer, 1972.
[63] N. Manders-Huits and M. Zimmer. “Values and Pragmatic Action: The Challenges of Introducing Ethical Intelligence in Technical Design Communities,” *International Review of Information Ethics*, pp 1-8, 2009.
[64] B. Tarnoff and M. Weigel. “Why Silicon Valley Can’t Fix Itself.” *The Guardian,* 3-May-2018.
[65] L. Stark. “Silicon Valley Wants to Improve Your ‘Digital Well-being’—And Collect More of Your Personal Data Along the Way.” *The Boston Globe*, 24-July-2018.
[66] V. Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin’s Press, 2018.
[67] S. Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.
[68] D. Harwell. “Facial Recognition May be Coming to a Police Body Camera Near You.” *The Washington Post,* 8-April-2018.
[69] S. Costanza-Chock. “Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice,” *Catalyst*, pp 1-14, 2018.
[70] L. Stark and A.L. Hoffmann. “Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Cultures. Forthcoming in *Journal of Cultural Analytics,* 2018.