Peter M. Asaro

Politicizing Data: AI Ethics as a Social Critique of Algorithms

FOR THIS SPECIAL ISSUE OF *SOCIAL RESEARCH*, "FRONTIERS OF SOCIAL INQUIRY," I want to examine the recent rise of AI ethics as a new domain of social inquiry directed at an emerging technology, artificial intelligence (AI). The interest in AI ethics has been growing over the past decade, greatly accelerating in the past year with the release of ChatGPT and the subsequent increased public awareness of the disruptive potential of these new technologies.

In order to understand the discourse around AI ethics and its significance, it is necessary to identify the various threads that have come together to form it. Most contributions to the field tend to draw upon only one or two of these threads. Moreover, there are two radically distinct discourses (one from critical social theory, the other from a techno-capitalist perspective) that have both embraced the general framework of AI ethics with very different conceptions of what it is and what it implies for the future development of AI. This fundamental tension within AI ethics means that in some ways it is the most powerful social critique of technology to date, and in other ways it has been co-opted as a strategy to avoid any substantive legal regulation of AI and the Big Tech companies leading its development.

To appreciate the power of the AI ethics critique, it is helpful to review some early history of the field of science and technology studies (STS), which has applied the techniques of social and cultural inquiry to the fields of science and technology, but not usually as a

form of critique. Similarly, in order to appreciate the frustration with AI ethics as a form of regulatory avoidance or even "ethics washing," it is necessary to understand both what alternatives are available for regulating technologies like AI and the more general shortcomings of "ethics" as a framework for the social control of technology.

The first section explores the general history of STS for the benefit of those not familiar with it. It also examines some early attempts to develop a general social critique of technology from within STS. The essay then considers labor critiques of technology that go back to the Luddites and Karl Marx. While these critiques have been highly developed and widely recognized, they are limited almost exclusively to the domain of work and the exploitation of labor. To flesh out the historical threads of AI ethics, the first section also provides a high-level summary of the history of engineering ethics, a field of applied ethics that emerged primarily as a means to improve safety in the design of public works. But it also adopted a framework of "ethics" that centered on the responsibility of individual engineers, encouraging them to do what they believed was right from an engineering perspective, even if they were being pressured by clients or managers to cut corners to save costs or meet deadlines.

The first section ends with a consideration of the critiques of software that emerged primarily from applying labor critiques of technology to computer and software systems in the workplace. While these have been mostly limited to IT systems and their impacts on workplace culture and labor politics, many aspects of these critiques can be generalized to other social domains. One of the key researchers in that field, Harvard business anthropologist Shoshana Zuboff, has recently developed one of the most powerful critiques of Big Tech and its data-driven targeted advertising and attention manipulation in what she has termed "surveillance capitalism" (2019). There is a clear line in her work from how technology transforms the power dynamics of the workplace to the more general concern about how the collection of vast amounts of data about society and individuals and its use in predicting and manipulating human behav-

ior is a political and economic revolution of the same scale as the Industrial Revolution.

The second section briefly explains the history and nature of AI. It also dispels some commonly held misconceptions about what AI is, at least currently, and concludes with a reflection on why this moment in history has become enthralled with AI and why it is in many ways a stand-in for technology in general and especially information technology in our social lives.

The third section looks at the rise of AI ethics proper, largely in the 2010s. This section also examines the reasons for the great success of AI ethics in capturing the attention of the public, government, and tech companies, a wave that may have crested recently with the US government and White House publishing an "AI Bill of Rights" and getting seven Big Tech companies to sign on to an agreement to build "ethical AI."

The final section explores the many weaknesses of the recent wave of AI ethics in terms of implementation, regulatory control, safety guidance, and provision for the democratic control and reform of technology.

SCIENCE AND TECHNOLOGY STUDIES AS SOCIAL CRITIQUE

The emergence of the field of science and technology studies (STS) in the 1970s and 1980s brought the methods and practices of social science to bear upon the natural sciences, engineering, and technology. What researchers found was that the construction of empirical facts and scientific knowledge was a highly disciplined set of human practices (Latour 1988). There have been great controversy and debate over whether applying social science methods to the natural sciences constituted a critique of the natural sciences as being fallible or flawed or a critique of the epistemic basis of the empirical sciences and thus a gateway to epistemic relativism moored only to social norms and human agency (Hilgartner 1997). But for all the controversy, STS and its practitioners largely accepted the natural sciences

as reliable forms of knowledge *because* of the social processes of peer review, empirical verification, and experimental replication inherent in their disciplined practices.

Early work in STS also revealed the many ways in which technologies were socially constructed and that social interests were imposed upon and shaped the emergence of technologies (Pinch and Bijker 1984). The social analysis of technological systems by STS found that technology is not determined purely or exclusively by social interests, nor solely by the forces of nature, some intrinsic natural order, or engineering constraints. Rather, social interests and desires lead to attempts to manipulate and control physical phenomena and forces, which meet real material resistance and affordances as well as social and political resistance, which in turn meet with accommodations and further attempts, ultimately tuning social and material forces together in a "dance of agency" (Pickering 1995) or failing to do so (Latour 1996). Importantly, there was also a recognition that all technologies are, in fact, socio-technical systems that entrain both social practices and material capabilities and constraints in order to be adopted and have an impact on the world. On the one hand, technologies only work because of many social inputs and accommodations, as well as maintenance—in other words, no technology functions completely independently of human society. On the other hand, technologies inevitably transform social practices and relations (and, importantly, political and economic relations); most technologies require such social transformations in order to be adopted or to even function; and, in many cases, these transformations are actually goals of introducing the new technology.[1]

Lacking from these social studies of science and technology was any analysis of power, or the ways in which scientific knowledge and material technologies are structured to serve the interests of specific individuals or social groups over others. Nor could they provide a normative analysis of what technologies might be "good" or "bad" apart from the interests of the actors who built and used them, which were not themselves subject to any moral analysis. Thus, while STS

could provide detailed accounts of the social construction of knowledge and the consequences of the implementation of technologies, it was largely unable to provide critical analyses that called out the politics or values inherent in these. Without the ability to provide reasons for preferring one technological configuration over another, it was left to merely describe how the emergent form of technology had come to be as it is. Such an analysis could not offer better alternatives or determine which among available alternatives was preferable. Nor could it suggest policies to regulate technologies in order to encourage better systems or systems that would promote some set of values or interests. In this sense STS remained descriptive rather than prescriptive. There were, however, some important insights that came from these descriptions that could prove valuable to developing a normative analysis of technology.

The work of Langdon Winner (1977, 1980) investigating whether technology was advancing autonomously from human control, or independently of social interests, introduced the idea that "things" have politics. By this he meant that the political interests of one group could be imposed upon others through material technologies and the design of social spaces. Often these politics leverage the material power of technological design, as when bridges are built too low for public buses to reach the most desirable beaches, thus keeping out the social classes that depend on public transit, or when obstructions are added to park benches to discourage unhoused people from lying down and sleeping on them. But Michel Foucault (1977) also provided numerous examples of social and psychological control through material technologies like architecture. In particular, the design of the panopticon amplifies the power of prison guards through providing a central guard tower wherein a small number of prison guards can instill in a large number of prisoners a sense of being watched over. That sense of being observed serves to discourage individuals from acting out or organizing a revolt or riot, even though the small number of guards could be easily overpowered by the prisoners if they acted together. Much of Foucault's work served to reveal these so-

cial mechanisms of the psychological control of individuals, however, rather than to further investigate how material technologies were being designed to amplify these effects.

Perhaps the most sophisticated critiques of technology came from the analysis of the narrow but important area of technological impacts on labor. Marx's analysis of labor showed how the owners of factories used machines to both increase productivity and deskill workers ([1867–94] 1959). Technology thus alienated workers from the products of their labor. Technologies were introduced to legitimize the shift of psychological norms and legal claims to ownership of the products of labor from the skilled craftsman to the capitalist. Even before that, the Luddite movement in the English garment industry in the early nineteenth century had recognized this process, and workers set about destroying the machines, mainly weaving looms and other fabric and garment machinery, which were the most exploitative of their labor, even as they spared the machines that did not seek to undermine their skills or further alienate their labor (Sadowski 2021). This line of critique was extended in the early twentieth century with analyses of Taylorism and still continues within the labor movement, which itself seems to be getting a fresh look in the current era of technological unemployment, alienation, and the political and social power of Big Tech and AI.[2] Still, this critique largely comes down to a critique in which the interests of owners and management are set against those of workers, and the political struggle is always and forever between these two forces. In reality, sometimes those interests align, and often there are many factions and differences of interests within those groups. Moreover, these critiques largely remain silent on the technological products that a business makes or on the interests of the greater society in which those technologies might function.

Another thread of the politics of technology grew from Habermasian theories of justice through democratic participation and discourse. Work in the critical theory of technology (Feenberg 1991) sought ways to embed democratic processes of participation into the

adoption of new technologies. Most examples from that time involved public hearings about large projects like powerplants and pipelines or citizen science to produce alternative analyses of the potential impacts of such projects. Researchers in participatory design and computer-supported cooperative work took the Marxist critique of workplace technologies further to include democratic participation in the design of new technologies (Asaro 2000). These researchers applied anthropological and ethnographic techniques to studying workers and the impacts of IT on the workplace (Zuboff 1988). By doing so, they recognized the ways social relations, labor relations, and power relations were all being transformed by the introduction of various IT systems and the design decisions they embodied. These efforts were at the forefront of understanding the need for a social critique of technology that had teeth—which could both describe the politics of things and point to preferable alternatives. Participatory design, in particular, began by trying to provide labor unions with technological designs that would empower workers, rather than deskill them.

These efforts fell somewhat short of the goal of providing a fully normative assessment of technological systems or design decisions. These had to be relativized to the interests of workers or their unions, and norms had to be determined by reverting to the processes of procedural democratic participation as the ultimate measure of justice, rather than attempting to define or evaluate the desirability of a given technology according to objective criteria. At best, such democratic participation might make technology responsive to some of the needs and interests of those participating in the design. At worst it fixed the design of technologies to the interests of what would inevitably be a small group of people, even while technologies were having an ever-greater impact on every aspect of life and every member of society. And while artifacts were seen to have politics and the process of design became a forum for democratic discourse, technological designs were still seen as static, not dynamically evolving in their material efficacy, their social uses, and their shifting political implications.

ENGINEERING ETHICS AND PUBLIC SAFETY

Within engineering itself, there has been a long history of engineering ethics. It emerged within professional engineering societies in the United States following a series of spectacular bridge collapses and dam failures in the late 1800s and early 1900s (Wikipedia 2023a). It started with the adoption of "codes of ethics" within these engineering societies and led to certification requirements for engineers, following the 1919 Great Molasses Flood in which a massive storage tank in Boston ruptured, releasing a fast-moving tidal wave of molasses that killed 21 people (Wikipedia 2023b). In its most common and basic form, engineering ethics aims to ensure that engineers do not develop or approve projects that they know could pose a risk to society, or use elements that do not meet the required specifications, or otherwise cut corners that introduce unacceptable risks (such as signing off on plans for a bridge that uses cheaper construction methods but risks catastrophic collapse, or whether the engineers who were pressured to allow the launch of the Space Shuttle Challenger should have done more to stop it when they knew the temperatures were too cold for the booster rocket O-rings [Perrow 1984]). Essentially, engineering ethics asks the engineer to consider the safety and social costs alongside the construction costs of various design alternatives and to not sacrifice safety even when there are economic or other pressures to do so. Ultimately, however, it assumes that the engineer is in fact aware of the risks, can accurately measure and compare them, and has the power to remedy them.

This tends to work well for things like structural engineering and physical safety but is not as effective in anticipating social and cultural harm or political implications—so the bridges will not collapse, but they may exclude public transit (Winner 1980). There are also obvious limits on the number of designs engineers might consider and on their creativity in imagining alternatives. Later efforts to reform engineering ethics have focused on trying to expand the design alternatives considered by engineers. It is not entirely fair to

expect engineers to assess designs that they have not considered, so how do we broaden the range of designs they do consider?

Human-centered design and value-centered design are the best examples of this trend, in that they ask engineers to not only consider human values in the costs and benefits of various designs but also actually develop designs with certain human-centered concerns or values as priorities, long before they get to the risk analysis of a set of design options. These methods also share a methodology with participatory design and other politically progressive approaches to technology, namely a brainstorming element. Simply, they ask individual engineers or engineering and design teams to conduct a deliberate exercise in imagining potential harms and then make efforts to mitigate these harms during their design process. They do not do much to guide that process normatively nor to provide any external forms of evaluation, accountability, or mechanisms for reforming designs once implemented.

For much of the history of both engineering ethics and STS, the focus has been on material technologies—transportation, medicine, materials, buildings, and so on. The development of computational technologies was largely seen as neutral, apart from the data in such systems, while the data was seen as very much subject to human social concerns, whether it was governmental census data, business and financial data, medical data, or demographic or personal data. Of course, when it comes to data, humans define the categories and metrics qualitatively, even if the ultimate product is a quantitative dataset (Bowker and Star 2000). But it has proven difficult to challenge the social biases and political interests that shape datasets, which in turn shape institutional decisions that are made about people and the distribution of resources.

This has been a key point of contention as the public becomes more concerned about the data being used to train AI systems. With the rise of the internet and social media, computer networks also subsumed mass media, with information shaping not only factories

and offices, but also public opinion, education, knowledge, and epistemic justification across society. Not only are social relations of work plastic and reconfigurable, but so are social relations in general, as well as the public consciousness. While twentieth-century studies of information systems focused primarily on flows of information, they did little to critique the data upon which these systems operated. A number of books published from 2016 on (O'Neil 2016; Eubanks 2017; Hicks 2017; Noble 2018; Benjamin 2019) and the ACM Fairness, Accountability and Transparency (FAccT) conferences that began in 2018 were instrumental in initiating an explicit political discourse around data bias. At best, real-world datasets represented a world of structural inequalities—racism, sexism, ableism, ageism, and more—and their use in decision-making systems served to entrench the status quo, if not amplify and exacerbate the inequalities inherent in data. And thus data itself became a matter for political struggle.

WHAT IS AI?

Artificial intelligence has multiple meanings. The term was coined by mathematician John McCarthy for a conference at Dartmouth College in the summer of 1956. AI was conceived of as the design of "thinking machines." This idea had been percolating for more than a decade within a scientific movement called cybernetics, which sought to explain the human brain and other complex systems with mathematical models and to simulate those models in machines. This same group had also been instrumental in the design of early digital computers and very much saw them as operating in some sense like a brain (Asaro 2006, 2011). By the 1950s digital computers and programming had developed to such an extent that AI was framed as developing computer programs that could perform tasks that were assumed to require an intelligent human to perform. As many of these researchers were mathematicians or engineers, what they considered "intelligent tasks" was mostly solving math problems, puzzles, and mazes, or playing rule-based games like checkers and chess. Historically, as various types of mathematical problems were

solved by AI researchers, these tasks were no longer seen as AI challenges, or even as AI.[3] As such, AI is not any particular technology, nor is it any static set of automated tasks. By definition it is evolving. In this sense it is better to think of it as a process by which tasks performed by humans are rendered as computational problems and solved/performed by computers. It is automation for the information age, though it also feeds back into robotics and industrial automation to further automate many tasks not already mechanized in the industrial age (e.g., the Amazon warehouse).

As it continued to develop in the 1960s and 1970s, AI aligned itself with work in computational linguistics, psychology, and neuroscience, leading to what came to be called the "cognitive revolution" in psychology. One vision of AI was that by programming computers to solve problems that usually required brains, we could better understand the human brain and psychology. Some thought we could use computers to model or simulate the brain, while others thought of AI as a form of engineering that could solve useful problems, regardless of any resemblance to or insights from natural systems. Historically speaking, we can also look at AI as a scientific and engineering discipline consisting of a set of techniques, canonical research, and professional practice. While all these aspects continue to play out in contemporary AI, it is the "AI as engineering" thread that has dominated recent work and led to increased public awareness of AI and its potential impacts on society.

From a technical perspective, there are many different approaches to AI. All of them, however, involve building computational models that represent, to some degree, an aspect of the world and allow computational operations on those representations, which can then be applied to the real world. Many different mathematical techniques have been used in computational modeling more generally, and AI has borrowed mathematical models from many different disciplines and application domains—from diffusion models in chemistry to structural models in architecture and civil engineering to financial models of markets, and many more. Most fundamental AI research,

however, has focused on two broad mathematical approaches, logic and statistics, which at various points have competed for research money and at others have been combined in various hybrid systems. Early AI focused more on logical techniques, which worked well for many mathematical puzzles and problems, especially with the limited computational power and structured datasets available at the time, but also proved rigid compared to the flexibility of human intelligence.

Within AI and cybernetics there were also threads that were focused on learning and biological adaptation and simulating these in machines (McCulloch and Pitts 1943). This developed into the AI subfield of machine learning, including the development of brain-inspired mathematical graphs called neural networks. These threads have also had a complex relationship to biological models and understanding of the human brain, sometimes drawing direct or loose inspiration from the brain and sometimes informing neuroscientific research (Asaro 2011). The development of statistical models began in the 1960s as perceptrons and evolved from parallel distributed processing in the 1970s into neural networks in the 1980s and 1990s. But it wasn't until the 2010s that a combination of massive computing resources (data centers and cloud computing) and the availability of massive datasets came together to make "deep neural networks" / "deep learning" possible. Both the data and the cloud computing were consequences of the internet and the availability of data collected from, and often about, the internet and its users, along with the need for giant data centers to keep the internet running. Deep learning was able to fulfill the promise of neural networks by allowing the creation of statistical models with millions, billions, and now trillions of parameters to be built and trained on enough real-world data, and for enough repetitions, to actually "tune" those trillions of parameters. The resulting models are thus able to perform in ways that replicate real-world systems without reverse engineering those systems or explicitly designing the network. They are essentially compact statistical models of real-world data, structured in a way that they can be used to predict real-world outcomes and thus solve problems and

perform tasks that once required human intelligence. When applied to linguistic datasets, containing almost all the texts readily available to be scraped from the internet, these techniques have been used to build what are called large language models (LLMs), with ChatGPT being only the most famous of these.

Many people talk about AI, or AIs, as if they were conscious thinking entities—artificial minds. While this has long been a goal of AI developers, it is unclear what would be required to achieve it, and philosophical debates abound around artificial consciousness, self-awareness, sentience, and life. In reality, AI is just software running on computers. To some extent this debate has shifted to concerns around artificial general intelligence, or AGI, which is conceived of as a program that can learn any task—or the full range of tasks a normal human is capable of (aka human-level intelligence). While neither is yet technologically feasible, there is also nothing that appears to preclude such a technological capability in the future. The main fear is that humanity might not be able to control such a technology if it is indeed much "smarter" than us and has better access to more information than we do—a possible superintelligence. But most of these discussions are based more on science fiction, or belief in almost magical abilities of AI, than on current technology or our scientific understanding of intelligence.

For the purpose of understanding AI ethics, it is best to think of AI as the process of automating human intelligence—perception, judgement, decision-making, and the performance of complex tasks. Not only is this an accurate way to look at it, but it also captures many of the concerns that experts and the public have about the potential risks and dangers of AI. While certain techniques and applications, like deep learning or LLMs, may have social, political, or ethical implications in a specific context, in general AI is not any specific technique or technology. Another way to look at it is as the design and implementation of software and technological systems that shape and will eventually perform many or most of the decisions and actions of businesses, institutions, governments, and individuals as they come

to rely on those systems. The fear of AI is thus also a fear of automation that goes beyond replacing manual labor and replaces intellectual labor. It represents a fear of a technocratic society in which human decisions, actions, and accountability are replaced with automation that cannot be appealed to or held accountable. As such, it is possible to view AI as the culmination of centuries of automation, as the process by which nearly all conceivable social, political, and economic relations may be reorganized through technological transformation and thus *the technology* that will have the greatest role in shaping our social, economic, and political lives going forward. This duality at the heart of AI also shapes both how Silicon Valley and AI proponents approach AI ethics and potential regulation, and how social critics of AI approach AI ethics.

THE RISE OF AI ETHICS

AI ethics is not really a distinct field or coherent discourse, but more of an amalgamation of different perspectives considering the potential implications of automated systems and algorithms making decisions with consequential impacts on human lives. Indeed, it may even be a misnomer insofar as many of the concerns are matters of social value or justice, rather than individual ethics. But there is value in considering how and why different stakeholders have developed different conceptions and labeled them AI ethics in order to suit their interests, as well as the ways in which these conceptions have illuminated important questions in the social critique of technology more generally. Indeed, I believe the divergent visions of AI ethics is the best explanation for the success of the term and the activities performed under it, despite its overall incoherence. The various stakeholders all have different commitments to ethical theories (mostly Western), as well as differing political and economic interests in AI technology and its application. These have mostly fixated on a few of the key social problems raised by AI, as well as efforts by technology companies to avoid regulation.

Historically, AI ethics also had its roots in machine ethics and robot ethics (Asaro and Wallach 2017). The earliest work in AI ethics was primarily concerned with abstract philosophical problems in metaethics, such as whether computers could be programmed to make ethical decisions, or whether ethical theories like utilitarianism are computationally tractable, or whether sufficiently complex or capable AI systems should be considered to have rights. Within software engineering more generally there was software ethics that grew directly out of the tradition of engineering ethics described above. This mostly dealt with managing "safety critical" software systems that could directly impact human lives and to some extent with growing concerns around data privacy and system security. Principles emerged such as warnings that software that was unstable or imprecise should not be entrusted with human lives (such as medical systems, flight control systems, nuclear powerplant control systems, and other safety critical systems) or that systems that contained personal or valuable information should have security mechanisms in place to prevent unauthorized access to such data (such as security by design).

These issues came more into public awareness in the 2010s with the promise and testing of self-driving cars, in which software is used to control technologies already responsible for killing tens of thousands of people every year and as data breaches rendered more and more people's information vulnerable to identity theft and other harms. Such concerns led to the revival of an old tool from metaethics, the trolley problem (Cowls 2017), a thought experiment on whether it is morally better to act to save several people by killing one person, or to simply allow several people to die through inaction. It was also used to show that your moral intuitions could be easily shifted or manipulated by adding additional information to the situation—whether to save a young person over an old person, or a "good" person over a "bad" person, and such. Though the new applications of the trolley problem were largely a distraction and misunderstanding of the original purpose of this philosophical thought experiment, it

raised awareness that autonomous systems making decisions based on algorithms would have to make ethically difficult decisions and that we should have public and professional discussion of how those decisions should be made and establish policies to regulate the process.

In 2013, Google entered discussions to acquire the UK-based machine learning company DeepMind. One stipulation of that acquisition, insisted on by DeepMind's three founders, was the creation of an internal ethics review board within Google (Legassick and Harding 2017). While few details about this review board have been made public, it was intended to be an integral part of Google, not just the DeepMind subsidiary. The only leaked information about the board as of 2017 (Hern 2017) was that it had convened, was bringing ethicists from other areas of science and technology "up to speed" on advances and future directions of AI technology, and included Nick Bostrom, the Swedish author of *Superintelligence* (2014). These tidbits suggest that the board was tasked primarily with considering the ethical issues stemming from advanced AI, AGI, or superintelligence, and the long-term threats to humans from creating an AI that is autonomous and beyond human control. This is somewhat ironic because Google was by that time already one of the pioneers of surveillance capitalism, and using data to target advertising to users of its search engine was its main business. The ethics of doing this would likely not be a topic for discussion by the ethics board.

A great deal of digital ink has been spilled discussing the potential long-term consequences of AI and the almost mystical properties of AI that transcend human intelligence (Asaro 2001). Some people working in AI ethics argue that these are the most important ethical issues because AI could lead to the end of human civilization or to AI-beings that go forth to colonize the solar system, the galaxy, or beyond (Bostrom 2014; Russell 2019). These ideas have gained a great deal of traction in Silicon Valley more generally. Most AI ethics researchers, however, are more concerned with near-term risks to society, jobs, and individuals. Indeed, the focus on far-off consequences, or "longtermism," is largely a distraction from the already existing

threats to human rights, fairness, justice, equality, and safety, and other more immediate, tangible, and avoidable risks. In this sense, longtermism is itself a form of "ethics washing" because it leads to avoiding consideration of these pressing issues in favor of considering problems with little impact on current business practices.

While the creation of such an ethics board at Google was innovative at the time, the secrecy around it was puzzling. One of the main functions of such boards, like institutional review boards for scientific research, is both to avoid potential legal liability in case things go wrong and to assure the public and/or customers and/or investors that there is meaningful oversight of any potentially harmful research, as well as to put in place internal processes for reducing the risks of doing such work. However, for it to fully function in these ways, particularly to build public trust, some degree of transparency is required, and even better would be public announcements of some major decisions made by the board. At the very least, an announcement of who is on the board could build confidence based on the reputation of the members, even if their actual authority and power are unknown or limited. Other companies followed suit in creating such boards, with some being more transparent, like the Texas startup Lucid.AI. Other Big Tech companies like Microsoft created AI ethics groups in 2017, only to disband them in 2023 (Belanger 2023), while Meta created an oversight board for ethical questions around content management on its social media platforms Facebook and Instagram (Wong and Floridi 2023).

After that, a series of companies issued "AI ethics principles." These were mostly abstract lists of ethical principles that companies touted to assure customers and the public that they would only build useful and beneficial AI and would avoid building anything that might harm society. Google led this trend with a set of principles that were released during an employee protest over the company's participation in Project Maven (Google n.d.), a Pentagon military project to use AI to analyze the vast quantities of video data collected by drones in the US military operations in Iraq and Afghanistan. Accordingly,

Google's principles included a principle to not develop "weapons" but did not go as far as to say they would not work for militaries nor make an effort to keep their technologies from being used as a component in some future weapon or in helping choose targets for weapons (Godz 2018; Suchman, Irani, and Asaro 2018). Other companies followed suit, and within a few years, there were well over 100 sets of "AI ethical principles" (Winfield 2019).

These sets of principles served the interests of technology companies to "ethics wash" their images and products by showing that they cared about developing the technology ethically. However, similar to the establishment of the ethics boards, there was little transparency about how the rules would be applied or interpreted. Even Google's CEO would not answer whether Project Maven met the criteria for being a component of a weapon system given the stated principle not to work on weapons, and while Google's leadership claimed they would not renew the Pentagon contract, they did not cancel it but continued working on the project for some time, as perhaps they still are. Intrinsically, any set of ethical principles, like the engineering codes of conduct, are abstract and require interpretation when being applied to concrete problems and cases. They are also nonbinding and unenforced; no one can really review how they are being applied or hold companies accountable if they are violated. Moreover, they do not really offer much guidance on how to design software or what kinds of considerations should influence design choices. More recent efforts by standard-setting bodies of professional organizations like the Association of Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE) developed more robust ethical principles meant to guide programmers and engineers working on AI and software projects, but these too are nonbinding (IEEE 2017; ACM 2018).

The ethics washing efforts were clear attempts to avoid any real legally binding regulation of the industry. There is yet no such regulation in the US, and the EU is just implementing its first attempt at creating a dedicated AI regulatory framework, while a few AI-relat-

ed clauses exist in previous regulations such as the EU's General Data Protection Regulation.

The other thread of AI ethics that emerged in the 2010s was a series of technological critiques and sociological analyses of technology that examined its differential effects on different people and groups. Namely, many technologies were being developed that imposed draconian social policies on the poor, discriminated against already marginalized groups based on race, age, class, gender, skin color, or religion, or otherwise showed biases in their effects on different groups. We can describe the various efforts in this thread as aiming for "data justice" or "algorithmic justice." Examples of this include the ways the implementation of AI and algorithmic technologies can be racist, sexist, or classist (see Asaro 2016; O'Neil 2016, 2023; Eubanks 2017; Hicks 2017; Noble 2018; Benjamin 2019). These trends were further explored in a more technical way through a series of annual ACM conferences on fairness, accountability, and transparency in algorithms that began in 2018. These conferences examined the ways existing social biases in datasets are replicated, amplified, and exacerbated by learning algorithms, and looked into developing strategies to detect, eliminate, and minimize these biases. For example, an algorithm for determining the salary to offer a new hire might be based on past salaries, which we know tend to be less for women than men and less for women of color than for white women. Such algorithms might then identify the gender and race of a job candidate and, based simply on these characteristics, offer them a lower salary than they would offer another candidate with other qualifications being the same. The algorithms would be simply identifying the historic trend and continuing it going forward. How to get them to *not* do that is a difficult and important area of research.

REFLECTIONS ON AI ETHICS FOR SOCIAL CRITIQUE AND TECHNOLOGY REGULATION

The rise of AI ethics has been fortunate in some ways and unfortunate in others. It has drawn a great deal of awareness to the social

issues raised by AI and rapid technological innovation more generally. It has also clearly been used by technology companies to avoid, or at least delay, governmental regulation and public scrutiny. Research into data justice and the fairness, accountability, and transparency of algorithms has perhaps made the greatest contribution in terms of focusing on social justice issues raised by AI and algorithmic technologies. All these are positive developments toward both a social critique of technology and its democratic regulation. But there remain significant weaknesses in terms of bringing these together into a coherent critique of technology that centers on the power dynamics of technology or offers a path forward in both the regulation and the progressive design of technology.

Perhaps the greatest weakness in this history is its reliance on "ethics" as a critical framework. Ethics is indeed a valuable normative theory, but it is not the only one. Indeed, much of the most valuable work in this area identifies its goal as "justice" or "fairness" rather than a strictly ethical or moral form of "the good." Even there, the goal seems to be limited to a distributive notion of justice, according to which there is equal opportunity or fair distribution based on merit. Typically, we think of ethics as reflecting the moral character, or decisions and actions, of an individual. Even if we apply that to organizations like the companies that build AI systems and adopt ethical principles, the moral character and motivations of a corporation are not our primary concern. Rather it is the social consequences of the technology that are of real concern. The "good" company might not make technology that is good for society. Most of the technologies harming society are already made by mostly good people—and by companies that believe their products are good for society. Moreover, a company or its engineers might not even be in a position to say what is good for society; that in itself is something that needs to emerge from an inclusive and democratic process of setting goals and establishing and prioritizing social values. It also requires independent observation, auditing, and normative assessment of social impacts. And ultimately, it needs the authority to impose sanctions

and restructure incentives in order to shape the development, implementation, use, and refinement of technologies. This is far beyond what ethics can be expected to do.

Historically, ethics concerns individuals. It has provided useful and powerful tools for thinking about the actions of an individual and their rights and duties with respect to others. But this approach has great difficulty dealing with social-level phenomena, from tragedies of the commons to challenging the status quo. There have been efforts to develop communitarian ethical theories, which are grounded in the community rather than an individual, but these are marginalized in contemporary philosophy and not often used in AI ethics. There has also been an effort to promote an ethics of care in robotics and AI ethics (van Wynsberghe 2013; Asaro 2019). The ethics of care is relational, rather than individualistic, and puts the care relationship, rather than individual actions or intentions, as the focus and key metric in normative assessment. As such it can give concrete normative guidance to assessing technological designs as well as their regulation.

If we focus instead on justice, there are clear paths toward eliminating forms of structural inequality. It has been argued that we need a new social contract, a bill of rights or even a Geneva Convention–type treaty for technology—evoking historical examples of collective agreement on just governance (Smith 2017; Caron and Gupta 2020; AIWS 2020; White House 2022). Theories of justice, such as the social contract, mainly look to justify the establishment of collective government (legitimacy) and the sublimation of our personal interests in order to attain collective respect for individual rights. But we still do not have a clear and comprehensive theory for thinking about how power is exercised and shifted through these systems or how we might collectively shape the development of these systems toward creating a more just, inclusive, and sustainable future. Without this, what should the content of a new technological contract or bill of rights contain? What is needed is to regulate corporations, markets, industries, and nations at a global scale (Verdegem 2021), not to stipulate an individual's sense of obligation or virtue. While moral

approval and disapproval, in the form of social norms, can be very powerful in shaping behavior, what should really concern us with AI going forward is *power*. And we need legal instruments to implement that power, as well as social critiques to give it content.

Indeed, any reliance on human values or social norms when dealing with a radically transformative technology like AI is fraught. AI, like a few other technologies such as bioengineering and synthetic biology, has the power to change who we are as human beings. It could shift our fundamental aesthetic and ethical values as well as social norms—what it means to make art, or what we owe to each other, or how we understand ourselves and the human condition. As these systems collect and analyze vast amounts of data about us, they will create new categories of people—categories that we do not yet see or recognize. We are likely to shape our own personal and social identities according to membership in those new categories (Hacking 1986). We will, in effect, become new types of humans in accordance with these AI programs. How are we to assess these transformations, and what do we owe to the new types of people our technologies will create? Moreover, how can we try to aim these processes toward a collectively more desirable world?

WHAT DO WE NEED INSTEAD?

Building on the long history of the labor critiques of technology, we need to extend these to other aspects of life and the ways in which they are impacted by technology. But we also need a theory of power that does not presuppose fixed social distinctions and adversaries or simplistic solutions. We also need to expand that critique to complex and dynamic social groups and to understand the exercise of technological power in complex ways. In this, we can also build upon STS and work in socio-technical systems in order to understand the ways sociality and technology are intertwined, dynamic, and emergent. In terms of finding our moral and political bearings, we can look to the social contract, theories of democratic participation, and participatory design. But we also need to recognize the ways democracy can

allow the majority to marginalize and subjugate minorities and how the formal processes of representational democracy can be taken over by a minority to subjugate the majority.

In some sense, we can still look to the social contract as a model, but we need a social contract with technology—and social control over its design, implementation, and reform. Governments might have the power to do this, but our current means of representational democracy are woefully inadequate for dealing with the complexity and rapid rate of change in technological systems. To the extent that these technologies develop and transform society rapidly, we also need more direct and responsive ways to reform it. Technology is also more fluid and mercurial than states; it does not respect borders and is built up in integrated modules, overlapping layers, and emergent powers that cannot be easily separated or controlled. Unlike the state, technology does not always have clear citizens and sovereigns, users and designers. Technologies are coproduced and have many indirect effects and consequences, and often we are unaware of what technologies are touching our lives.

We stand at the cusp of a technological revolution with the potential to permeate and transform our social and political world to a far greater extent than previous technological revolutions have done. Insofar as we can apply lessons learned from those technological revolutions and improve on our processes of innovation and mechanisms of regulation, we have an opportunity to shape this emerging revolution toward desirable social goals. Failing to do so would mean abdicating responsibility and allowing powerful political and commercial interests, and the technocratic values held by engineers, to shape the social values and political goals that are built into these technologies and that will reshape the social and physical world according to those values. This, for me, is the most critical and challenging frontier in social inquiry and practice today—and for the foreseeable future. There will not be a simple theory or regulatory fix that resolves it. But if we have a critical theory of technology that centers power and has a

moral compass guided by shared values, we can guide regulation and shape innovation toward a better future.

NOTES

1. One has to look no further than the Facebook motto, often applied to Silicon Valley more generally, of "Move fast and break things" to see that the point of many software companies is to "disrupt" existing social and business practices and relations in order to insert a new piece of software technology, which becomes both socially necessary and economically beneficial to the software company. Companies like Uber and Lyft disrupt the taxi business by making their app the social norm for hailing a ride in a hired car, while simultaneously redirecting taxi revenue from existing taxi companies around the world to their own company, all while avoiding actually investing capital in cars and their maintenance as a traditional taxi company must do.
2. For example, the current (at the time of writing) labor strikes by the Screen Actors Guild, Writers Guild of America, and Directors Guild of America are putting the issue of their labor being automated by AI (which is also being trained on data scraped from their previous work) as a top issue in their contract negotiations (Watercutter 2023). And new forms of labor organizing are focusing on the technological products workers are instructed to make, rather than just their working conditions, such as protests at Google over their work on the Pentagon's Project Maven to improve drone strike targeting (Godz 2018).
3. For instance, finding the most efficient route through a map or graph was a problem studied by early AI researchers. Yet today, when this task is performed by Google Maps, most people do not think of this as AI, but as simply the application of an efficient route-finding algorithm. Similarly, speech recognition and optical character recognition were problems long studied in AI, but these are now seen as simply computer programs for those tasks.

REFERENCES

ACM. 2018. "ACM Code of Ethics and Professional Conduct." Association of Computing Machinery. https://ethics.acm.org/.

AIWS. 2020. "A Social Contract for the AI Age." AI World Society, Sept. 9. https://bostonglobalforum.org/wp-content/uploads/Social-Contract-for-the-AI-Age-8-9-2020-official.pdf.

Asaro, Peter M. 2000. "Transforming Society by Transforming Technology: The Science and Politics of Participatory Design." *Accounting, Management and Information Technologies* 10 (4): 257–90.

Asaro, Peter M. 2001. Review of *Mere Machine to Transcendent Mind* by Hans Moravec. *Minds and Machines* 11 (1): 143–47.

Asaro, Peter M. 2006. "On the Origins of the Synthetic Mind: Working Models, Mechanisms, and Simulations." PhD diss., U. Illinois at Urbana-Champaign.

Asaro, Peter M. 2011. "Computers as Models of the Mind: On Simulations, Brains and the Design of Early Computers." In *The Search for a Theory of Cognition: Early Mechanisms and New Ideas*, ed. Stefano Franchi and Francesco Bianchini, 89–114. Amsterdam: Rodopi.

Asaro, Peter M. 2016. "Will #BlackLivesMatter to RoboCop?" Paper presented at WeRobot: Conference on Legal and Policy Issues Relating to Robotics, U. Miami School of Law, April 1–2. https://robots.law.miami.edu/2016/wp-content/uploads/2015/07/Asaro_Will-BlackLivesMatter-to-Robocop_Revised_DRAFT.pdf.

Asaro, Peter M. 2019. "AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care." *IEEE Technology & Society Magazine* 38 (2): 40–53.

Asaro, Peter M., and Wendell Wallach. 2017. "An Introduction to Machine Ethics and Robot Ethics." In *Machine Ethics and Robot Ethics*, ed. Wendell Wallach and Peter M. Asaro. New York: Routledge.

Belanger, Ashley. 2023. "Amid BING Chat Controversy: Microsoft Cut a Key AI Ethics Team." *Ars Technica*, March 14. https://arstechnica.com/tech-policy/2023/03/amid-bing-chat-controversy-microsoft-cut-an-ai-ethics-team-report-says/.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge, MA: Polity.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford U. Press.

Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Caron, Mika Snyder, and Abhisek Gupta. 2020. "The Social Contract for AI." arXiv, June 15. https://arxiv.org/abs/2006.08140.

Cowls, Josh. 2017. "AI and the 'Trolley Problem' Problem." *Medium*, Aug. 9. https://medium.com/josh-cowls/ai-and-the-trolley-problem-problem-ef48582b49bf.

Eubanks, Virginia. 2017. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Feenberg, Andrew. 1991. *Critical Theory of Technology*. New York: Oxford U. Press.

Foucault, Michel. 1977. *Discipline and Punish: The Birth of the Prison*. New York: Random House.

Godz, Polina. 2018. "Tech Workers versus the Pentagon: An Interview with Kim." *Jacobin*, June 6. https://jacobin.com/2018/06/google-project-maven-military-tech-workers.

Google. n.d. "Google AI Principles." https://ai.google/responsibility/principles/.

Hacking, Ian. 1986. "Making Up People." In *Reconstructing Individualism: Autonomy, Individuality, and the Self in Western Thought*, ed. Thomas C. Heller, Morton Sosna, and David E. Wellbery, 222–36. Stanford, CA: Stanford U. Press.

Hern, Alex. 2017. "Whatever Happened to the DeepMind AI Ethics Board Google Promised?" *Guardian*, Jan. 26. https://www.theguardian.com/technology/2017/jan/26/google-deepmind-ai-ethics-board.

Hicks, Marie. 2017. *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing*. Cambridge, MA: MIT Press.

Hilgartner, Stephen. 1997. "The Sokal Affair in Context." *Science, Technology, & Human Values* 22 (4): 506–22.

IEEE Standards Association. 2017. "Ethically Aligned Design—Version II." https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.

Latour, Bruno. 1988. *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard U. Press.

Latour, Bruno. 1996. *Aramis, or the Love of Technology*. Cambridge, MA: Harvard U. Press.

Legassick, Sean, and Verity Harding. 2017. "Why We Launched DeepMind Ethics & Society." Google DeepMind, Oct. 3. https://www.deepmind.com/blog/why-we-launched-deepmind-ethics-society.

Marx, Karl. [1867–94] 1959. *Capital: A Critique of Political Economy*. Chicago: Regnery Publishing.

McCulloch, Warren S., and Walter H. Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115–33. http://dx.doi.org/10.1007/BF02478259.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York U. Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

O'Neil, Lorena. 2023. "These Women Tried to Warn Us about AI." *Rolling Stone*, Aug. 12. https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/.

Perrow, Charles. 1984. *Normal Accidents: Living with High-Risk Technologies*. Princeton, NJ: Princeton U. Press.

Pickering, Andrew. 1995. *The Mangle of Practice: Time, Agency, and Science*. Chicago: U. Chicago Press.

Pinch, Trevor J., and Wiebe E. Bijker. 1984. "The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other." *Social Studies of Science* 14 (3): 399–441. https://doi.org/10.1177/030631284014003004.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin.

Sadowski, Jathan. 2021. "I'm a Luddite. You Should Be One Too." *The Conversation*, Aug. 9. https://theconversation.com/im-a-luddite-you-should-be-one-too-163172.

Smith, Brad. 2017. "The Need for a Digital Geneva Convention." Microsoft blog, Feb. 2. https://blogs.microsoft.com/on-the-issues/2017/02/14/need-digital-geneva-convention/.

Suchman, Lucy, Lilly Irani, and Peter M. Asaro. 2018. "Google's March to the Business of War Must be Stopped." *Guardian*, May 16. https://www.theguardian.com/commentisfree/2018/may/16/google-business-war-project-maven.

Van Wynsberghe, Aimee. 2013. "Designing Robots for Care: Care Centered Value-Sensitive Design." *Science and Engineering Ethics* 19 (2): 407–33.

Verdegem, Pieter. 2021. "Dismantling AI Capitalism: The Commons as an Alternative to the Power Concentration of Big Tech." *AI & Society*. https://doi.org/10.1007/s00146-022-01437-8.

Watercutter, Angela. 2023. "The Hollywood Actors Strike Will Revolutionize the AI Fight." *Wired*, July 14. https://www.wired.com/story/hollywood-sag-strike-artificial-intelligence/.

White House. 2022. *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. White House Office of Science and Technology Policy. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

Wikipedia. 2023a. S.v. "Engineering ethics." Last edited Aug. 7, 16:46. https://en.wikipedia.org/wiki/Engineering_ethics.

Wikipedia. 2023b. S.v. "Great Molasses Flood." Last edited Sept. 3, 19:09. https://en.wikipedia.org/wiki/Great_Molasses_Flood.

Winfield, Alan. 2019. "An Updated Round up of Ethical Principles of Robotics and AI." *Robohub*, May 12. https://robohub.org/an-updated-round-up-of-ethical-principles-of-robotics-and-ai/.

Winner, Langdon. 1977. *Autonomous Technology*. Cambridge, MA: MIT Press.

Winner, Langdon. 1980. "Do Artifacts Have Politics?" *Daedalus* 109 (1): 121–36.

Wong, David, and Luciano Floridi. 2023. "Meta's Oversight Board: A Review and Critical Assessment." *Minds and Machines* 33: 261–84.

Zuboff, Shoshana. 1988. *In the Age of the Smart Machine: The Future of Work and Power*. New York: Basic Books.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.