CHAPTER 9

Critical Perspectives on Governance Mechanisms for AI/ML Systems

Luke Stark, Daniel Greene, and Anna Lauren Hoffmann

As the use of artificial intelligence (AI) systems grounded in various forms of machine learning (ML) and statistical inference has grown, the hype around these technologies has grown faster. AI/ML technologies will, according to their extollers, usher in a "Fourth Industrial Revolution," purportedly providing data-driven insights and data-driven efficiencies reshaping labor, medicine, urban life, and consumption (Schwab, 2017) is highest. Enthusiasm for AI/ML technologies is rampant in global corporations, national governments, and even nongovernmental organizations. Critiques of the adverse societal impacts of these systems have intensified

L. Stark (✉)
University of Western Ontario, London, ON, Canada
e-mail: cstark23@uwo.ca

D. Greene
University of Maryland, College Park, MD, USA
e-mail: dgreene1@umd.edu

A. L. Hoffmann
University of Washington, Seattle, WA, USA
e-mail: alho@uw.edu

J. Roberge and M. Castelle (eds.), *The Cultural Life of Machine Learning*, https://doi.org/10.1007/978-3-030-56286-1_9

in response to this upsurge in AI boosterism (Benjamin, 2019; Eubanks, 2018) but have at times struggled to gain traction with policymakers (Crawford et al., 2019).

In this chapter, we provide a critical overview of some of the proposed mechanisms for the ethical governance of contemporary AI systems. These strategies include technical solutions intended to mitigate bias or unfairness in the design of AI systems as well as legal, regulatory, and other social mechanisms intended to guide those systems as they are built and deployed. Academics and industry teams have developed technical tools for the development of fair, trustworthy, and interpretable AI systems; socio-legal governance mechanisms include projects from civil society groups, local, state and supranational governments, and industry actors. These latter solutions include high-level values statements and sets of principles around AI ethics, promulgated by actors in all three of the above categories; AI-specific laws and regulations from governments, alongside voluntary standards proposals from business and civil society groups; and the application of existing human rights frameworks and discourses of "securitization" to the governance of AI/ML technology.

Focusing on these interventions primarily in their North American and European contexts, we describe various proposed mechanisms for AI/ML governance in turn, arguing each category of intervention has in practice supported the broader regimes of corporate and state power under which AI/ML technologies are being developed. The various AI/ML governance mechanisms being proposed by states and corporate actors do not function independently of each other. As Nissenbaum (2011) observes, "law and technology both have the power to organize and impose order on society" (p. 1373). Technical and social governance mechanisms act together as sociotechnical systems, and understanding how these elements interact in the hands of state and corporate actors is critical in ensuring that the governance of AI/ML is not only most effective but also most just. The mutual interdependence of various material, regulatory, and rhetorical governance mechanisms can work together for less than ideal ends: to subvert one form of effective governance by undercutting it through other means, or to confound, confuse, and delay the exercise of oversight through emphasis on a different governance mechanism. The interrelationship between governance mechanisms can thus do as much to hinder as to help the causes of equality and justice. Here, we critique many of the proposed solutions for AI/ML governance as supporting a narrow, unjust, and undemocratic set of norms around

these technologies' design and deployment, grounded in the exigencies of both computational media and neoliberal capital (Deleuze, 1990). We conclude by highlighting alternative perspectives—including labor movements such as the Tech Won't Build It campaign and social justice groups such as the Movement for Black Lives—committed to dismantling and transforming both the AI/ML technologies supporting the broader twenty-first-century neoliberal surveillance economy, and that economy itself (Zuboff, 2019).

Governance Through Tools

Some of the most commonly proposed solutions to the shortcomings of contemporary AI systems have been technical fixes, i.e., changes to machine learning processes and practices to diagnose and diminish statistical biases or omissions in these systems' outputs (Narayanan, 2018). As "Big" data analysis and AI/ML have become ubiquitous, the long-standing critiques of social bias expressed via digital systems drawn from science and technology studies (STS) (Friedman & Nissenbaum, 1996) have gained traction in computer science and data science. Philosophers and social theorists of technology such as Nissenbaum (2010), Gandy (2009), Johnson (2007), Pfaffenberger (1992), and Winner (1988) have variously argued technical specifications can and should be only one element of a broader, multidisciplinary assessment of digital technologies and their social impacts. In 2013, Dwork (a computer scientist) and Mulligan (a legal scholar) argued for "greater attention to the values embedded and reflected in classifications, and the roles they play in shaping public and private life," observing digital analytics "promised—or threatened—to bring classification to an increasing range of human activity" (Dwork & Mulligan, 2013, p. 35). With the increasing ubiquity of AI/ML technologies, computer scientists began to follow the lead of Dwork, Friedman, Mulligan, Nissenbaum, and others in working to analyze, and in some cases formalize abstract values such as fairness, accountability, transparency, and interpretability, particularly in the context of machine learning systems. Yet while necessary for addressing the governance of AI systems, these tools address only a small fraction of the governance challenges provoked by these technologies.

Techniques for diminishing technical definitions of bias and unfairness have been developed by corporate (Zhang, Lemoine, & Mitchell, 2018), scholarly (Kearns, Roth, & Wu, 2017), and civil society actors (Duarte,

2017). These efforts have historical parallels in the social sciences, particularly around quantitative educational, vocational (Hutchinson & Mitchell, 2019), and psychometric testing (Lussier, 2018). Fairness, Accountability, and Transparency in Machine Learning (FATML) workshops, held yearly in conjunction with the International Conference on Machine Learning (ICML) from 2014 to 2018, were organized by a group comprised largely of computer scientists. In 2018, the first dedicated FAT* Conference was held in New York City, and in 2019 the second edition of the conference became associated with the Association of Computing (ACM) conference series, a testament to the quick increase of interest in the field from machine learning practitioners and other computer science researchers.

Technical tools for mitigating bias in AI systems have coalesced around strategies for ensuring fairness (Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018)—though what the term "fairness" means, and how that definition subsequently shapes computational bias solutions, has remained an open question (Narayanan, 2018). Technical solutions have been developed to ensure both individual statistical fairness (Hardt, Price, & Srebro, 2016) and statistical fairness across pre-defined groups (Verma & Rubin, 2018), while related work has explored human perceptions of algorithmic bias and fairness in everyday contexts (Lee, 2018; Woodruff, Fox, Rousso-Schindler, & Warshaw, 2018). Scholarship has also begun to connect computational models of fairness with extant formalizations from philosophical theories of political and economic equality (Heidari, Loi, Gummadi, & Krause, 2018).

A second cluster of technical work has centered on ensuring AI algorithms and statistical models are accountable and explainable—or "transparent"—to human users and auditors (Miller, 2019). Much like fairness, the definitions of these terms lack consensus (Caplan, Donovan, Hanson, & Matthews, 2018). Tools intended to enable such accountability and transparency tend to focus on demonstrating either the provenance of training data sets or the decision-making processes through which machine-learning models make use of such data, defining terms like accountability and transparency narrowly as referring to a system's usability or the clarity of its interface design (Ananny & Crawford, 2017). Scholars have noted further that transparency and accountability are, like fairness, fundamentally social concepts (Brown, Chouldechova, Putnam-Hornstein, Tobin, & Vaithianathan, 2019; Veale, Van Kleek, & Binns, 2018).

The focus on narrow technical definitions of terms like accountability, bias, or unfairness on the parts of computer scientists and especially large digital technology firms has led many scholars to note such tools are necessary, but entirely insufficient, to address the full spectrum of sociotechnical problems created by AI systems (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). These scholars draw on the STS and critical legal studies literatures noted above, alongside scholarship in critical race studies, women's, gender, disability, and queer studies, and other fields interrogating social/institutional structures of power and domination. Scholars such as Noble (2018), Eubanks (2018), Broussard (2018), and Benjamin (2019) have documented and analyzed the failures of digital tools to properly account for race, gender, sexuality, and other facets of human diversity. Focusing specifically on fairness, Hoffmann (2019) observes that discourses around rights, due process, and antidiscrimination often fail to overcome animus and have at times even "hindered … the transformative and lasting structural change that social justice demands" (p. 901). Costanza-Chock (2018) calls for the application of design justice, or "theory and practice that is concerned with how the design of objects and systems influences the distribution of risks, harms, and benefits among various groups of people," to AI/ML and other digital systems.

These scholars and others have both challenged the emphasis on technical solutions as the primary mechanisms to govern AI/ML systems and the contention that such solutions are value neutral in their execution. No technical solution to the governance of technical systems such as AI/ML is advisable on its own; as Nissenbaum (2011) observes, "however well-designed, well-executed, and well-fortified … [technical] systems are, incipient weaknesses are inevitable and pose a threat to their programmed action" (p. 1386). Such "weaknesses" are the elements of the system that predispose it toward particular normative outcomes that are incommensurate with social values like fairness, democratic accountability, and justice; as Nissenbaum notes further, "an important role for regulation is to remove the temptation to exploit these [technical] weaknesses" (p. 1386). Yet many of the regulatory efforts around AI/ML to date have the opposite valence. As in other recent cases involving novel technologies, the interests and concerns of these technologies' powerful proponents have sought to capture the discourse around social, legal, and regulatory responses to AI/ML technologies.

Governance Through Principles

In conjunction with work on technical mechanisms for governing AI/ML systems, industry players and nonprofit groups have also produced high-level AI "values statements," articulating guidelines for the development and deployment of these technologies. In recent work (Greene, Hoffmann, & Stark, 2019), we interrogated a sample of these statements released by high-profile actors in the area, such as nonprofit AI research company Open AI, industry group the Partnership for AI, and the tenets of the AI Ethics Board for Public Safety of Axon Corporation (formerly Taser). In the interim, a slew of similar statements from corporations, governments, and civil society groups have been announced. Jobin, Ienca, and Vayena (2019) observe some high-level principles such as transparency are commonly invoked across many of these statements, while others, like sustainability and solidarity, are far less common. The diversity of principles and codes around the world makes synthetic analysis of such statements complex, with discourse in the Anglosphere tending to overlook principle sets in languages other than English or not easily accessible via conventional digital channels.

We concluded in (2019) that the AI values statements we analyzed offered a deterministic, expert-driven vision of AI/ML governance, the challenges and pitfalls of which are best addressed through technical and design—not social or political—solutions. In the interim, little has changed. These statements both reflect and reify what Abend (2014) calls the "moral background" for AI/ML development, or the parameters under which ethics are understood and delimited. Perhaps unsurprisingly given the involvement of tech sector companies in many of these statements, there is little acknowledgement in these documents that AI/ML can be limited or constrained by social exigencies or democratic oversight. The rush to apply AI/ML during the COVID-19 pandemic to digital contact tracing, automatic temperature tracking, and other technical interventions with little constraint is a sobering case in point.

High-profile AI vision statements are distinguished by several shared traits. These include an insistence that the positive and negative impacts of AI are a matter of universal concern amenable to a common ethical language; an emphasis on AI governance as an elite, expert project of technical and legal oversight despite a desire to pay lip service to broad stakeholder input; a paradoxical insistence on the inevitability of AI technologies while placing the ethical onus for their governance on humans;

and a focus on technical solutions and design elements such as transparency, and not on the broader political economy in which these systems are embedded, as the necessary locus of ethical scrutiny and AI governance (Greene et al., 2019). These values statements are couched in the descriptive language of STS and the philosophy of technology, indicating that these critical fields have had at least some rhetorical effect.

Recently released AI vision and value statements have deviated little from the core themes described above. The Organization for Economic Cooperation and Development (OECD)'s Principles on AI, adopted in mid-2019, emphasize the development of "trustworthy AI" in order to fuel "inclusive growth" and argue for the facilitation of AI development, not its regulation or potential curtailment ("OECD Principles on Artificial Intelligence," n.d.). Even the Vatican's recent "Call for an AI Ethics," signed in Rome in February 2020, echoes already extant AI principles, calling for AI to be transparent, inclusive, reliable, secure, and impartial—values described in the call's press release as "fundamental elements of good innovation," but with seemingly little connection to Christian ethical traditions (Pontifical Academy for Life, 2020). The Vatican document was cosigned by both IBM and Microsoft, suggesting the effort was as much a public relations exercise as a serious attempt to grapple with the social impacts of AI and automation more broadly.

While the ethical design parameters suggested by AI vision statements share some of the elements and framing of STS and other critical fields, they differ implicitly in normative ends, with explicit goals around social justice or equitable human flourishing often missing. The "moral background" of these ethical AI/ML statements is thus closer to conventional business ethics (Metcalf, Heller, & Boyd, 2016; Moss & Metcalf, 2020), typified by codes of ethical conduct that foreground protecting and consolidating professional and corporate interests (Stark & Hoffmann, 2019). Indeed, the focus on developing high-level principles around AI ethics could be considered strong evidence for the field's attempt to consolidate itself—with many technical divisions and differences across practitioners, agreements around high-minded yet abstract principles can potentially serve not only as a sop to governance efforts from other actors, but also as mechanisms to signal professional membership and insider status.[1] AI/ML ethics statements buttress business-as-usual approaches within technical fields while helping to strengthen the professional clout of AI/ML practitioners.

Governance Through Regulations and Standards

Most national governments responding explicitly to the growth of AI research and development to date have done so primarily through national AI strategies, documents that frequently echo the broad principles of corporate and civil-society vision statements described above (Dutton, Barron, & Boskovic, 2018). Canada was the first country to announce a nationally funded AI strategy in March of 2017. Some of these strategies are paired with increased financial investment in various aspects of AI research and development. As of the end of 2018, the Canadian Institute for Advanced Research (CIFAR) lists nine national AI strategies with funding commitments, while another twenty countries had produced or were at work on AI guidance documents (Dutton et al., 2018, pp. 5–7). In the European Union, both various national governments and the European Commission have produced AI strategic planning documents, the latter including the European Commission's *Ethical Guidelines for Trustworthy AI* (AI HLEG, 2019) developed in 2019. While some national jurisdictions have begun to move forward on binding regulatory regimes for AI and automated systems, this progress has been slow. In tandem with its *Ethical Guidelines*, the European Commission published a white paper in February 2020 on "a European approach" to artificial intelligence intended as the groundwork for binding EU regulations (European Commission, 2020), though critics have noted the EC's recommendations seek to regulate, but not ban, certain applications of AI such as facial recognition (Baraniuk, 2020).

Regional and municipal governments have been more active and successful at developing regulatory responses to AI. Local regulation of AI-enabled facial recognition technologies (FRTs) has been a particularly active area of policymaking. Full or partial bans and moratoria on the deployment of FRTs by local governments and law enforcement agencies have been passed in jurisdictions in the United States, Europe, and elsewhere (Leong, 2019) as the dangers of these technologies to human equality (Stark, 2019) and civil liberties (Hartzog & Selinger, 2018) have become more widely recognized. These regulatory moves are partial ones; they generally do not cover private sector deployment of FRTs (Wright, 2019), nor the deployment of these technologies in educational institutions (Andrejevic & Selwyn, 2019). Moreover, these regulations often fail to address the wide range of AI-equipped analytic technologies designed to surveil elements of human bodies and behavior alongside the

face—such as gait recognition or emotion analytics. Nonetheless, local regulation of AI technologies is a welcome regulatory first step—and has often been catalyzed by the work of social justice groups such as the Movement for Black Lives (more below).

Another potentially promising set of mechanisms for regulating the application of AI systems are algorithmic impact assessments (AIAs) (Reisman, Schultz, Crawford, & Whittaker, 2018; Selbst, 2017): procedural mechanisms through which institutions systematically assess the potential risks and outcomes of automated decision systems before they are deployed. Based on a variety of similar impact assessment processes in environmental regulation, human rights law, and more recently digital privacy scholarship (Bamberger & Mulligan, 2008), AIAs have attracted interest from varying levels of government, including the New York City municipal and the Canadian federal governments (Cardoso, 2019). New York City formed an Automated Decision Systems Task Force in 2017 to assess how such mechanisms were being used in the municipal context and provide recommendations for their regulation. The Task Force's report (New York City, 2019) released in late 2019, advocated for the creation of a city Algorithms Management and Policy Officer but was criticized for its lack of other specific policy suggestions (Lecher, 2019); a Shadow Report from the AI Now Institute (Richardson, 2019) argued for broader and more rigorous application of AIA processes in all levels of government, including broad applicability and periodic external reviews of their effectiveness. The uneven implementation of AIAs suggests the challenge of moving to binding regulation around automated systems—not so much because of technical difficulties as due to political pressures to ensure the powerful continue to benefit from the deployment of these technologies.

High-profile national AI policies and supranational statements of principle are also quickly being supplemented by more granular corporate mechanisms for the governance of AI/ML systems: internationally recognized technical standards and sets of ethical design principles. These mechanisms extend the technocratic, processes-grounded and expert-focused themes often found in AI vision statements. As such, they are a means to cement such solutions to the societal problems posed by AI/ML systems as the main field of debate for practitioners and policymakers. Standards, unlike regulation, implicitly work within the growth plans of industry and serve to coordinate individual enterprises around interoperability and consistency; they support, rather than hinder, the notion

of regulation as an enabler of the tool-focused approach preferred by AI/ML companies.

A variety of private and public organizations at the national and international level have begun the process of developing standards around AI (Cihon, 2019). The International Organization for Standardization (ISO), an independent, nongovernmental international organization, has begun to develop standards around AI in conjunction with the International Electrotechnical Commission (IEC) through Subcommittee 42 of the two organizations' Joint Technical Committee (JTC) 1, the latter formed in 1987 to develop global digital technology standards. The ISO/IEC JTC 1/SC 42 process is in its early stages and has produced a number of drafts currently being developed in committee around AI topics including ISO/IEC WD 22989: Artificial intelligence—Concepts and terminology and ISO/IEC WD 23053: Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML).

The AI standards-making activities of the Institute of Electrical and Electronics Engineers (IEEE), which describes itself as "the world's largest technical professional organization for the advancement of technology" are somewhat more advanced. As part of its *Global Initiative on Ethics of Autonomous and Intelligent Systems*, the organization has published an omnibus set of high-level AI ethics principles, *Ethically Aligned Design* (IEEE, 2019) and is in the process of developing particular standards through a variety of focused working groups on topics such as Transparency of Autonomous Systems (P7001), a Standard for Child and Student Data Governance (P7004), and a Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems (P7008).[2] Of particular relevance to contemporary discussions is working group P7013 on Inclusion and Application Standards for Automated Facial Analysis Technology. The group seeks to create "phenotypic and demographic definitions that technologists and auditors can use to assess the diversity of face data used for training and benchmarking algorithmic performance" as well as "a rating system to determine contexts in which automated facial analysis technology should not be used."[3] The IEEE has also begun an Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS), the goal of which is to "offer a process and define a series of marks by which organizations can seek certifications for the processes around the A/IS products, systems, and services they provide."[4] As the IEEE notes, participation in this metrification of AI governance requires an IEEE Standards Association Corporate Membership.

In the United States, the White House Office of Science and Technology Policy under the Obama administration identified the AI-adjacent field of "big data" as both a strategic priority and an area of legal and ethical concern through a series of reports between 2014 and 2016 (Muñoz, Smith, & Patil, 2016). In early 2019, the Trump administration released an Executive Order on "Maintaining American Leadership in Artificial Intelligence," which laid out some elements common to other national AI strategies (White House, 2019). The US Commerce Department's National Institute of Standards and Technology (NIST) was charged with developing national AI standards in a February 2019 Executive Order; they produced a plan for federal engagement in the topic in August of 2019 (NIST, 2019). The plan's recommendations focus narrowly on developing benchmarks, tools, and metrics for AI systems with little attention to the broader societal impacts of AI and automation.

Governance Through Human Rights

The pitfalls apparent in partial technical solutions, nonbinding national strategies, and voluntary standard setting as AI governance mechanisms have prompted many civil society actors to call for an approach grounded in global human rights discourse. According to Latonero (2018), AI's "design and deployment should avoid harms to fundamental human values," while "international human rights provide a robust and global formulation of those [same] values" (p. 5). Donahoe and Metzger (2019) argue human rights provide "a framework that can claim global buy-in and that addresses the roles and responsibilities of both government and the private sector when it comes to accountability for the impact of AI-based decisions" (p. 118). The authors cite the 2011 UN Guiding Principles on Business and Human Rights (known colloquially as the Ruggie Principles) as a key mechanism to ensure private companies apply human rights law to their own products, services, and operations.

Other scholars have advanced more cautious endorsements of the application of human rights frameworks to AI governance. Daniel Munro (2019) observes that while the "shared ethical language" of extant human rights conventions can "help to overcome the challenge of coordinating multiple, and sometimes incompatible, frameworks," he also warns of three pitfalls to such an approach. The first concerns enforcing human rights covenants against the private businesses and other entities

manufacturing and selling AI systems. Munro notes the Ruggie Principles do provide some guidance on the parameters of such an approach but that concrete enforcement mechanisms remain inadequate. Second, high-level human rights frameworks suffer from some of the same problems around the interpretation of abstract principles as other extant AI ethics codes.

Third, and to our mind most critically, Munro shares the concern of philosophers such as Mathias Risse, who observe the "minimal standards" approach to much human rights discourse is insufficient to account for positive values of both distributive and relational justice and equality (Anderson, 1999). This critique—that at best such rights frameworks are conceptually too narrow, trading positive visions of radical social and economic justice for a cramped vision of negative political liberty—is just one of many leveled by progressives at human rights as an ethical framework.[5] Critics often note the centrality of political rights to human rights discourse, in contrast to the relative paucity of effort to protect economic and social rights. Others argue the international legal regimes enabling and enforcing such rights are imperialistic, exporting Western values around the world (Anghie, 2005, 2013), and that human rights are explicitly neoliberal, guaranteeing the survival of populations and individuals only so that they can be further exploited by global capital (Anghie, 2005, 2013; Moyn, 2011, 2013, 2018). Moyn (2018) also observes that human rights principles often fail to guarantee human flourishing and diminish human suffering in practice precisely because they lack reliable mechanisms for dispute by which they can be activated.

Unfortunately, suggested applications of human rights frameworks to AI governance often duplicate many of the limitations we identified in other high-level AI ethics principles described above. These limitations include a focus on design as the locus of ethical activity in work around human-rights-by-design. While salutary, such a design focus does not grapple sufficiently with the systemic logics behind many of AI's deleterious effects (Penney, McKune, Gill, & Deibert, 2018). A focus in human-rights-by-design on the need for technical and business transparency is also paralleled in high-level ethics codes. Yet emphasis on transparency as a procedural virtue is insufficient when considering the real costs to human flourishing produced by many AI-driven technologies. In sum, governance strategies for AI/ML based on traditional human rights frameworks have not yet avoided the insufficiencies of similar high-level statements of principle; major companies such as Salesforce have

already grounded their AI policies in human rights principles, a testatment to how closely corporate and human rights ethics discourses can potentially overlap.

Our aim here, however, is not to further denigrate human rights discourse. As Latonero observes, human rights provide a familiar framework legible to a wide array of actors around the world. It is notable that the surge in academic and public conversations around AI has returned human rights to the fore in a historical period in which the human rights apparatus, always more honored in the breach than in the observance, has been under attack from many quarters. If human rights frameworks are applied both carefully and radically to the governance of artificial intelligence, both those systems and human rights discussions themselves stand to be improved and strengthened as a result.

Governance Through Securitization

Security policymakers have long sought to frame the digital world in terms with which they are already familiar, drawing on a broader discourse around "cyberspace" already extant in popular literature and society from the mid-1980s (Eriksson, 2001). Newly developed AI technologies have been slotted neatly into this vision by policymakers and defense pundits: frequent references in the media to a new "AI arms race," particularly between China and the United States, have threatened to recast Cold War patterns of competition as models for contemporary AI governance. Artificial intelligence technologies are certainly of major interest to the defense sector, sparking reasonable anxieties around AI/ML systems as military assets, either to augment conventional warfare, such as through their use in unmanned aerial vehicle (UAV) or drone technologies, or as aides in cyberattacks. Moreover, reports that international technology companies are willing to work closely with, or in some cases directly as proxies for, national governments have also associated national AI policies with espionage.

However, this emphasis on AI through the lens of already existing cybersecurity discourses and as a national security problem risks further entrenching AI policy in the hands of the few. Securitization theory argues that discourses around "security" deploy particular "grammars of securitization" to connect "referent objects, threats, and securitizing actors together" to both depoliticize and control particular sociotechnical arenas (Hansen & Nissenbaum, 2009). AI's securitization shares similar traits

with those of cybersecurity more broadly (which itself draws strongly from Cold War discourse around nuclear arms). AI's "grammar of securitization" both reinforces AI's status as an elite technical discourse and distracts from the effects of that same discourse on AI policymaking across the board. Hansen and Nissenbaum (2009) identify three "securitization modalities" which they suggest are particularly powerful in broader cybersecurity discourse: hyper-securitization or "a tendency both to exaggerate threats and to resort to excessive countermeasures"; everyday security practices, i.e., "situations in which private organizations and businesses mobilize individual experiences" to "make hyper-securitization scenarios more plausible by linking elements of the disaster scenario to experiences familiar from everyday life"; and technification, "a particular constitution of epistemic authority and political legitimacy" whereby some subjects are depoliticized into the realm of the "expert" (p. 1164).

The concept of technification aligns closely with our observations in past work (Greene et al., 2019) and above: that most AI vision statements or sets of ethical principles seek to restrict debates around the societal impacts of AI to a coterie of technical experts whose positions are posited, chiefly by themselves, as technocratic and thus apolitical. Securitization rhetoric thus allows both industry and governments free reign to set the parameters for AI governance on their own terms. In its simultaneous insistence on the inevitability of AI technologies and that experts' hold responsibility for ethical AI governance, much AI discourse also performs a flavor of hyper-securitization, hyping the disruption being produced by AI systems while disavowing the role of policy choices in accepting the deployment of problematic technologies. And in line with other surveillance technologies such as CCTV cameras, institutional actors are rapidly securitizing AI as a threat in itself if deployed by the wrong hands, while also advocating its use as a means to identify threats to the national polity (Cassiano, 2019; Chen & Cheung, 2017).

The securitization of AI discourses and technologies also further exacerbates existing patterns of digital inequality and power asymmetries between the global North and South. Rumman Chowdhury and Abeba Birhane argue the extractive data practices of many AI firms constitute "algorithmic colonialism," a digital analogue to the exploitative material extraction of natural resources and human capital to which the Global South has been subjected for centuries (Chowdhury, 2019). Algorithmic colonialism entails both the depredations of neocolonial powers (in particular the United States and China) and those of local elites supported by

broader networks of global capital (Couldry & Mejias, 2019); it also includes algorithmic settler colonialism, through which settler colonial states look inward to exploit and dispossess the data heritage of indigenous populations (TallBear, 2013). Finally, it relies on globalized digital infrastructures to outsource and occlude the immense amounts of human labor of which "AI" services often actually consist: workers around the world are paid meagerly to clean and tag data, train models, and facilitate other purportedly automated services, while the more privileged developers of these systems reap the lion's share of the financial reward (Gray & Suri, 2019; Poster, 2019a, 2019b).

Securitization discourse, particularly around technification, thus seeks to push AI/ML systems out of the hands of citizens in both the global North and South and to install them as yet another weapon in the game of neoliberal great power politics. More broadly, securitization supports and underpins the various regulatory, technical, and discursive proposals outlined above designed to narrow the range of people and institutions able to have a say in the governance of AI/ML—to varying degrees, the proposed governance mechanisms for these technologies ensure their fate lies with a particular subset of more or less privileged individuals.

Conclusion: Future Alternatives for AI/ML Governance

None of the various existing or proposed mechanisms for AI/ML governance, ranging from computational tools to global regimes around human rights and securitization, are entirely antithetical to the broader systems of neoliberal governance and capitalist accumulation responsible for the development and deployment of contemporary AI/ML technologies—a fact hardly surprising. What are our proposed alternatives? At a baseline, truly just and equitable governance of AI/ML technologies will have to more or less radically transform the development and deployment of those technologies themselves; in turn, the future development of these transformed AI/ML technologies cannot be grounded in the values of the Deleuzian "society of control" (Deleuze, 1990), wherein societal life is modulated through digital manipulation overseen by state and corporate power. This is a tall order, requiring the work of many hands. Elsewhere, we have variously described some of the strategies required, including foregrounding data justice (Hoffmann, 2019), dissecting digital inequality (Greene, forthcoming), overhauling professional ethics in the

2-3
2 notes:
04
nick merrill

AI/ML space (Stark & Hoffmann, 2019), and engaging with a diverse array of queer ethical traditions (Stark & Hawkins, 2019).

One through-line across this and much related scholarship is the necessity of [illegible] grounding and centering the expertise of communities affected by AI/ML systems in their design and use (Costanza-Chock, 2018; Forlano & Mathew, 2013; Madaio, Stark, Wortman Vaughan, & Wallach, 2020; Sloan, Moss, Awomolo, & Forlano, 2020). Politics is a social system of collective deliberation around decision making and collective distribution around resource matching. Parallel political projects for technological reform and governance thus suggest alternative ways both to govern who makes decisions about technologies such as AI/ML and who benefits from their effects. Workplace labor movements such as the Tech Won't Build It campaign and the abolitionist and antisurveillance work of the Movement for Black Lives (MBL) are thus exemplary of alternative paradigms for AI/ML governance.

The Tech Won't Build It campaign coalesced in 2018 around Google employee labor action against the company's involvement in the United States Department of Defense's Maven contract for targeted autonomous weapons systems. Workers from multiple companies, including Microsoft, Salesforce, Amazon, and Palantir, began to protest similar entanglements by their firms with both military and immigration enforcement efforts by the US government. These protests have grown to include work stoppages, open letters illustrating the gap between espoused corporate values and actual corporate practices, and support for other social justice organizations.

Tech Won't Build It is broadly a movement focused on cultivating workplace democracy through labor action, holding that workers developing AI/ML should have a say in how such technologies are deployed. This focus is distinct from the narrow focus on expertise in AI ethics statements and other governance mechanisms—tech sector workers here claim a seat at the table not because of their qualifications but because of their position in the production process. The labor power of technology workers is both a statement of legitimacy and a threat due to their strategic position, with an explicit focus on reducing profits (or at least redirecting them away from morally indefensible ends).

The Black Lives Matter movement was begun in 2013 by Alicia Garza, Patrisse Cullors, and Opal Tometi after George Zimmerman's acquittal for the murder of Trayvon Martin; it has since become an international movement committed to creating a world where police and

prisons—violent institutions that exacerbate social problems—are unnecessary. As an abolitionist project, the Movement for Black Lives' policy platform provides a useful contrast to ethical design manifestoes. Under the section "An End to the Mass Surveillance of Black Communities, and the End to the Use of Technologies that Criminalize and Target Our Communities," the MBL policy platform addresses many of the same technologies addressed by ethics principles and other governance mechanisms. However, in contrast to the narrow, elite-focused stance of most governance mechanisms, MBL frames the solution to AI/ML governance as increased democratic oversight of technological procurement and deployment.

The abolitionist perspective of the MBL differs from other governance mechanisms described above in at least two ways. First, the project of defining harm and redress takes a longer and more contextualized view; while everyone might benefit from MBL's policy prescriptions, prison abolition and police violence are historically specific problems afflicting Black people in colonial societies. Second, rather than beginning with abstract ideal principles, the MBL platform is grounded in its vision of the ideal community—one in which prison abolition is a reality—and works backward to identify general principles and then the specific policy stances and campaigns they imply. Unlike many of the governance mechanisms we describe, this is also a mass mobilization document that embraces conflict and agonism as mechanisms for democratic change.

Not only do these movements encourage a broader governance conversation focused explicitly on social justice and AI/ML—urgently needed as these technologies become ubiquitous—they also point toward ways in which technical and social governance can interoperate for the benefit of all. What are our various collective visions of the ideal community, and how can the governance of AI/ML systems play a part? And how can members of diverse communities, often with asymmetric access to wealth and power, work together to ensure justice, equality, and fairness exist not just in principle but also in practice? We hope these questions can open a more radically inclusive and democratic conversation around AI governance—one that surpasses technical fixes and narrow expert guidelines to embrace the heterogeneity of needs and desires centered on human intelligence, ability, and solidarity.

Notes

1. We are grateful to Bart Simon for crystallizing this observation.
2. See https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.
3. https://standards.ieee.org/project/7013.html.
4. https://standards.ieee.org/industry-connections/ecpais.html.
5. Conservative critiques invariably assail the notion of human rights as a mechanism through which tenets of natural law are subverted, leading to various forms of social malaise and decay.

References

Abend, G. (2014). *The moral background: An inquiry into the history of business ethics*. Princeton University Press.

AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Ananny, M., & Crawford, K. (2017). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645.

Anderson, E. S. (1999). What is the point of equality? *Ethics*, *109*(2), 287–337. http://www.philosophy.rutgers.edu/joomlatools-files/docman-files/4ElizabethAnderson.pdf.

Anghie, A. (2005). *Imperialism, sovereignty and the making of international law*. Cambridge University Press.

Anghie, A. (2013). Whose Utopia? Human rights, development, and the Third World. *Qui Parle, 22*(1), 63–69. https://doi.org/10.5250/quiparle.22.1.0063.

Andrejevic, M., & Selwyn, N. (2019). Facial recognition technology in schools: Critical questions and concerns. *Learning, Media and Technology* (2), 1–14. http://doi.org/10.1080/17439884.2020.1686014.

Bamberger, K. A., & Mulligan, D. K. (2008). Privacy decisionmaking in administrative agencies. *Chicago Law Review, 75*(1), 75–107.

Baraniuk, C. (2020, February 19). EU to tackle AI "Wild West"—But still to say how. *BBC News*. https://www.bbc.com/news/technology-51559010.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Wiley.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). *Toward algorithmic accountability in public services* (pp. 1–12). Presented at the 2019 CHI Conference, New York, NY, USA: ACM Press. http://doi.org/10.1145/3290605.3300271.

Caplan, R., Donovan, J., Hanson, L., & Matthews, J. (2018). *Algorithmic accountability: A primer*. Data & Society Research Institute.

Cardoso, T. (2019, May 28). Federal government unveiling risk assessment tool for artificial intelligence. *The Globe & Mail*. https://www.theglobeandmail.com/politics/article-federal-government-unveiling-risk-assessment-tool-for-artificial/.

Cassiano, M. S. (2019). China's Hukou platform: Windows into the family. *Surveillance & Society, 17*(1/2), 232–239.

Chen, Y., & Cheung, A. S. Y. (2017). The transparent self under big data profiling: Privacy and Chinese legislation on the social credit system. *The Columbia Science & Technology Law Review, 12*(2), 356–378. https://doi.org/10.2139/ssrn.2992537.

Chowdhury, R. (2019). *AI ethics and algorithmic colonialism*. https://www.mcgill.ca/igsf/channels/event/rumman-chowdhury-ai-ethics-and-algorithmic-colonialism-300414.

Cihon, P. (2019). *Standards for AI governance: International standards to enable global coordination in AI research & development*. Future of Humanity Institute, University of Oxford.

Costanza-Chock, S. (2018). *Design justice: Towards an intersectional feminist framework for design theory and practice*. Presented at the Design Research Society. http://doi.org/10.21606/dma.2017.679.

Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media, 20*(4), 336–349. https://doi.org/10.1177/1527476418796632.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., … Raji, D. (2019). *AI now 2019 report*. AI Now Institute.

Deleuze, G. (1990). Postscript on control societies. In *Negotiations, 1972–1990* (pp. 177–182) (M. Joughin, Trans.). Columbia University Press.

Donahoe, E., & Metzger, M. M. (2019). Artificial intelligence and human rights. *Journal of Democracy, 30*(2), 115–126. https://doi.org/10.1353/jod.2019.0029.

Duarte, N. (2017, August 8). *Digital decisions tool*. Center for Democracy & Technology. https://cdt.org/insights/digital-decisions-tool/.

Dutton, T., Barron, B., & Boskovic, G. (2018). *Building an AI world*. Canadian Institute for Advanced Research.

Dwork, C., & Mulligan, D. K. (2013). It's not privacy, and it's not fair. *Stanford Law Review Online, 66,* 35–40.

Eriksson, J. (2001). Cyberplagues, IT, and security: Threat politics in the information age. *Journal of Contingencies and Crisis Management, 9*(4), 211–222.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Commission. (2020). *Artificial intelligence—A European approach to excellence and trust* (No. COM[2020] 65 final). European Commission. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Forlano, L., & Mathew, A. (2013). *The designing policy toolkit*. Urban Communication Foundation.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347.

Gandy, O. H. (2009). Engaging rational discrimination: Exploring reasons for placing regulatory constraints on decision support systems. *Ethics and Information Technology, 12*(1), 29–42. https://doi.org/10.1007/s10676-009-9198-6.

Gray, M. L., & Suri, S. (2019). *Ghost work*. Houghton Mifflin Harcourt.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer. In T. X. Bui & R. H. Sprague (Eds.), (pp. 2122–2131). Presented at the Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS). https://hdl.handle.net/10125/59651.

Hansen, L., & Nissenbaum, H. (2009). Digital disaster, cyber security, and the Copenhagen School. *International Studies Quarterly, 53*, 1155–1175.

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. arXiv:1610.02413 [cs.LG], pp. 1–9.

Hartzog, W., & Selinger, E. (2018, August 2). Facial recognition is the perfect tool for oppression. *Medium*. https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66.

Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2018, September 10). *A moral framework for understanding of fair ML through economic models of equality of opportunity*. arXiv:1809.03400 [cs.LG]. https://arxiv.org/abs/1809.03400.

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society, 22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912.

Hutchinson, B., & Mitchell, M. (2019). *50 years of test (un)fairness* (pp. 49–58). Presented at the Conference on Fairness, Accountability, and Transparency 2019, New York, NY, USA: ACM Press. http://doi.org/10.1145/3287560.3287600.

IEEE/The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design, first edition*. https://ethicsinaction.ieee.org/.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1–11. http://doi.org/10.1038/s42256-019-0088-2.

Johnson, D. G. (2007). Ethics and technology "in the making": An essay on the challenge of nanoethics. *Nanoethics, 1*(1), 21–30. https://doi.org/10.1007/s11569-007-0006-7.

Kearns, M., Roth, A., & Wu, Z. S. (2017). *Meritocratic fairness for cross-population selection* (pp. 1–9). Presented at the Proceedings of the International Conference on Machine Learning, Sydney, Australia.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings, 108*, 22–27. https://doi.org/10.1257/pandp.20181018.

Latonero, M. (2018). *Governing artificial intelligence*. Data & Society Research Institute. https://datasociety.net/output/governing-artificial-intelligence/.

Lecher, C. (2019, November 20). NYC's algorithm task force was "a waste," member says. *The Verge*. https://www.theverge.com/2019/11/20/20974379/nyc-algorithm-task-force-report-de-blasio.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1). http://doi.org/10.1177/2053951718756684.

Leong, B. (2019). Facial recognition and the future of privacy: I always feel like . . . somebody's watching me. *Bulletin of the Atomic Scientists, 75*(3), 109–115. http://doi.org/10.1080/00963402.2019.1604886.

Lussier, K. (2018). Temperamental workers: Psychology, business, and the Humm-Wadsworth Temperament Scale in interwar America. *History of Psychology*, 1–22. http://doi.org/10.1037/hop0000081.

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). *Co-designing checklists to understand organizational challenges and opportunities around fairness in AI* (pp. 1–20). Presented at the CHI 2020: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Honolulu, HI. http://doi.org/10.1145/3313831.3376445.

Metcalf, J., Heller, E. F., & Boyd, D. (2016). *Perspectives on big data, ethics, and society*. The Council for Big Data, Ethics, and Society.

Miller, T. (2019). But why? Understanding explainable artificial intelligence. *XRDS: Crossroads, the ACM Magazine for Students, 25*(3), 20–25. http://doi.org/10.1145/3313107.

Moss, E., & Metcalf, J. (2020). *Ethics owners: A new model of organizational responsibility in data-driven technology companies*. New York: Data & Society Research Institute. https://datasociety.net/pubs/Ethics-Owners.pdf.

Moyn, S. (2011). *The last Utopia*. Belknap Press.
Moyn, S. (2013). The continuing perplexities of human rights. *Qui Parle, 22*(1), 95–115. https://doi.org/10.5250/quiparle.22.1.0095.
Moyn, S. (2018). *Not enough: Human rights in an unequal world*. Belknap Press.
Munro, D. (2019, July 12). *Artificial intelligence needs an ethics framework*. Centre for International Governance Innovation. https://www.cigionline.org/articles/artificial-intelligence-needs-ethics-framework.
Muñoz, C., Smith, M., & Patil, D. J. (2016). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President.
Narayanan, A. (2018). *Translation tutorial: 21 fairness definitions and their politics*. Presented at the FAT* 2018, New York.
National Institute of Standards and Technology (NIST). (2019). *U.S. leadership in AI: A plan for federal engagement in developing technical standards and related tools*. https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.
New York City. (2019, November). *Automated decision systems task force report*. https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf.
Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford Law Books.
Nissenbaum, H. (2011). From preemption to circumvention. *Berkeley Technology Law Journal, 26*(3), 1367–1386.
Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
OECD Principles on Artificial Intelligence. (n.d.). https://www.oecd.org/going-digital/ai/principles/.
Penney, J., McKune, S., Gill, L., & Deibert, R. J. (2018, December 20). Advancing human-rights-by-design in the dual-use technology industry. *Journal of International Affairs*. https://jia.sipa.columbia.edu/advancing-human-rights-design-dual-use-technology-industry.
Pfaffenberger, B. (1992). Technological dramas. *Science, Technology and Human Values, 17*(3), 282–312.
Pontifical Academy for Life. (2020). *Rome call 2020*. https://romecall.org/romecall2020/.
Poster, W. R. (2019a). Racialized surveillance in the digital service economy. In R. Benjamin (Ed.), *Captivating technology: Race, technoscience, and the carceral imagination* (pp. 133–169). Duke University Press.
Poster, W. R. (2019b). Sound bites, sentiments, and accents: Digitizing communicative labor in the era of global outsourcing. In D. Ribes & J. Vertesi (Eds.), *DigitalSTS: A field guide for science technology studies* (pp. 240–262). Princeton University Press.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). *Algorithmic impact assessments*. AI Now Institute.

Richardson, R. (Ed.) (2019). *Confronting black boxes: A shadow report of the New York City automated decision system task force*. AI Now Institute. https://ainowinstitute.org/ads-shadowreport-2019.html.

Schwab, K. (2017). *The fourth industrial revolution*. Portfolio Penguin.

Selbst, A. D. (2017). Disparate impact in big data policing. *Georgia Law Review, 52,* 109–195.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). *Fairness and abstraction in sociotechnical systems* (pp. 59–68). Presented at the Proceedings of the Conference on Fairness, Accountability, and Transparency, New York, NY, USA: Association for Computing Machinery. http://doi.org/10.1145/3287560.3287598.

Sloan, M., Moss, E., Awomolo, O., & Forlano, L. (2020). *Participation is not a design fix for machine learning* (pp. 1–7). Presented at the Proceedings of the International Conference on Machine Learning, Vienna, Austria.

Stark, L. (2019). Facial recognition is the plutonium of AI. *XRDS: Crossroads, the ACM Magazine for Students*, *25*(3), 50–55. http://doi.org/10.1145/3313129.

Stark, L., & Hawkins, B. (2019, 9 December). *Queering AI ethics: Pedagogy and practice*. Thirty-third Conference on Neural Information Processing Systems (NeurIPS), Queer in AI Workshop, Vancouver, BC.

Stark, L., & Hoffmann, A. L. (2019). Data is the new what? Popular metaphors & professional ethics in emerging data culture. *Journal of Cultural Analytics*, 1–22. http://doi.org/10.22148/16.036.

TallBear, K. (2013). Genomic articulations of indigeneity. *Social Studies of Science, 43*(4), 509–533. https://doi.org/10.1177/0306312713483893.

Veale, M., Van Kleek, M., & Binns, R. (2018). *Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making* (pp. 1–14). Presented at the Extended Abstracts of the 2018 CHI Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/3173574.3174014.

Verma, S., & Rubin, J. (2018). *Fairness definitions explained* (pp. 1–7). Presented at the 2018 ACM/IEEE International Workshop on Software Fairness, New York, New York, USA: ACM Press. http://doi.org/10.1145/3194770.3194776.

White House. (2019, 11 February). *Executive Order on maintaining American leadership in artificial intelligence*. https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.

Winner, L. (1988). Do artifacts have politics? In *The whale and the reactor* (pp. 19–39). University of Chicago Press.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). *A qualitative exploration of perceptions of algorithmic fairness* (pp. 1–14). Presented at the Extended Abstracts of the 2018 CHI Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/3173574.3174230.

Wright, E. (2019). The future of facial recognition is not fully known: Developing privacy and security regulatory mechanisms for facial recognition in the retail sector. *Fordham Intellectual Property, Media & Entertainment Law Journal, 29*(2). https://ir.lawnet.fordham.edu/iplj/vol29/iss2/6.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. arXiv:1801.07593 [cs.LG]. https://arxiv.org/abs/1801.07593.

Zuboff, S. (2019). *The age of surveillance capitalism*. PublicAffairs and Hachette.

Critical Perspectives on Governance Mechanisms for AI / ML Systems

Stark, Luke; Greene, Daniel; Hoffmann, Anna Lauren

01 nick merrill Page 1

6/12/2020 0:45

This chapter provides a critical overview of proposed mechanisms for the ethical design and governance of contemporary artificial intelligence (AI) and machine learning (ML) systems. We describe various proposed mechanisms for AI/ML governance, paying particular attention to the possibilities and limits of these mechanisms for realizing truly just and equitable societal outcomes. In doing so we argue each category of intervention has reinforced and supported the broader regimes of corporate and state power under which AI/ML technologies are being developed, and that reformist initiatives relying on these mechanism risk cooptation and failure. We conclude by highlighting the abolitionist approach to AI-driven surveillance technology taken by the Movement for Black Lives and the workplace democracy approach taken by the #TechWontBuildIt technology workers campaign as alternative paradigms for AI/ML governance.

02 nick merrill Page 15

6/12/2020 0:45

radical modes of governance for AI and ML

03 nick merrill Page 15

10/12/2020 20:25

04 nick merrill Page 15

10/12/2020 20:25

05 nick merrill Page 16

10/12/2020 20:25

06 nick merrill Page 16

10/12/2020 20:25

07 nick merrill Page 16

10/12/2020 20:25

08 nick merrill Page 16

10/12/2020 20:25