AI Now Report 2018

Meredith Whittaker, AI Now Institute, New York University, Google Open Research
Kate Crawford, AI Now Institute, New York University, Microsoft Research
Roel Dobbe, AI Now Institute, New York University
Genevieve Fried, AI Now Institute, New York University
Elizabeth Kaziunas, AI Now Institute, New York University
Varoon Mathur, AI Now Institute, New York University
Sarah Myers West, AI Now Institute, New York University
Rashida Richardson, AI Now Institute, New York University
Jason Schultz, AI Now Institute, New York University School of Law
Oscar Schwartz, AI Now Institute, New York University

With research assistance from Alex Campolo and Gretchen Krueger (AI Now Institute, New York University)

DECEMBER 2018

AINOW

CONTENTS

ABOUT THE AI NOW INSTITUTE

The AI Now Institute at New York University is an interdisciplinary research institute dedicated to understanding the social implications of AI technologies. It is the first university research center focused specifically on AI's social significance. Founded and led by Kate Crawford and Meredith Whittaker, AI Now is one of the few women-led AI institutes in the world.

AI Now works with a broad coalition of stakeholders, including academic researchers, industry, civil society, policy makers, and affected communities, to identify and address issues raised by the rapid introduction of AI across core social domains. AI Now produces interdisciplinary research to help ensure that AI systems are accountable to the communities and contexts they are meant to serve, and that they are applied in ways that promote justice and equity. The Institute's current research agenda focuses on four core areas: bias and inclusion, rights and liberties, labor and automation, and safety and critical infrastructure.

Our most recent publications include:

- **Litigating Algorithms**, a major report assessing recent court cases focused on government use of algorithms
- **Anatomy of an AI System**, a large-scale map and longform essay produced in partnership with SHARE Lab, which investigates the human labor, data, and planetary resources required to operate an Amazon Echo
- **Algorithmic Impact Assessment (AIA) Report**, which helps affected communities and stakeholders assess the use of AI and algorithmic decision-making in public agencies
- **Algorithmic Accountability Policy Toolkit**, which is geared toward advocates interested in understanding government use of algorithmic systems

We also host expert workshops and public events on a wide range of topics. Our workshop on **Immigration, Data, and Automation in the Trump Era**, co-hosted with the Brennan Center for Justice and the Center for Privacy and Technology at Georgetown Law, focused on the Trump Administration's use of data harvesting, predictive analytics, and machine learning to target immigrant communities. **The Data Genesis Working Group** convenes experts from across industry and academia to examine the mechanics of dataset provenance and maintenance. Our roundtable on **Machine Learning, Inequality and Bias**, co-hosted in Berlin with the Robert Bosch Academy, gathered researchers and policymakers from across Europe to address issues of bias, discrimination, and fairness in machine learning and related technologies.

Our annual public symposium convenes leaders from academia, industry, government, and civil society to examine the biggest challenges we face as AI moves into our everyday lives. The **AI Now 2018 Symposium** addressed the intersection of AI ethics, organizing, and accountability, examining the landmark events of the past year. Over 1,000 people registered for the event, which was free and open to the public. Recordings of the program are available on our **website**.
More information is available at **www.ainowinstitute.org**.

RECOMMENDATIONS

1. **Governments need to regulate AI by expanding the powers of sector-specific agencies to oversee, audit, and monitor these technologies by domain.** The implementation of AI systems is expanding rapidly, without adequate governance, oversight, or accountability regimes. Domains like health, education, criminal justice, and welfare all have their own histories, regulatory frameworks, and hazards. However, a national AI safety body or general AI standards and certification model will struggle to meet the sectoral expertise requirements needed for nuanced regulation. We need a sector-specific approach that does not prioritize the technology, but focuses on its application within a given domain. Useful examples of sector-specific approaches include the United States Federal Aviation Administration and the National Highway Traffic Safety Administration.

2. **Facial recognition and affect recognition need stringent regulation to protect the public interest.** Such regulation should include national laws that require strong oversight, clear limitations, and public transparency. Communities should have the right to reject the application of these technologies in both public and private contexts. Mere public notice of their use is not sufficient, and there should be a high threshold for any consent, given the dangers of oppressive and continual mass surveillance. Affect recognition deserves particular attention. Affect recognition is a subclass of facial recognition that claims to detect things such as personality, inner feelings, mental health, and "worker engagement" based on images or video of faces. These claims are not backed by robust scientific evidence, and are being applied in unethical and irresponsible ways that often recall the pseudosciences of phrenology and physiognomy. Linking affect recognition to hiring, access to insurance, education, and policing creates deeply concerning risks, at both an individual and societal level.

3. **The AI industry urgently needs new approaches to governance. As this report demonstrates, internal governance structures at most technology companies are failing to ensure accountability for AI systems.** Government regulation is an important component, but leading companies in the AI industry also need internal accountability structures that go beyond ethics guidelines. This should include rank-and-file employee representation on the board of directors, external ethics advisory boards, and the implementation of independent monitoring and transparency efforts. Third party experts should be able to audit and publish about key systems, and companies need to ensure that their AI infrastructures can be understood from "nose to tail," including their ultimate application and use.

4. **AI companies should waive trade secrecy and other legal claims that stand in the way of accountability in the public sector.** Vendors and developers who create AI and automated decision systems for use in government should agree to waive any trade secrecy or other legal claim that inhibits full auditing and understanding of their software. Corporate secrecy

laws are a barrier to due process: they contribute to the "black box effect" rendering systems opaque and unaccountable, making it hard to assess bias, contest decisions, or remedy errors. Anyone procuring these technologies for use in the public sector should demand that vendors waive these claims before entering into any agreements.

5. **Technology companies should provide protections for conscientious objectors, employee organizing, and ethical whistleblowers.** Organizing and resistance by technology workers has emerged as a force for accountability and ethical decision making. Technology companies need to protect workers' ability to organize, whistleblow, and make ethical choices about what projects they work on. This should include clear policies accommodating and protecting conscientious objectors, ensuring workers the right to know what they are working on, and the ability to abstain from such work without retaliation or retribution. Workers raising ethical concerns must also be protected, as should whistleblowing in the public interest.

6. **Consumer protection agencies should apply "truth-in-advertising" laws to AI products and services.** The hype around AI is only growing, leading to widening gaps between marketing promises and actual product performance. With these gaps come increasing risks to both individuals and commercial customers, often with grave consequences. Much like other products and services that have the potential to seriously impact or exploit populations, AI vendors should be held to high standards for what they can promise, especially when the scientific evidence to back these promises is inadequate and the longer-term consequences are unknown.

7. **Technology companies must go beyond the "pipeline model" and commit to addressing the practices of exclusion and discrimination in their workplaces.** Technology companies and the AI field as a whole have focused on the "pipeline model," looking to train and hire more diverse employees. While this is important, it overlooks what happens once people are hired into workplaces that exclude, harass, or systemically undervalue people on the basis of gender, race, sexuality, or disability. Companies need to examine the deeper issues in their workplaces, and the relationship between exclusionary cultures and the products they build, which can produce tools that perpetuate bias and discrimination. This change in focus needs to be accompanied by practical action, including a commitment to end pay and opportunity inequity, along with transparency measures about hiring and retention.

8. **Fairness, accountability, and transparency in AI require a detailed account of the "full stack supply chain."** For meaningful accountability, we need to better understand and track the component parts of an AI system and the full supply chain on which it relies: that means accounting for the origins and use of training data, test data, models, application program interfaces (APIs), and other infrastructural components over a product life cycle. We call this accounting for the "full stack supply chain" of AI systems, and it is a necessary condition for a

more responsible form of auditing. The full stack supply chain also includes understanding the true environmental and labor costs of AI systems. This incorporates energy use, the use of labor in the developing world for content moderation and training data creation, and the reliance on clickworkers to develop and maintain AI systems.

9. **More funding and support are needed for litigation, labor organizing, and community participation on AI accountability issues.** The people most at risk of harm from AI systems are often those least able to contest the outcomes. We need increased support for robust mechanisms of legal redress and civic participation. This includes supporting public advocates who represent those cut off from social services due to algorithmic decision making, civil society organizations and labor organizers that support groups that are at risk of job loss and exploitation, and community-based infrastructures that enable public participation.

10. **University AI programs should expand beyond computer science and engineering disciplines.** AI began as an interdisciplinary field, but over the decades has narrowed to become a technical discipline. With the increasing application of AI systems to social domains, it needs to expand its disciplinary orientation. That means centering forms of expertise from the social and humanistic disciplines. AI efforts that genuinely wish to address social implications cannot stay solely within computer science and engineering departments, where faculty and students are not trained to research the social world. Expanding the disciplinary orientation of AI research will ensure deeper attention to social contexts, and more focus on potential hazards when these systems are applied to human populations.

EXECUTIVE SUMMARY

At the core of the cascading scandals around AI in 2018 are questions of accountability: who is responsible when AI systems harm us? How do we understand these harms, and how do we remedy them? Where are the points of intervention, and what additional research and regulation is needed to ensure those interventions are effective? Currently there are few answers to these questions, and the frameworks presently governing AI are not capable of ensuring accountability. As the pervasiveness, complexity, and scale of these systems grow, the lack of meaningful accountability and oversight – including basic safeguards of responsibility, liability, and due process – is an increasingly urgent concern.

Building on our 2016 and 2017 reports, the AI Now 2018 Report contends with this central problem and addresses the following key issues:

1. The growing accountability gap in AI, which favors those who create and deploy these technologies at the expense of those most affected
2. The use of AI to maximize and amplify surveillance, especially in conjunction with facial and affect recognition, increasing the potential for centralized control and oppression
3. Increasing government use of automated decision systems that directly impact individuals and communities without established accountability structures
4. Unregulated and unmonitored forms of AI experimentation on human populations
5. The limits of technological solutions to problems of fairness, bias, and discrimination

Within each topic, we identify emerging challenges and new research, and provide recommendations regarding AI development, deployment, and regulation. We offer practical pathways informed by research so that policymakers, the public, and technologists can better understand and mitigate risks. Given that the AI Now Institute's location and regional expertise is concentrated in the U.S., this report will focus primarily on the U.S. context, which is also where several of the world's largest AI companies are based.

The AI accountability gap is growing: The technology scandals of 2018 have shown that the gap between those who develop and profit from AI—and those most likely to suffer the consequences of its negative effects—is growing larger, not smaller. There are several reasons for this, including a lack of government regulation, a highly concentrated AI sector, insufficient governance structures within technology companies, power asymmetries between companies and the people they serve, and a stark cultural divide between the engineering cohort responsible for technical research, and the vastly diverse populations where AI systems are deployed. These gaps are producing growing concern about bias, discrimination, due process, liability, and overall responsibility for harm. This report emphasizes the urgent need for stronger, sector-specific research and regulation.

AI is amplifying widespread surveillance: The role of AI in widespread surveillance has expanded immensely in the U.S., China, and many other countries worldwide. This is seen in the growing use of sensor networks, social media tracking, facial recognition, and affect recognition. These expansions not only threaten individual privacy, but accelerate the automation of surveillance, and thus its reach and pervasiveness. This presents new dangers, and magnifies many longstanding concerns. The use of affect recognition, based on debunked pseudoscience, is also on the rise. Affect recognition attempts to read inner emotions by a close analysis of the face and is connected to spurious claims about people's mood, mental health, level of engagement, and guilt or innocence. This technology is already being used for discriminatory and unethical purposes, often without people's knowledge. Facial recognition technology poses its own dangers, reinforcing skewed and potentially discriminatory practices, from criminal justice to education to employment, and presents risks to human rights and civil liberties in multiple countries.

Governments are rapidly expanding the use of automated decision systems without adequate protections for civil rights: Around the world, government agencies are procuring and deploying automated decision systems (ADS) under the banners of efficiency and cost-savings. Yet many of these systems are untested and poorly designed for their tasks, resulting in illegal and often unconstitutional violations of individual rights. Worse, when they make errors and bad decisions, the ability to question, contest, and remedy these is often difficult or impossible. Some agencies are attempting to provide mechanisms for transparency, due process, and other basic rights, but trade secrecy and similar laws threaten to prevent auditing and adequate testing of these systems. Drawing from proactive agency efforts, and from recent strategic litigation, we outline pathways for ADS accountability.

Rampant testing of AI systems "in the wild" on human populations: Silicon Valley is known for its "move fast and break things" mentality, whereby companies are pushed to experiment with new technologies quickly and without much regard for the impact of failures, including who bears the risk. In the past year, we have seen a growing number of experiments deploying AI systems "in the wild" without proper protocols for notice, consent, or accountability. Such experiments continue, due in part to a lack of consequences for failure. When harms occur, it is often unclear where or with whom the responsibility lies. Researching and assigning appropriate responsibility and liability remains an urgent priority.

The limits of technological fixes to problems of fairness, bias, and discrimination: Much new work has been done designing mathematical models for what should be considered "fair" when machines calculate outcomes, aimed at avoiding discrimination. Yet, without a framework that accounts for social and political contexts and histories, these mathematical formulas for fairness will almost inevitably miss key factors, and can serve to paper over deeper problems in ways that ultimately increase harm or ignore justice. Broadening perspectives and expanding research into AI fairness and bias beyond the merely mathematical is critical to ensuring we are capable of addressing the core issues and moving the focus from parity to justice.

The move to ethical principles: This year saw the emergence of numerous ethical principles and guidelines for the creation and deployment of AI technologies, many in response to growing concerns about AI's social implications. But as studies show, these types of ethical commitments have little measurable effect on software development practices if they are not directly tied to structures of accountability and workplace practices. Further, these codes and guidelines are rarely backed by enforcement, oversight, or consequences for deviation. Ethical codes can only help close the AI accountability gap if they are truly built into the processes of AI development and are backed by enforceable mechanisms of responsibility that are accountable to the public interest.

The following report develops these themes in detail, reflecting on the latest academic research, and outlines seven strategies for moving forward:

1. Expanding AI fairness research beyond a focus on mathematical parity and statistical fairness toward issues of justice
2. Studying and tracking the full stack of infrastructure needed to create AI, including accounting for material supply chains
3. Accounting for the many forms of labor required to create and maintain AI systems
4. Committing to deeper interdisciplinarity in AI
5. Analyzing race, gender, and power in AI
6. Developing new policy interventions and strategic litigation
7. Building coalitions between researchers, civil society, and organizers within the technology sector

These approaches are designed to positively recast the AI field and address the growing power imbalance that currently favors those who develop and profit from AI systems at the expense of the populations most likely to be harmed.

INTRODUCTION

The Social Challenges of AI in 2018

The past year has seen accelerated integration of powerful artificial intelligence systems into core social institutions, against a backdrop of rising inequality, political populism, and industry scandals.[1] There have been major movements from both inside and outside technology companies pushing for greater accountability and justice. The AI Now 2018 Report focuses on these themes and examines the gaps between AI ethics and meaningful accountability, and the role of organizing and regulation.

In short, it has been a dramatic year in AI. In any normal year, Cambridge Analytica seeking to manipulate national elections in the US and UK using social media data and algorithmic ad targeting would have been the biggest story.[2] But in 2018, it was just one of many scandals. Facebook had a series of disasters, including a massive data breach in September,[3] multiple class action lawsuits for discrimination,[4] accusations of inciting ethnic cleansing in Myanmar,[5] potential violations of the Fair Housing Act,[6] and hosting masses of fake Russian accounts.[7] Throughout the year, the company's executives were frequently summoned to testify, with Mark Zuckerberg facing the US Senate in April and the European Parliament in May.[8] Zuckerberg mentioned AI technologies over 30 times in his Congressional testimony as the cure-all to the company's problems, particularly in the complex areas of censorship, fairness, and content moderation.[9]

But Facebook wasn't the only one in crisis. News broke in March that Google was building AI systems for the Department of Defense's drone surveillance program, Project Maven.[10] The news kicked off an unprecedented wave of technology worker organizing and dissent across the industry.[11] In June, when the Trump administration introduced the family separation policy that forcibly removed immigrant children from their parents, employees from Amazon, Salesforce, and Microsoft all asked their companies to end contracts with U.S. Immigration and Customs Enforcement (ICE).[12] Less than a month later, it was revealed that ICE modified its own risk assessment algorithm so that it could only produce one result: the system recommended "detain" for 100% of immigrants in custody.[13]

Throughout the year, AI systems continued to be tested on live populations in high-stakes domains, with some serious consequences. In March, autonomous cars killed drivers and pedestrians.[14] Then in May, a voice recognition system in the UK designed to detect immigration fraud ended up cancelling thousands of visas and deporting people in error.[15] Documents leaked in July showed that IBM Watson was producing "unsafe and incorrect" cancer treatment recommendations.[16] And an investigation in September revealed that IBM was also working with the New York City Police Department (NYPD) to build an "ethnicity detection" feature to search faces based on race, using police camera footage of thousands of people in the streets of New York taken without their knowledge or permission.[17]

This is just a sampling of an extraordinary series of incidents from 2018.[18] The response has included a growing wave of criticism, with demands for greater accountability from the technology industry and the systems they build.[19] In turn, some companies have made public calls for the U.S. to regulate technologies like facial recognition.[20] Others have published AI ethics principles and increased efforts to produce technical fixes for issues of bias and discrimination in AI systems. But many of these ethical and technical approaches define the problem space very narrowly, neither contending with the historical or social context nor providing mechanisms for public accountability, oversight, and due process. This makes it nearly impossible for the public to validate that any of the current problems have, in fact, been addressed.

As numerous scholars have noted, one significant barrier to accountability is the culture of industrial and legal secrecy that dominates AI development.[21] Just as many AI technologies are "black boxes", so are the industrial cultures that create them.[22] Many of the fundamental building blocks required to understand AI systems and to ensure certain forms of accountability – from training data, to data models, to the code dictating algorithmic functions, to implementation guidelines and software, to the business decisions that directed design and development – are rarely accessible to review, hidden by corporate secrecy laws.

The current accountability gap is also caused by the incentives driving the rapid pace of technical AI research. The push to "innovate," publish first, and present a novel addition to the technical domain has created an accelerated cadence in the field of AI, and in technical disciplines more broadly. This comes at the cost of considering empirical questions of context and use, or substantively engaging with ethical concerns.[23] Similarly, technology companies are driven by pressures to "launch and iterate," which assume complex social and political questions will be handled by policy and legal departments, leaving developers and sales departments free from the responsibility of considering the potential downsides. The "move fast and break things" culture provides little incentive for ensuring meaningful public accountability or engaging the communities most likely to experience harm.[24] This is particularly problematic as the accelerated application of AI systems in sensitive social and political domains presents risks to marginalized communities.

The challenge to create better governance and greater accountability for AI poses particular problems when such systems are woven into the fabric of government and public institutions. The lack of transparency, notice, meaningful engagement, accountability, and oversight creates serious structural barriers for due process and redress for unjust and discriminatory decisions.

In this year's report, we assess many pressing issues facing us as AI tools are deployed further into the institutions that govern everyday life. We focus on the biggest industry players, because the number of companies able to create AI at scale is very small, while their power and reach is global. We evaluate the current range of responses from industry, governments, researchers,

activists, and civil society at large. We suggest a series of substantive approaches and make ten specific recommendations. Finally, we share the latest research and policy strategies that can contribute to greater accountability, as well as a richer understanding of AI systems in a wider social context.

1. THE INTENSIFYING PROBLEM SPACE

In identifying the most pressing social implications of AI this year, we look closely at the role of AI in widespread surveillance in multiple countries around the world, and at the implications for rights and liberties. In particular, we consider the increasing use of facial recognition, and a subclass of facial recognition known as affect recognition, and assess the growing calls for regulation. Next, we share our findings on the government use of automated decision systems, and what questions this raises for fairness, transparency, and due process when such systems are protected by trade secrecy and other laws that prevent auditing and close examination.[25] Finally, we look at the practices of deploying experimental systems "in the wild," testing them on human populations. We analyze who has the most to gain, and who is at greatest risk of experiencing harm.

1.1 AI is Amplifying Widespread Surveillance

This year, we have seen AI amplify large-scale surveillance through techniques that analyze video, audio, images, and social media content across entire populations and identify and target individuals and groups. While researchers and advocates have long warned about the dangers of mass data collection and surveillance,[26] AI raises the stakes in three areas: automation, scale of analysis, and predictive capacity. Specifically, AI systems allow automation of surveillance capabilities far beyond the limits of human review and hand-coded analytics. Thus, they can serve to further centralize these capabilities in the hands of a small number of actors. These systems also exponentially scale analysis and tracking across large quantities of data, attempting to make connections and inferences that would have been difficult or impossible before their introduction. Finally, they provide new predictive capabilities to make determinations about individual character and risk profiles, raising the possibility of granular population controls.

China has offered several examples of alarming AI-enabled surveillance this year, which we know about largely because the government openly acknowledges them. However, it's important to note that many of the same infrastructures already exist in the U.S. and elsewhere, often produced and promoted by private companies whose marketing emphasizes beneficial use cases. In the U.S. the use of these tools by law enforcement and government is rarely open to public scrutiny, as we will review, and there is much we do not know. Such infrastructures and capabilities could easily be turned to more surveillant ends in the U.S., without public disclosure and oversight, depending on market incentives and political will.

In China, military and state-sanctioned automated surveillance technology is being deployed to monitor large portions of the population, often targeting marginalized groups. Reports include installation of facial recognition tools at the Hong Kong-Shenzhen border,[27] using flocks of robotic dove-like drones in five provinces across the country,[28] and the widely reported social credit monitoring system,[29] each of which illustrates how AI-enhanced surveillance systems can be mobilized as a means of far-reaching social control.[30]

The most oppressive use of these systems is reportedly occuring in the Xinjiang Autonomous Region, described by *The Economist* as a "police state like no other."[31] Surveillance in this Uighur ethnic minority area is pervasive, ranging from physical checkpoints and programs where Uighur households are required to "adopt" Han Chinese officials into their family, to the widespread use of surveillance cameras, spyware, Wi-Fi sniffers, and biometric data collection, sometimes by stealth. Machine learning tools integrate these streams of data to generate extensive lists of suspects for detention in re-education camps, built by the government to discipline the group. Estimates of the number of people detained in these camps range from hundreds of thousands to nearly one million.[32]

These infrastructures are not unique to China. Venezuela announced the adoption of a new smart card ID known as the "carnet de patria," which, by integrating government databases linked to social programs, could enable the government to monitor citizens' personal finances, medical history, and voting activity.[33] In the United States, we have seen similar efforts. The Pentagon has funded research on AI-enabled social media surveillance to help predict large-scale population behaviors,[34] and the U.S. Immigration and Customs Enforcement (ICE) agency is using an Investigative Case Management System developed by Palantir and powered by Amazon Web Services in its deportation operations.[35] The system integrates public data with information purchased from private data brokers to create profiles of immigrants in order to aid the agency in profiling, tracking, and deporting individuals.[36] These examples show how AI systems increase integration of surveillance technologies into data-driven models of social control and amplify the power of such data, magnifying the stakes of misuse and raising urgent and important questions as to how basic rights and liberties will be protected.

The faulty science and dangerous history of affect recognition

We are also seeing new risks emerging from unregulated facial recognition systems. These systems facilitate the detection and recognition of individual faces in images or video, and can be used in combination with other tools to conduct more sophisticated forms of surveillance, such as automated lip-reading, offering the ability to observe and interpret speech from a distance.[37]

Among a host of AI-enabled surveillance and tracking techniques, facial recognition raises particular civil liberties concerns. Because facial features are a very personal form of biometric identification that is extremely difficult to change, it is hard to subvert or "opt out" of its operations.

And unlike other tracking tools, facial recognition seeks to use AI for much more than simply recognizing faces. Once identified, a face can be linked with other forms of personal records and identifiable data, such as credit score, social graph, or criminal record.

Affect recognition, a subset of facial recognition, aims to interpret faces to automatically detect inner emotional states or even hidden intentions. This approach promises a type of emotional weather forecasting: analyzing hundreds of thousands of images of faces, detecting "micro-expressions," and mapping these expressions to "true feelings."[38] This reactivates a long tradition of physiognomy – a pseudoscience that claims facial features can reveal innate aspects of our character or personality. Dating from ancient times, scientific interest in physiognomy grew enormously in the nineteenth century, when it became a central method for scientific forms of racism and discrimination.[39] Although physiognomy fell out of favor following its association with Nazi race science, researchers are worried about a reemergence of physiognomic ideas in affect recognition applications.[40] The idea that AI systems might be able to tell us what a student, a customer, or a criminal suspect is really feeling or what type of person they intrinsically are is proving attractive to both corporations and governments, even though the scientific justifications for such claims are highly questionable, and the history of their discriminatory purposes well-documented.

The case of affect detection reveals how machine learning systems can easily be used to intensify forms of classification and discrimination, even when the basic foundations of these theories remain controversial among psychologists. The scientist most closely associated with AI-enabled affect detection is the psychologist Paul Ekman, who asserted that emotions can be grouped into a small set of basic categories like anger, disgust, fear, happiness, sadness, and surprise.[41] Studying faces, according to Ekman, produces an objective reading of authentic interior states—a direct window to the soul. Underlying his belief was the idea that emotions are fixed and universal, identical across individuals, and clearly visible in observable biological mechanisms regardless of cultural context. But Ekman's work has been deeply criticized by psychologists, anthropologists, and other researchers who have found his theories do not hold up under sustained scrutiny.[42] The psychologist Lisa Feldman Barrett and her colleagues have argued that an understanding of emotions in terms of these rigid categories and simplistic physiological causes is no longer tenable.[43] Nonetheless, AI researchers have taken his work as fact, and used it as a basis for automating emotion detection.[44]

Contextual, social, and cultural factors — how, where, and by whom such emotional signifiers are expressed — play a larger role in emotional expression than was believed by Ekman and his peers. In light of this new scientific understanding of emotion, any simplistic mapping of a facial expression onto basic emotional categories through AI is likely to reproduce the errors of an outdated scientific paradigm. It also raises troubling ethical questions about locating the arbiter of someone's "real" character and emotions outside of the individual, and the potential abuse of

power that can be justified based on these faulty claims. Psychiatrist Jamie Metzl documents a recent cautionary example: a pattern in the 1960s of diagnosing Black people with schizophrenia if they supported the civil rights movement.[45] Affect detection combined with large-scale facial recognition has the potential to magnify such political abuses of psychological profiling.

In the realm of education, some U.S. universities have considered using affect analysis software on students.[46] The University of St. Thomas, in Minnesota, looked at using a system based on Microsoft's facial recognition and affect detection tools to observe students in the classroom using a webcam. The system predicts the students' emotional state. An overview of student sentiment is viewable by the teacher, who can then shift their teaching in a way that "ensures student engagement," as judged by the system. This raises serious questions on multiple levels: what if the system, with a simplistic emotional model, simply cannot grasp more complex states? How would a student contest a determination made by the system? What if different students are seen as "happy" while others are "angry"—how should the teacher redirect the lesson? What are the privacy implications of such a system, particularly given that, in the case of the pilot program, there is no evidence that students were informed of its use on them?

Outside of the classroom, we are also seeing personal assistants, like Alexa and Siri, seeking to pick up on the emotional undertones of human speech, with companies even going so far as to patent methods of marketing based on detecting emotions, as well as mental and physical health.[47] The AI-enabled emotion measurement company Affectiva now promises it can promote safer driving by monitoring "driver and occupant emotions, cognitive states, and reactions to the driving experience...from face and voice."[48] Yet there is little evidence that any of these systems actually work across different individuals, contexts, and cultures, or have any safeguards put in place to mitigate concerns about privacy, bias, or discrimination in their operation. Furthermore, as we have seen in the large literature on bias and fairness, classifications of this nature not only have direct impacts on human lives, but also serve as data to train and influence other AI systems. This raises the stakes for any use of affect recognition, further emphasizing why it should be critically examined and its use severely restricted.

Facial recognition amplifies civil rights concerns

Concerns are intensifying that facial recognition increases racial discrimination and other biases in the criminal justice system. Earlier this year, the American Civil Liberties Union (ACLU) disclosed that both the Orlando Police Department and the Washington County Sheriff's department were using Amazon's Rekognition system, which boasts that it can perform "real-time face recognition across tens of millions of faces" and detect "up to 100 faces in challenging crowded photos."[49] In Washington County, Amazon specifically worked with the Sheriff's department to create a mobile app that could scan faces and compare them against a database

of at least 300,000 mugshots.[50] An Amazon representative recently revealed during a talk that they have been considering applications where Orlando's network of surveillance cameras could be used in conjunction with facial recognition technology to find a "person of interest" wherever they might be in the city.[51]

In addition to the privacy and mass surveillance concerns commonly raised, the use of facial recognition in law enforcement has also intersected with concerns of racial and other biases. Researchers at the ACLU and the University of California (U.C.) Berkeley tested Amazon's Rekognition tool by comparing the photos of sitting members in the United States Congress with a database containing 25,000 photos of people who had been arrested. The results showed significant levels of inaccuracy: Amazon's Rekognition incorrectly identified 28 members of Congress as people from the arrest database. Moreover, the false positives disproportionately occurred among non-white members of Congress, with an error rate of nearly 40% compared to only 5% for white members.[52] Such results echo a string of findings that have demonstrated that facial recognition technology is, on average, better at detecting light-skinned people than dark-skinned people, and better at detecting men than women.[53]

In its response to the ACLU, Amazon acknowledged that "the Rekognition results can be significantly skewed by using a facial database that is not appropriately representative."[54] Given the deep and historical racial biases in the criminal justice system, most law enforcement databases are unlikely to be "appropriately representative."[55] Despite these serious flaws, ongoing pressure from civil rights groups, and protests from Amazon employees over the potential for misuse of these technologies, Amazon Web Services CEO Andrew Jassy recently told employees that "we feel really great and really strongly about the value that Amazon Rekognition is providing our customers of all sizes and all types of industries in law enforcement and out of law enforcement."[56]

Nor is Amazon alone in implementing facial recognition technologies in unaccountable ways. Investigative journalists recently disclosed that IBM and the New York City Police Department (NYPD) partnered to develop such a system that included "ethnicity search" as a custom feature, trained on thousands of hours of NYPD surveillance footage.[57] Use of facial recognition software in the private sector has expanded as well.[58] Major retailers and venues have already begun using these technologies to detect shoplifters, monitor crowds, and even "scan for unhappy customers," using facial recognition systems instrumented with "affect detection" capabilities.[59]

These concerns are amplified by a lack of laws and regulations. There is currently no federal legislation that seeks to provide standards, restrictions, requirements, or guidance regarding the development or use of facial recognition technology. In fact, most existing federal legislation looks to promote the use of facial recognition for surveillance, immigration enforcement, employment verification, and domestic entry-exit systems.[60] The laws that we do have are piecemeal, and none specifically address facial recognition. Among these is the Biometric Information Privacy Act, a 2008 Illinois law that sets forth stringent rules regarding the collection of biometrics. While the law does not mention facial recognition, given that the technology was

not widely available in 2008, many of its requirements, such as obtaining consent, are reasonably interpreted to apply.[61] More recently, several municipalities and a local transit system have adopted ordinances that seek to create greater transparency and oversight of data collection and use requirements regarding the acquisition of surveillance technologies, which would include facial recognition based on the expansive definition in these ordinances.[62]

Opposition to the use of facial recognition tools by government agencies is growing. Earlier this year, AI Now joined the ACLU and over 30 other research and advocacy organizations calling on Amazon to stop selling facial recognition software to government agencies after the ACLU uncovered documents showing law enforcement use of Amazon's Rekognition API.[63] Members of Congress are also pushing Amazon to provide more information.[64]

Some have gone further, calling for an outright ban. Scholars Woodrow Hartzog and Evan Selinger argue that facial recognition technology is a "tool for oppression that's perfectly suited for governments to display unprecedented authoritarian control and an all-out privacy-eviscerating machine," necessitating extreme caution and diligence before being applied in our contemporary digital ecosystem.[65] Critiquing the Stanford "gaydar" study that claimed its deep neural network was more accurate than humans at predicting sexuality from facial images,[66] Frank Pasquale wrote that "there are some scientific research programs best not pursued - and this might be one of them."[67]

Kade Crockford, Director of the Technology for Liberty Program at ACLU of Massachusetts, also wrote in favor of a ban, stating that "artificial intelligence technologies like face recognition systems fundamentally change the balance of power between the people and the government...some technologies are so dangerous to that balance of power that they must be rejected."[68] Microsoft President Brad Smith has called for government regulation of facial recognition, while Rick Smith, CEO of law enforcement technology company Axon, recently stated that the "accuracy thresholds" of facial recognition tools aren't "where they need to be to be making operational decisions."[69]

The events of this year have strongly underscored the urgent need for stricter regulation of both facial and affect recognition technologies. Such regulations should severely restrict use by both the public and the private sector, and ensure that communities affected by these technologies are the final arbiters of whether they are used at all. This is especially important in situations where basic rights and liberties are at risk, requiring stringent oversight, audits, and transparency. Linkages should not be permitted between private and government databases. At this point, given the evidence in hand, policymakers should not be funding or furthering the deployment of these systems in public spaces.

1.2 The Risks of Automated Decision Systems in Government

Over the past year, we have seen a substantial increase in the adoption of Automated Decision Systems (ADS) across government domains, including criminal justice, child welfare, education, and immigration. Often adopted under the theory that they will improve government efficiency or cost-savings, ADS seek to aid or replace various decision-making processes and policy determinations. However, because the underlying models are often proprietary and the systems frequently untested before deployment, many community advocates have raised significant concerns about lack of due process, accountability, community engagement, and auditing.[70]

Such was the case for Tammy Dobbs, who moved to Arkansas in 2008 and signed up for a state disability program to help her with her cerebral palsy.[71] Under the program, the state sent a qualified nurse to assess Tammy to determine the number of caregiver hours she would need. Because Tammy spent most of her waking hours in a wheelchair and had stiffness in her hands, her initial assessment allocated 56 hours of home care per week. Fast forward to 2016, when the state assessor arrived with a new ADS on her laptop. Using a proprietary algorithm, this system calculated the number of hours Tammy would be allotted. Without any explanation or opportunity for comment, discussion, or reassessment, the program allotted Tammy 32 hours per week, a massive and sudden drop that Tammy had no chance to prepare for and that severely reduced her quality of life.

Nor was Tammy's situation exceptional. According to Legal Aid of Arkansas attorney Kevin De Liban, hundreds of other individuals with disabilities also received dramatic reductions in hours, all without any meaningful opportunity to understand or contest their allocations. Legal Aid subsequently sued the State of Arkansas, eventually winning a ruling that the new algorithmic allocation program was erroneous and unconstitutional. Yet by then, much of the damage to the lives of those affected had been done.[72]

The Arkansas disability cases provide a concrete example of the substantial risks that occur when governments use ADS in decisions that have immediate impacts on vulnerable populations. While individual assessors may also suffer from bias or flawed logic, the impact of their case-by-case decisions has nowhere near the magnitude or scale that a single flawed ADS can have across an entire population.

The increased introduction of such systems comes at a time when, according to the World Income Inequality Database, the United States has the highest income inequality rate of all western countries.[73] Moreover, Federal Reserve data shows wealth inequalities continue to grow, and racial wealth disparities have more than tripled in the last 50 years, with current policies set to exacerbate such problems.[74] In 2018 alone, we have seen a U.S. executive order cutting funding for social programs that serve the country's poorest citizens,[75] alongside a proposed federal

budget that will significantly reduce low-income and affordable housing,[76] the implementation of onerous work requirements for Medicaid,[77] and a proposal to cut food assistance benefits for low-income seniors and people with disabilities.[78]

In the context of such policies, agencies are under immense pressure to cut costs, and many are looking to ADS as a means of automating hard decisions that have very real effects on those most in need.[79] As such, many ADS systems are often implemented with the goal of doing more with less in the context of austerity policies and cost-cutting. They are frequently designed and configured primarily to achieve these goals, with their ultimate effectiveness being evaluated based on their ability to trim costs, often at the expense of the populations such tools are ostensibly intended to serve.[80] As researcher Virginia Eubanks argues, "What seems like an effort to lower program barriers and remove human bias often has the opposite effect, blocking hundreds of thousands of people from receiving the services they deserve."[81]

When these problems arise, they are frequently difficult to remedy. Few ADS are designed or implemented in ways that easily allow affected individuals to contest, mitigate, or fix adverse or incorrect decisions. Additionally, human discretion and the ability to intervene or override a system's determination is often substantially limited or removed from case managers, social workers, and others trained to understand the context and nuance of a particular person and situation.[82] These front-line workers become mere intermediaries, communicating inflexible decisions made by automated systems, without the ability to alter them.

Unlike the civil servants who have historically been responsible for such decisions, many ADS come from private vendors and are frequently implemented without thorough testing, review, or auditing to ensure their fitness for a given domain.[83] Nor are these systems typically built with any explicit form of oversight or accountability. This makes discovery of problematic automated outcomes difficult, especially since such errors and evidence of discrimination frequently manifest as collective harms, only recognizable as a pattern across many individual cases. Detecting such problems requires oversight and monitoring. It also requires access to data that is often neither available to advocates and the public nor monitored by government agencies.

For example, the Houston Federation of Teachers sued the Houston Independent School District for procuring a third-party ADS to use student test data to make teacher employment decisions, including which teachers were promoted and which were terminated. It was revealed that no one in the district – not a single employee – could explain or even replicate the determinations made by the system, even though the district had access to all the underlying data.[84] Teachers who sought to contest the determinations were told that the "black box" system was simply to be believed and could not be questioned. Even when the teachers brought a lawsuit, claiming constitutional, civil rights, and labor law violations, the ADS vendor fought against providing any access to how its system worked. As a result, the judge ruled that the use of this ADS in public employee cases could run afoul of constitutional due process protections, especially when trade secrecy blocked employees' ability to understand how decisions were made. The case has subsequently been settled, with the District agreeing to abandon the third-party ADS.

Similarly, in 2013, Los Angeles County adopted an ADS to assess imminent danger or harm to children, and to predict the likelihood of a family being re-referred to the child welfare system within 12 to 18 months. The County did not perform a review of the system or assess the efficacy of using predictive analytics for child safety and welfare. It was only after the death of a child whom the system failed to identify as at-risk that County leadership directed a review, which raised serious questions regarding the system's validity. The review specifically noted that the system failed to provide a comprehensive picture of a given family, "but instead focus[ed] on a few broad strokes without giving weight to important nuance."[85] Virginia Eubanks found similar problems in her investigation of an ADS developed by the same private vendor for use in Allegheny County, PA. This system produced biased outcomes because it significantly oversampled poor children from working class communities, especially communities of color, in effect subjecting poor parents and children to more frequent investigation.[86]

Even in the face of acknowledged issues of bias and the potential for error in high-stakes domains, these systems are being rapidly adopted. The Ministry of Social Development in New Zealand supported the use of a predictive ADS system to identify children at risk of maltreatment, despite their recognizing that the system raised "significant ethical concerns." They defended this on the grounds that the benefits "plausibly outweighed" the potential harms, which included reconfiguring child welfare as a statistical issue.[87]

These cases not only highlight the need for greater transparency, oversight, and accountability in the adoption, development, and implementation of ADS, but also the need for examination of the limitations of these systems overall, and of the economic and policy factors that accompany the push to apply such systems. Virginia Eubanks, who investigated Allegheny County's use of an ADS in child welfare, looked at this and a number of case studies to show how ADS are often adopted to avoid or obfuscate broader structural and systemic problems in society – problems that are often beyond the capacity of cash-strapped agencies to address meaningfully.[88]

Other automated systems have also been proposed as a strategy to combat pre-existing problems within government systems. For years, criminal justice advocates and researchers have pushed for the elimination of cash bail, which has been shown to disproportionately harm individuals based on race and socioeconomic status while at the same time failing to enhance public safety.[89] In response, New Jersey and California recently passed legislation aimed at addressing this concern. However, instead of simply ending cash bail, they replaced it with a pretrial assessment system designed to algorithmically generate "risk" scores that claim to predict whether a person should go free or be detained in jail while awaiting trial.[90]

The shift from policies such as cash bail to automated systems and risk assessment scoring is still relatively new, and is proceeding even without substantial research examining the potential to amplify discrimination within the criminal justice system. Yet there are some early indicators that raise concern. New Jersey's law went into effect in 2017, and while the state has experienced a decline in its pretrial population, advocates have expressed worry that racial disparities in the risk

assessment system persist.[91] Similarly, when California's legislation passed earlier this year, many of the criminal justice advocates who pushed for the end of cash bail, and supported an earlier version of the bill, opposed its final version due to the risk assessment requirement.[92]

Education policy is also feeling the impact of automated decision systems. A University College London professor is among those who argued for AI to replace standardized testing, suggesting that UCL Knowledge Lab's AIAssess can be "trusted...with the assessment of our children's knowledge and understanding," and can serve to replace or augment more traditional testing.[93] However, much like other forms of AI, there is a growing body of research that shows automated essay scoring systems may encode bias against certain linguistic and ethnic groups in ways that replicate patterns of marginalization.[94] Unfair decisions based on automated scores assigned to students from historically and systemically disadvantaged groups are likely to have profound consequences on children's lives, and to exacerbate existing disparities in access to employment opportunities and resources.[95]

The implications of educational ADS go beyond testing to other areas, such as school assignments and even transportation. The City of Boston was in the spotlight this year after two failed efforts to address school equity via automated systems. First, the school district adopted a geographically-driven school assignment algorithm, intended to provide students access to higher quality schools closer to home. The city's goal was to increase the racial and geographic integration in the school district, but a report assessing the impact of the system determined that it did the opposite: while it shortened student commutes, it ultimately reduced school integration.[96] Researchers noted that this was, in part, because it was impossible for the system to meet its intended goal given the history and context within which it was being used. The geographic distribution of quality schools in Boston was already inequitable, and the pre-existing racial disparities that played a role in placement at these schools created complications that could not be overcome by an algorithm.[97]

Following this, the Boston school district tried again to use an algorithmic system to improve inequity, this time designing it to reconfigure school start times – aiming to begin high school later, and middle school earlier. This was done in an effort to improve student health and performance based on a recognition of students' circadian rhythms at different ages, and to optimize use of school buses to produce cost savings. It also aimed to increase racial equity, since students of color primarily attended schools with inconvenient start times compounded by long bus rides. The city developed an ADS that optimized for these goals. However, it was never implemented because of significant public backlash, which ultimately resulted in the resignation of the superintendent.[98]

In this case, the design process failed to adequately recognize the needs of families, or include them in defining and reviewing system goals. Under the proposed system, parents with children in both high school and middle school would need to reconfigure their schedules for vastly different start and end times, putting strain on those without this flexibility. The National Association for the Advancement of Colored People (NAACP) and the Lawyers' Committee for Civil Rights and

Economic Justice opposed the plan because of the school district's failure to appreciate that parents of color and lower-income parents often rely on jobs that lack work schedule flexibility and may not be able to afford additional child care.[99]

These failed efforts demonstrate two important issues that policymakers must consider when evaluating the use of these systems. First, unaddressed structural and systemic problems will persist and will likely undermine the potential benefits of these systems if they are not addressed prior to a system's design and implementation. Second, robust and meaningful community engagement is essential before a system is put in place and should be included in the process of establishing a system's goals and purpose.

In AI Now's Algorithmic Impact Assessment (AIA) framework, community engagement is an integral part of any ADS accountability process, both as part of the design stage as well as before, during, and after implementation.[100] When affected communities have the opportunity to assess and potentially reject the use of systems that are not acceptable, and to call out fundamental flaws in the system before it is put in place, the validity and legitimacy of the system is vastly improved. Such engagement serves communities and government agencies: if parents of color and lower-income parents in Boston were meaningfully engaged in assessing the goals of the school start time algorithmic intervention, their concerns might have been accounted for in the design of the system, saving the city time and resources, and providing a much-needed model of oversight.

Above all, accountability in the government use of algorithmic systems is impossible when the systems making recommendations are "black boxes." When third-party vendors insist on trade secrecy to keep their systems opaque, it makes any path to redress or appeal extremely difficult.[101] This is why vendors should waive trade secrecy and other legal claims that would inhibit the ability to understand, audit, or test their systems for bias, error, or other issues. It is important for both people in government and those who study the effects of these systems to understand why automated recommendations are made, and to be able to trust their validity. It is even more critical that those whose lives are negatively impacted by these systems be able to contest and appeal adverse decisions.[102]

Governments should be cautious: while automated decision systems may promise short-term cost savings and efficiencies, it is governments, not third party vendors, who will ultimately be held responsible for their failings. Without adequate transparency, accountability, and oversight, these systems risk introducing and reinforcing unfair and arbitrary practices in critical government determinations and policies.[103]

1.3 Experimenting on Society: Who Bears the Burden?

Over the last ten years, the funding and focus on technical AI research and development has accelerated. But efforts at ensuring that these systems are safe and non-discriminatory have not

received the same resources or attention. Currently, there are few established methods for measuring, validating, and monitoring the effects of AI systems "in the wild". AI systems tasked with significant decision making are effectively tested on live populations, often with little oversight or a clear regulatory framework.

For example, in March 2018, a self-driving Uber was navigating the Phoenix suburbs and failed to "see" a woman, hitting and killing her.[104] Last March, Tesla confirmed that a second driver had been killed in an accident in which the car's autopilot technology was engaged.[105] Neither company suffered serious consequences, and in the case of Uber, the person minding the autonomous vehicle was ultimately blamed, even though Uber had explicitly disabled the vehicle's system for automatically applying brakes in dangerous situations.[106] Despite these fatal errors, Alphabet Inc.'s Waymo recently announced plans for an "early rider program" in Phoenix.[107] Residents can sign up to be Waymo test subjects, and be driven automatically in the process.

Many claim that the occasional autonomous vehicle fatality needs to be put in the context of the existing ecosystem, in which many driving-related deaths happen without AI.[108] However, because regulations and liability regimes govern humans and machines differently, risks generated from machine-human interactions do not cleanly fall into a discrete regulatory or accountability category. Strong incentives for regulatory and jurisdictional arbitrage exist in this and many other AI domains. For example, the fact that Phoenix serves as the site of Waymo and Uber testing is not an accident. Early this year, Arizona, perhaps swayed by a promise of technology jobs and capital, made official what the state allowed in practice since 2015: fully autonomous vehicles without anyone behind the wheel are permitted on public roads. This policy was put in place without any of the regulatory scaffolding that would be required to contend with the complex issues that are raised in terms of liability and accountability. In the words of the *Phoenix New Times*: "Arizona has agreed to step aside and see how this technology develops. If something goes wrong, well, there's no plan for that yet."[109] This regulatory accountability gap is clearly visible in the Uber death case, apparently caused by a combination of corporate expedience (disabling the automatic braking system) and backup driver distraction.[110]

While autonomous vehicles arguably present AI's most straightforward non-military dangers to human safety, other AI domains also raise serious concerns. For example, IBM's Watson for Oncology is already being tested in hospitals across the globe, assisting in patient diagnostics and clinical care. Increasingly, its effectiveness, and the promises of IBM's marketing, are being questioned. Investigative reporters gained access to internal documents that paint a troubling picture of IBM's system, including its recommending "unsafe and incorrect cancer treatments." While this system was still in its trial phase, it raised serious concerns about the incentives driving the rush to integrate such technology, and the lack of clinical validation and peer-reviewed research attesting to IBM's marketing claims of effectiveness.[111]

Such events have not slowed AI deployment in healthcare. Recently, the U.S. Food and Drug Administration (FDA) issued a controversial decision to clear the new Apple Watch, which features a built-in electrocardiogram (EKG) and the ability to notify a user of irregular heart

rhythm, as safe for consumers.[112] Here, concerns that the FDA may be moving too quickly in an attempt to keep up with the pace of innovation have joined with concerns around data privacy and security.[113] Similarly, DeepMind Health's decision to move its Streams Application, a tool designed to support decision-making by nurses and health practitioners, under the umbrella of Google, caused some to worry that DeepMind's promise to not share the data of patients would be broken.[114]

Children and young adults are frequently subjects of such experiments. Earlier this year, it was revealed that Pearson, a major AI-education vendor, inserted "social-psychological interventions" into one of its commercial learning software programs to test how 9,000 students would respond. They did this without the consent or knowledge of students, parents, or teachers.[115] The company then tracked whether students who received "growth-mindset" messages through the learning software attempted and completed more problems than students who did not. This psychological testing on unknowing populations, especially young people in the education system, raises significant ethical and privacy concerns. It also highlights the growing influence of private companies in purportedly public domains, and the lack of transparency and due process that accompany the current practices of AI deployment and integration.

Here we see not only examples of the real harms that can come from biased and inaccurate AI systems, but evidence of the AI industry's willingness to conduct early releases of experimental tools on human populations. As Amazon recently responded when criticized for monetizing people's wedding and baby registries with deceptive advertising tactics, "we're constantly experimenting."[116] This is a repeated pattern when market dominance and profits are valued over safety, transparency, and assurance. Without meaningful accountability frameworks, as well as strong regulatory structures, this kind of unchecked experimentation will only expand in size and scale, and the potential hazards will grow.

2. EMERGING SOLUTIONS IN 2018

2.1 Bias Busting and Formulas for Fairness: the Limits of Technological "Fixes"

Over the past year, we have seen growing consensus that AI systems perpetuate and amplify bias, and that computational methods are not inherently neutral and objective. This recognition comes in the wake of a string of examples, including evidence of bias in algorithmic pretrial risk assessments and hiring algorithms, and has been aided by the work of the Fairness, Accountability, and Transparency in Machine Learning community.[117] The community has been at the center of an emerging body of academic research on AI-related bias and fairness, producing insights into the nature of these issues, along with methods aimed at remediating bias. These approaches are now being operationalized in industrial settings.

In the search for "algorithmic fairness", many definitions of fairness, along with strategies to achieve it, have been proposed over the past few years, primarily by the technical community.[118] This work has informed the development of new algorithms and statistical techniques that aim to diagnose and mitigate bias. The success of such techniques is generally measured against one or another computational definition of fairness, based on a mathematical set of results. However, the problems these techniques ultimately aim to remedy have deep social and historical roots, some of which are more cleanly captured by discrete mathematical representations than others. Below is a brief survey of some of the more prominent approaches to understanding and defining issues involving algorithmic bias and fairness.

- *Allocative harms* describe the effects of AI systems that unfairly withhold services, resources, or opportunities from some. Such harms have captured much of the attention of those dedicated to building technical interventions that ensure fair AI systems, in part because it is (theoretically) possible to quantify such harms and their remediation.[119] However, we have seen less attention paid to fixing systems that amplify and reproduce *representational harms*: the harm caused by systems that reproduce and amplify harmful stereotypes, often doing so in ways that mirror assumptions used to justify discrimination and inequality.

  In a keynote of the 2017 Conference on Neural Information Processing (NeurIPS), AI Now cofounder Kate Crawford described the way in which historical patterns of discrimination and classification, which often construct harmful representations of people based on perceived differences, are reflected in the assumptions and data that inform AI systems, often resulting in allocative harms.[120] This perspective requires one to move beyond locating biases in an algorithm or dataset, and to consider "the role of AI in harmful representations of human identity," and the way in which such harmful representations are both shaped, and shape, our social and cultural understandings of ourselves and each other.[121]

- *Observational fairness strategies* attempt to diagnose and mitigate bias by considering a dataset (either data used for training an AI model, or the input data processed by such a model), and applying methods to the data aimed at detecting whether it encodes bias against individuals or groups based on characteristics such as race, gender, or socioeconomic standing. These characteristics are typically referred to as protected or sensitive attributes. The majority of observational fairness approaches can be categorized as being a form of either anti-classification, classification parity, or calibration, as proposed by Sam Corbett-Davies and Sharad Goel.[122] Observational fairness strategies have increasingly emerged through efforts from the community to contend with the limitations of technical fairness work and to provide entry points for other disciplines.[123]

- *Anti-classification strategies* declare a machine learning model to be fair if it does not depend on protected attributes in the data set. For instance, this strategy considers a

pretrial risk assessment of two defendants who differ based on race or gender but are identical in terms of their other personal information to be "fair" if they are assigned the same risk. This strategy often requires omitting all protected attributes and their "proxies" from the data set that is used to train a model (proxies being any attributes that are correlated to protected attributes, such as ZIP code being correlated with race).[124]

- *Classification parity* declares a model fair when its predictive performance is equal across groupings that are defined by protected attributes. For example, classification parity would ensure that the percentage of people an algorithm turns down for a loan when they are actually creditworthy (its "false negative" rate) is the same for both Black and white populations. In practice, this strategy often results in decreasing the "accuracy" for certain populations in order to match that of others.

- *Calibration strategies* look less at the data and more at the outcome once an AI system has produced a decision or prediction. These approaches work to ensure that outcomes do not depend on protected attributes. For example, in the case of pretrial risk assessment, applying a calibration strategy would aim to make sure that among a pool of defendants with a similar risk score, the proportion who actually do reoffend on release is the same across different protected attributes, such as race.

Several scholars have identified limitations with these approaches to observational fairness. With respect to anti-classification, some argue that there are important cases where protected attributes—such as race or gender—*should* be included in data used to train and inform an AI system in order to ensure equitable decisions.[125] For example, Corbett-Davies and Goel discuss the importance of including gender in pretrial risk assessment. As women reoffend less often than men in many jurisdictions, gender-neutral risk assessments tend to overstate the recidivism risk of women, "which can lead to unnecessarily harsh judicial decisions." As a result, some jurisdictions use gender-specific risk assessment tools. These cases counter a widespread view that deleting sufficient information from data sets will eventually "debias" an AI system. Since correlations between variables in a dataset almost always exist, removing such variables can result in very little information, and thus poor predictive performance without the ability to measure potential harms post hoc.

Secondly, some have argued that different mathematical fairness criteria are mutually exclusive. Hence, it is generally not possible, except in highly constrained cases, to simultaneously satisfy both calibration and any form of classification parity.[126] These "impossibility results" show how each fairness strategy makes implicit assumptions about what is and is not fair. They also highlight the inherent mathematical trade-offs facing those aiming to mitigate various forms of bias based on one or another fairness definition. Ultimately, these findings serve to complicate the broader policy debate focused on solving bias issues with mathematical fairness tools. What they make clear is that solving complex policy issues related to bias and discrimination by indiscriminately applying one or more fairness metrics is unlikely to be successful. This does not mean that such metrics are not useful: observational criteria may help understanding around

whether datasets and AI systems meet various notions of fairness and bias and subsequently help inform a richer discussion about the goals one hopes to achieve when deploying AI systems in complex social contexts.

The proliferation of observational fairness methods also raises concerns over the potential to provide a false sense of assurance. While researchers often have a nuanced sense of the limitations of their tools, others who might implement them may ignore such limits when looking for quick fixes. The idea that, once "treated" with such methods, AI systems are free of bias and safe to use in sensitive domains can provide a dangerous sense of false security—one that relies heavily on mathematical definitions of fairness without looking at the deeper social and historical context. As legal scholar Frank Pasquale observes, "algorithms alone can't meaningfully hold other algorithms accountable."[127]

While increased attention to the problems of fairness and bias in AI is a positive development, some have expressed concern over a "mathematization of ethics."[128] As Shira Mitchell has argued:

> "As statistical thinkers in the political sphere we should be aware of the hazards of supplanting politics by an expert discourse. In general, every statistical intervention to a conversation tends to raise the technical bar of entry, until it is reduced to a conversation between technical experts…are we speaking statistics to power? Or are we merely providing that power with new tools for the marginalization of unquantified political concerns?"[129]

Such concerns are not new. Upcoming work by Hutchinson and Mitchell surveys over fifty years of attempts to construct quantitative fairness definitions across multiple disciplines. Their work recalls a period between 1964 and 1973 when researchers focused on defining fairness for educational assessments in ways that echo the current AI fairness debate. Their efforts stalled after they were unable to agree on "broad technical solutions to the issues involved in fairness." These precedents emphasize what the Fairness, Accountability and Transparency in Machine Learning community has been discovering: without a "tight connection to real world impact," the added value of new fairness metrics and algorithms in the machine learning community could be minimal.[130] In order to arrive at more meaningful research on fairness and algorithmic bias, we must continue to pair the expertise and perspectives of communities outside of technical disciplines to those within.

Broader approaches

Dobbe et al. have drawn on the definition of bias proposed in the early value-sensitive design (VSD) literature to propose a broader view of fairness.[131] VSD, as theorized in the nineties by Batya Friedman and Helen Nissenbaum, asserts that bias in computer systems pre-exists the system itself.[132] Such bias is reflected in the data that informs the systems and embedded in the assumptions made during the construction of a computer system. This bias manifests during the

operation of the systems due to feedback loops and dissonance between the system and our dynamic social and cultural contexts.[133] The VSD approach is one way to bring a broader lens to these issues, emphasizing the interests and perspectives of direct and indirect stakeholders throughout the design process.

Another approach is a "social systems analysis" first described by Kate Crawford and Ryan Calo in *Nature*.[134] This is a method that combines quantitative and qualitative research methods by forensically analyzing a technical system while also studying the technology once it is deployed in social settings. It proposes that we engage with social impacts at every stage—conception, design, deployment, and regulation of a technology, across the life cycle.

We have also seen increased focus on examining the provenance and construction of the data used to train and inform AI systems. This data shapes AI systems' "view of the world," and an understanding of how it is created and what it is meant to represent is essential to understanding the limits of the systems that it informs.[135] As an initial remedy to this problem, a group of researchers led by Timnit Gebru proposed "Datasheets for Datasets," a standardized form of documentation meant to accompany datasets used to train and inform AI systems.[136] A follow-up paper looks at standardizing provenance for AI models.[137] These approaches allow AI practitioners and those overseeing and assessing the applicability of AI within a given context to better understand whether the data that shapes a given model is appropriate, representative, or potentially possessing legal or ethical issues.

Advances in bias-busting and fairness formulas are strong signs that the field of AI has accepted that these concerns are real. However, the limits of narrow mathematical models will continue to undermine these approaches until broader perspectives are included. Approaches to fairness and bias must take into account both allocative and representational harms, and those that debate the definitions of fairness and bias must recognize and give voice to the individuals and communities most affected.[138] Any formulation of fairness that excludes impacted populations and the institutional context in which a system is deployed is too limited.

2.2 Industry Applications: Toolkits and System Tweaks

This year, we have also seen several technology companies operationalize fairness definitions, metrics, and tools. In the last year, four of the biggest AI companies released bias mitigation tools. IBM released the "AI Fairness 360" open-source tool kit, which includes nine different algorithms and many other fairness metrics developed by researchers in the Fairness, Accountability and Transparency in Machine Learning community. The toolkit is intended to be integrated into the software development pipeline from early stages of data pre-processing, to the training process itself, through the use of specific mathematical models that deploy bias mitigation strategies.[139] Google's People + AI Research group (PAIR) released the open-source "What-If" tool, a dashboard allowing researchers to visualize the effects of different bias mitigation strategies and metrics, as well as a tool called "Facets" that supports decision-making around which fairness metric to

use.[140] Microsoft released fairlearn.py, a Python package meant to help implement a binary classifier subject to a developer's intended fairness constraint.[141] Facebook announced the creation and testing of a tool called "Fairness Flow", an internal tool for Facebook engineers that incorporates many of the same algorithms to help identify bias in machine learning models.[142] Even Accenture, a consulting firm, has developed internal software tools to help clients understand and "essentially eliminate the bias in algorithms."[143]

Industry standards bodies have also taken on fairness efforts in response to industry and public sector requests for accountability assurances. The Institute of Electrical and Electronics Engineers (IEEE) recently announced an Ethics Certification Program for Autonomous and Intelligent Systems in the hopes of creating "marks" that can attest to the broader public that an AI system is transparent, accountable, and fair.[144] While this effort is new, and while IEEE has not published the certification's underlying methods, it is hard to see, given the complexity of these issues, how settling on one certification standard across all contexts and all AI systems would be possible—or ultimately reliable—in ensuring that systems are used in safe and ethical ways. Similar concerns have arisen in other contexts, such as privacy certification programs.[145]

In both the rapid industrial adoption of academic fairness methods, and the rush to certification, we see an eagerness to "solve" and "eliminate" problems of bias and fairness using familiar approaches and skills that avoid the need for significant structural change, and which fail to interrogate the complex social and historical factors at play. Combining "academically credible" technical fairness fixes and certification check boxes runs the risk of instrumenting fairness in ways that lets industry say it has fixed these problems and may divert attention from examining ongoing harms. It also relieves companies of the responsibility to explore more complex and costly forms of review and remediation. Rather than relying on quick fixes, tools, and certifications, issues of bias and fairness require deeper consideration and more robust accountability frameworks, including strong disclaimers about how "automated fairness" cannot be relied on to truly eliminate bias from AI systems.

2.3 Why Ethics is Not Enough

A top-level recommendation in the AI Now 2017 Report advised that "ethical codes meant to steer the AI field should be accompanied by strong oversight and accountability mechanisms."[146] While we have seen a rush to adopt such codes, in many instances offered as a means to address the growing controversy surrounding the design and implementation of AI systems, we have not seen strong oversight and accountability to backstop these ethical commitments.

After it was revealed that Google was working with the Pentagon on Project Maven—developing AI systems for drone surveillance—the debate about the role of AI in weapons systems grew in intensity. Project Maven generated significant protest among Google's employees, who successfully petitioned the company's leadership to end their involvement with the program when the current contract expired.[147] By way of response, Google's CEO Sundar Pichai released a public

set of seven "guiding principles" designed to ensure that the company's work on AI will be socially responsible.[148] These ethical principles include the commitment to "be socially beneficial," and to "avoid creating or reinforcing unfair bias." They also include a section titled, "AI applications we will not pursue," which includes "weapons and other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people"—a direct response to the company's decision not to renew its contract with the Department of Defense. But it is not clear to the public who would oversee the implementation of the principles, and no ethics board has been named.

Google was not alone. Other companies, including Microsoft, Facebook, and police body camera maker Axon, also assembled ethics boards, advisors, and teams.[149] In addition, technical membership organizations moved to update several of their ethical codes. The IEEE reworked its code of ethics to reflect the challenges of AI and autonomous systems, and researchers in the Association for Computing Machinery (ACM) called for a restructuring of peer review processes, requiring the authors of technical papers to consider the potential adverse uses of their work, which is not a common practice.[150] Universities including Harvard, NYU, Stanford, and MIT offered new courses on the ethics and ethical AI development practices aimed at identifying issues and considering the ramifications of technological innovation before it is implemented at scale.[151] The University of Montreal launched a wide-ranging process to formulate a declaration for the responsible development of AI that includes both expert summits and open public deliberations for input from citizens.[152]

Such developments are encouraging, and it is noteworthy that those at the heart of AI development have declared they are taking ethics seriously. Ethical initiatives help develop a shared language with which to discuss and debate social and political concerns. They provide developers, company employees, and other stakeholders a set of high-level value statements or objectives against which actions can be later judged. They are also educational, often doing the work of raising awareness of particular risks of AI both within a given institution, and externally, amongst the broader concerned public.[153]

However, developing socially just and equitable AI systems will require more than ethical language, however well-intentioned it may be. We see two classes of problems with this current approach to ethics. The first has to do with enforcement and accountability. Ethical approaches in industry implicitly ask that the public simply take corporations at their word when they say they will guide their conduct in ethical ways. While the public may be able to compare a post hoc decision made by a company to its guiding principles, this does not allow insight into decision making, or the power to reverse or guide such a decision. In her analysis of Google's AI Principles, Lucy Suchman, a pioneering scholar of human computer interaction, argues that without "the requisite bodies for deliberation, appeal, and redress" vague ethical principles like "don't be evil" or "do the right thing" are "vacuous."[154]

This "trust us" form of corporate self-governance also has the potential to displace or forestall more comprehensive and binding forms of governmental regulation. Ben Wagner of the Vienna

University of Economics and Business argues, "Unable or unwilling to properly provide regulatory solutions, ethics is seen as the "easy" or "soft" option which can help structure and give meaning to existing self-regulatory initiatives."[155] In other words, ethical codes may deflect criticism by acknowledging that problems exist, without ceding any power to regulate or transform the way technology is developed and applied. The fact that a former Facebook operations manager claims, "We can't trust Facebook to regulate itself," should be taken into account when evaluating ethical codes in industry.[156]

A second problem relates to the deeper assumptions and worldviews of the designers of ethical codes in the technology industry. In response to the proliferation of corporate ethics initiatives, Greene et al. undertook a systematic critical review of high-profile "vision statements for ethical AI."[157] One of their findings was that these statements tend to adopt a technologically deterministic worldview, one where ethical agency and decision making was delegated to experts, "a narrow circle of who can or should adjudicate ethical concerns around AI/ML" on behalf of the rest of us. These statements often assert that AI promises both great benefits and risks to a universal humanity, without acknowledgement of more specific risks to marginalized populations. Rather than asking fundamental ethical and political questions about whether AI systems should be built, these documents implicitly frame technological progress as inevitable, calling for better building.[158]

Empirical study of the use of these codes is only beginning, but preliminary results are not promising. One recent study found that "explicitly instructing [engineers] to consider the ACM code of ethics in their decision making had no observed effect when compared with a control group."[159] However, these researchers did find that media or historical accounts of ethical controversies in engineering, like Volkswagen's Dieselgate, may prompt more reflective practice.

Perhaps the most revealing evidence of the limitations of these emerging ethical codes is how corporations act after they formulate them. Among the list of applications Google promises not to pursue as a part of its AI Principles are "technologies whose purpose contravenes widely accepted principles of international law and human rights."[160] That was tested earlier this year after investigative journalists revealed that Google was quietly developing a censored version of its search engine (which relies extensively on AI capabilities) for the Chinese market, code-named Dragonfly.[161] Organizations condemned the project as a violation of human rights law, and as such, a violation of Google's AI principles. Google employees also organized against the effort.[162] As of writing, the project has not been cancelled, nor has its continued development been explained in light of the clear commitment in the company's AI Principles, although Google's CEO has defended it as "exploratory."[163]

There is an obvious need for accountability and oversight in the industry, and so far the move toward ethics is not meeting this need. This is likely in part due to the market-driven incentives working against industry-driven implementations: a drastic (if momentary) drop in Facebook and Twitter's share price occurred after they announced efforts to combat misinformation and increase spending on security and privacy efforts.[164]

This is no excuse not to pursue a more ethically driven agenda, but it does suggest that we should be wary of relying on companies to implement ethical practices voluntarily, since many of the incentives governing these large, publicly traded technology corporations penalize ethical action. For these mechanisms to serve as meaningful forms of accountability requires that external oversight and transparency be put into place to ensure that there exists an external system of checks and balances in addition to the cultivation of ethical norms and values within the engineering profession and technology companies.

3. WHAT IS NEEDED NEXT

When we released our AI Now 2016 Report, fairness formulas, debiasing toolkits, and ethical guidelines for AI were rare. The fact that they are commonplace today shows how far the field has come. Yet much more needs to be done. Below, we outline seven strategies for future progress on these issues.

3.1 From Fairness to Justice

Any debate about bias and fairness should approach issues of power and hierarchy, looking at who is in a position to produce and profit from these systems, whose values are embedded in these systems, who sets their "objective functions," and which contexts they are intended to work within.[165] Echoing the Association for Computing Machinery (ACM) researcher's call for an acknowledgement of "negative implications" as a requirement for peer review, much more attention must be paid to the ways that AI can be used as a tool for exploitation and control.[166] We must also be cautious not to reframe political questions as technical concerns.[167]

When framed as technical "fixes," debiasing solutions rarely allow for questions about the appropriateness or efficacy of an AI system altogether, or for an interrogation of the institutional context into which the "fixed" AI system will ultimately be applied. For example, a "debiased" predictive algorithm that accurately forecasts where crime will occur, but that is being used by law enforcement to harass and oppress communities of color, is still an essentially unfair system.[168] To this end, our definitions of "fairness" must expand to encompass the structural, historical, and political contexts in which an algorithmic systems is deployed.

Furthermore, fairness is a term that can be easily co-opted: important questions such as "Fair to whom? And in what context?" should always be asked. For example, making a facial recognition system perform equally on people with light and dark skin may be a type of technical progress in terms of parity, but if that technology is disproportionately used on people of color and low-income communities, is it really "fair?" This is why definitions of fairness face a hard limit if they remain purely contained within the technical domain: in short, "parity is not justice."[169]

3.2 Infrastructural Thinking

In order to better understand and track the complexities of AI systems, we need to look beyond the technology and the hype to account for the broader context of how AI is shaping and shaped by social and material forces. As Edwards et al. argue: "When dealing with infrastructures, we need to look to the whole array of organizational forms, practices, and institutions which accompany, make possible, and inflect the development of new technology."[170] Doing so requires both experimental methodological approaches and theory building, expanding beyond narrow analyses of individual systems in isolation to consider them on a local and global scale. It also requires considering ways in which technologies are entangled in social relations, material dependencies, and political purposes.[171]

In "Anatomy of an AI System," a 2018 essay and large-scale map, AI Now cofounder Kate Crawford and Professor Vladan Joler took a single Amazon Echo and analyzed all the forms of environmental and labor resources required to develop, produce, maintain, and finally dispose of this sleek and seemingly simple object. When you ask Alexa to play your favorite song, you have drawn on a massive interlinked chain of extractive processes. It involves lithium mining in Bolivia, clickworkers creating large-scale training datasets in southeast Asia, container ships and international logistics, and vast data extraction and analysis by Alexa Voice Service (AVS) across distributed data centers. The process ends in the final resting place of all AI consumer gadgets: in e-waste rubbish heaps in Ghana, Pakistan, and China.

The "Anatomy of an AI System" project points to approaches we can employ in contending with the global implications of AI, and the multi-layered nature of value extraction and exploitation from the developing world to the developed world. This helps to illuminate the darker corners that are rarely considered in analysis of AI systems.[172]

In particular, an infrastructural analysis of AI shows that there are black boxes within black boxes: not just at the algorithmic level, but also at the levels of trade secrecy laws, labor practices, and untraceable global supply chains for rare earth minerals used to build consumer AI devices. These obscure not only the material impacts of AI systems, but the intensive human work of maintaining and repairing them through practices like content moderation and data training.[173] As Nick Seaver puts it, "If you cannot see a human in the loop, you just need to look for a bigger loop."[174]

Only by tracing across these sociotechnical layers can we understand what we are calling the "full stack supply chain" of AI—the human and nonhuman components that make up the global scale of AI systems. There are many sociotechnical data infrastructures needed to make AI function: these include training data, test data, APIs, data centers, fiber networks, undersea cables, energy use, labor involved in content moderation and training set creation, and a constant reliance on

clickwork to develop and maintain AI systems. We cannot see the global environmental and labor implications of these tools of everyday convenience, nor can we meaningfully advocate for fairness, accountability, and transparency in AI systems, without an understanding of this full stack supply chain.

3.3 Accounting for Hidden Labor in AI Systems

Another emerging research area where we expect to see greater impact focuses on the underpaid and unrecognized workers who help build, maintain, and test AI systems. This hidden human labor takes many forms—from supply chain work, to digital crowdsourced "clickwork," to traditional service industry jobs. Hidden labor exists at all stages of the AI pipeline, from producing and transporting the raw minerals required to create the core infrastructure of AI systems, to providing the invisible human work that often backstops claims of AI "magic" once these systems are deployed in products and services.[175] Communications scholar Lilly Irani refers to such hidden labor as "human-fueled automation."[176] Her research draws attention to the experiences of clickworkers or "microworkers" who perform the repetitive digital tasks that underlie AI systems, like labeling training data and reviewing flagged content, as "workers hidden in the technology."[177]

While this labor is essential to making AI systems "work," it is usually very poorly compensated. A 2018 study from the United Nations' International Labor Organization (ILO) surveyed 3,500 microworkers from 75 countries who routinely offered their labor on popular microtask platforms like Mechanical Turk, Crowdflower, Microworker, and Clickworker. The report found that a substantial number of people earned below their local minimum wage (despite 57% of respondents having advanced degrees specializing in science and technology).[178] Similarly, those who do content moderation work, screening problematic content posted on social media platforms and news feeds, are also paid poorly, in spite of their essential and emotionally difficult labor.[179]

This has not been lost on some in the technical AI research community, who have begun to call attention to the crucial and marginalized role of this labor, and to consider their own responsibility to intervene. Silberman and others discuss how researchers conducting AI studies are increasingly dependent upon cheap crowdsourced labor.[180] They note that, between the years 2008 and 2016, the term "crowdsourcing" went from appearing in less than 1,000 scientific articles to over 20,000. With online microworkers unregulated by current labor laws, researchers are being asked to reconsider what counts as "ethical conduct" in the AI research community. Silberman et al. argue for treating crowdworkers as coworkers, paying them minimum wage determined by the client's location, and the need for additional Institutional Review Board (IRB) oversight.

The practice of examining hidden human labor draws on a lineage of feminist research. The concept of "invisible work," for instance, originated with studies of unpaid women's care work and investigations into organizational settings that relied upon "emotional labor," particularly traditionally "feminized" fields like nursing and flight attendants.[181] Researchers found that common activities taken on by female workers, such as soothing anxious patients or managing unruly customers, were not formally recognized or compensated as work, in spite of their being essential. The feminist legacy of invisible work is useful for contextualizing these new forms of labor, and in understanding the characterization of this work, which, while essential, is often written out of the AI narrative, rarely counted or compensated.

In her article, "The Automation Charade," Astra Taylor proposes the term "fauxtomation" to call attention to the gap between the marketing rhetoric of AI as a seamless product or service and the messy, lived reality of automation, which frequently relies on such unsung human labor. "Automation," Taylor cautions, "has an ideological function as well as a technological dimension."[182] In making this case, she critiques popular narratives around the future of labor, which posit a near-horizon where workers will be replaced by machines. She sees such claims as functioning to disempower workers: what leverage do workers have to demand better wages and benefits in the face of impending automation? We saw this narrative deployed in 2016 by former McDonald's CEO Ed Rensi, who cited the growing "Fight for $15" movement as the impetus for the company's introduction of automated kiosks to replace cashiers.[183] Workers who fought for better pay would ultimately be worse off, he reasoned, as their demand for living wages would force the company to automate and eliminate them. Examining his claim two years on, we see that this is not entirely true. Automation or no, workers are still needed: after McDonald's added kiosks to its Chicago flagship store, the location reopened with more employees than before the kiosks were introduced.[184]

The integration of automation and AI in the workplace is aimed not only at automating worker tasks, but at managing, monitoring, and assessing workers themselves. Alex Rosenblatt's 2018 ethnography of Uber drivers details the precarity and uncertainty produced by depending on the whims of a centralized, AI-enabled platform for one's livelihood. The algorithmic logic that governs ride-sharing applications can arbitrarily bar drivers from work, result in unreliable wages and unexpected costs, and nudge people into working longer hours, resulting in unsafe driving conditions.[185] Such platforms isolate workers from each other, making concerted activity and labor organizing difficult. They also function to create significant information asymmetries between data-rich companies aiming to extract value from workers, and the workers themselves. Even so, 2018 has seen increasing dissent from such workers. Some prominent examples of worker-driven protest include on-demand delivery riders striking alongside UK fast food industry employees and rideshare drivers calling for job protections.[186]

Silicon Valley contractors working in security, food, and janitorial services within major technology companies have also organized, seeking a living wage and other protections.[187] They are among thousands of workers who labor alongside their full-time technology worker peers, but are classified as independent contractors. Under this designation, they are often paid low wages, and

provided few benefits and protections. They are also rarely counted in official employee numbers, even though they make up a large portion of most technology industry workforces, and perform essential work. For example, as of this year, contract workers outnumber Google's direct employees for the first time in the company's history.[188] This increasing wave of dissent makes visible the social tensions at the heart of the practice of hiding and marginalizing important forms of labor.

The physical, emotional, and financial costs of treating workers like "bits of code" and devaluing their work and well-being has been highlighted in recent news articles describing the conditions of Amazon warehouse workers and contracted Prime delivery drivers.[189] Amazon warehouse workers recently went on strike in Europe, protesting harsh conditions. According to one striking worker, "You start at the company healthy and leave it as a broken human," with many workers requiring surgeries related to workplace conditions.[190]

Recognizing all of the labor required to "make AI work" can help us better understand the implications of its development and use. Research in these areas also helps us reexamine the focus on technical talent in narratives describing AI's creation and recognize that technical skills account for only a portion of a much larger effort. This enables us to question numerous labor policies, such as the focus on pushing workers to acquire coding or data science skills as a way to ensure they are counted and compensated. They also help us identify who is likely to benefit, and who, along the AI production and deployment pipeline, is likely to be harmed.

3.4 Deeper Interdisciplinarity

AI researchers and developers are engaged in building technologies that have significant implications for diverse populations in broad fields like law, sociology, and medicine. Yet much of this development happens far removed from the experience and expertise of these groups. This has led to a call to expand the disciplinary makeup of those engaged in AI design, development, and critique, beyond purely technical expertise.[191] Since then, we have seen some movement in this direction. Recently, MIT announced plans to establish a new college of computing that aims to "advance pioneering work on AI's ethical use and societal impact" by fostering integrated cross-disciplinary training, "educating the bilinguals of the future," as MIT President L. Rafael Reif described it.[192]

Such initiatives are critical: as AI becomes more deeply embedded in areas like healthcare, criminal justice, hiring, housing, and educational systems, experts from these domains are essential if we are to ensure AI works as envisioned. In integrating these disciplinary perspectives, it is important that they are not merely "languages" to be acquired by computer scientists and engineers seeking to expand their work into new areas—especially when other disciplines have been leading that work. Instead, social science and the humanities should be centered as contributors to the AI field's foundational knowledge and future direction, enabling us to leverage new modes of analysis and methodological intervention.[193]

3.5 Race, Gender and Power in AI

This year, a groundswell of political action emerged around issues of discrimination, harassment, and inequity in the technology industry, especially in the AI field.[194] This rising concern weaves together a number of related issues, from the biases in AI systems, to failed diversity and inclusion efforts within industry and academia, to the grassroots efforts to confront sexual harassment and the abuse of power in workplaces and classrooms.

Resonating with the broader #MeToo movement, we saw issues relating to diversity and inclusion in artificial intelligence rise on the public agenda:

- Following the 2017 Conference on Neural Information Processing Systems, members of the artificial intelligence and machine learning communities began voicing concerns about long standing problems of harassment and discrimination in conference settings, leading to #ProtestNIPS, a movement aimed at highlighting examples of toxicity in the community and the need to address them.[195] Among other things, this provoked a change to the conference acronym, a longstanding subject of sensitivity for its gendered and historical connotations. The conference, which was previously referred to as NIPS, now goes by NeurIPS.[196]

- We also saw renewed focus on initiatives devoted to creating platforms for inclusion in the field, such as Black in AI, Women in Machine Learning, Latinx in AI, and Queer in AI, alongside the appointment of Diversity and Inclusion chairs and a series of other changes to the design of NeurIPS intended to foster equity and inclusion among participants.[197]

- Across the industry, we saw a growing technology worker movement that intersected with these issues. The Google Walkout, in particular, took on a worker-driven agenda that acknowledged that race, class, and sexuality are intertwined with forms of gender-based discrimination. The walkout explicitly aimed to center the needs of the company's temporary contract workers and vendors who lack the job security and benefits of more privileged technology workers.[198] These efforts have led to some significant structural changes—notably, the end to forced arbitration for sexual harassment claims across a number of the largest companies in the AI industry.[199]

- In other arenas, corporate boards have ignored or otherwise refused to address shareholder proposals targeting discriminatory workspaces. This year, Google dismissed a proposal that would tie executive compensation to progress made on diversity and inclusion, while in 2016, Apple refused a mandate that would require it to diversify its board and senior management.[200]

Across these efforts, advocates of diversity in AI are finding intersections between the move to address gender and race-based harassment and abuse within the technology community, and other forms of inequity and abuses of power. But this is still an uphill battle: while there is increased attention to problems of bias in AI systems, we have yet to see much research within the fairness and bias debate focused on the state of equity and diversity in the AI field itself. Indeed, reliable figures on representation in AI are difficult to come by, although some limited data does exist.

A recent estimate produced by *WIRED* and Element AI found that only 12% of researchers who contributed to the three leading machine learning conferences in 2017 were women. This gender gap is replicated at large technology firms like Facebook and Google, whose websites show that only 15% and 10% of their AI research staff are women.[201] And there is no reliable data on the state of racial diversity in the field, or retention rates for people of color.[202] Collectively, the limited evidence suggests that AI, as a field, is even less diverse than computer science as a whole, which is itself at a historic low point: women make up only 18% of computer science majors in the United States, a decline from a high point of 37% in 1984.[203]

These trends are even more dramatic when compared to other STEM fields in which gender diversity has shown a marked improvement.[204] Yet these are not new problems: the *WIRED*/Element AI survey is not significantly different from a study of the AI field that was published by *IEEE Expert* in 1992, which found that only 13% of published authors in the journal over the prior four years were women.[205] And in the 1980s, female grad students at MIT's Computer Science and Artificial Intelligence Labs thoroughly documented their experiences with toxic working environments in the report "Barriers to Equality in Academia: Women in Computer Science at MIT."[206]

It is time to address the connection between discrimination and harassment in the AI community, and bias in the technical products that are produced by the community. Scholars in science and technology studies have long observed that the values and beliefs of those who create technologies shape the technologies they create.[207] Expanding the field's frame of reference to recognize this connection will ensure it is better equipped to address the problems raised by its rapid proliferation into sensitive social domains. As one AI researcher put it, "Bias is not just in our datasets, it's in our conferences and community."[208]

A recent example illustrates these connections, and how discriminatory practices within the culture that produces an AI system can be mirrored and amplified in the system itself. Amazon recently developed an experimental AI system to help it rank job candidates. It trained the system on data reflecting the company's historical hiring preferences, hoping to more efficiently identify qualified candidates.[209] But the system didn't work as expected: based on the company's historical hiring, it showed a distinct bias against women candidates, downgrading resumes from candidates who attended two all-women's colleges, and even penalizing resumes that contained the word "woman." After uncovering this bias, the company attempted to fix the system, adjusting the algorithm to treat these terms more fairly. This did not work, and the project was eventually

scrapped. Gender-based discrimination was embedded too deeply within the system – a system built to reflect Amazon's past hiring practices – to be uprooted using the "debiasing" approach commonly adopted within the AI field.

As scholars like Safiya Noble and Mar Hicks have observed, there is a clear through-line connecting longstanding patterns of discrimination and harassment in AI to the ways artificial intelligence technologies can amplify and contribute to marginalization and social inequity.[210] Patterns of cultural discrimination are often embedded in AI systems in complex and meaningful ways, and we need to better understand how these effects are felt by different communities.[211]

This is a space that has too long been overlooked and where research is sorely needed. AI Now will be publishing a dedicated report on these issues, and we have a multi-year research project dedicated to examining these challenges.

3.6 Strategic Litigation and Policy Interventions

This year saw an increase in court challenges to the use of automated systems, particularly when government agencies use them in decisions that affect individual rights. In a recent AI Now Report called "Litigating Algorithms," we documented five recent case studies of litigation involving the use of automated systems: in Medicaid and disability benefits cases, public teacher employment evaluations, juvenile criminal risk assessment, and criminal DNA analysis.[212] The findings brought to light several emerging trends. First, these cases provided concrete evidence that governments are routinely adopting automated decision systems (ADS) as measures to produce "cost savings" or to streamline work. Yet, they are failing to assess how these systems might disproportionately harm the populations they are meant to serve, particularly those who are the most vulnerable and who have little recourse or even knowledge that these systems are deeply affecting their lives. In many cases, there was not a single government employee who could explain the automated decision, correct errors, or audit the results of its determination. Through a series of vendor and contractor agreements, almost all avenues for understanding or contesting the impact of these systems were shielded by legal protections such as trade secret law.

Second, few government agencies had invested real efforts to ensure that fairness and due process protections remained in place when switching from human-driven decisions to algorithmically-driven ones. The typical audit, appeals, and accountability mechanisms were totally absent from automated system design. Fortunately, successful strategic litigation by lawyers from the American Civil Liberties Union (ACLU) of Idaho, Legal Aid of Arkansas, the Houston Federation of Teachers, The Legal Aid Society of New York, and various public defenders were able to secure victories for their clients and challenge these unlawful uses based, in part, on constitutional and administrative due process litigation claims.

The playbook for how to litigate algorithms is still being written, but our report uncovered several useful strategies to support long-term solutions and protections. First, arguments based on procedural due process presented serious challenges to the trade secrecy claims of private vendors, with the vast majority of judges ruling that the right to assert constitutional or civil rights protections outweighs any risk of intellectual property misappropriation. Second, a failure to notify affected individuals and communities matters: agencies who neglected to engage community groups concerning the use of these systems were often judged to have failed to appropriately provide the opportunity for public notice and comment, meaning that their implementation of AI systems was potentially unconstitutional. Third, interdisciplinary collaboration is important when trying to determine where these systems fail, especially when submitting evidence to judges. In cases in which lawyers worked closely with technical and social science experts, judges were able to learn about the scientific flaws in these systems as well as the social ramifications and harms.

Looking forward, we anticipate future strategic litigation cases will produce many more lessons. These interventions generate greater understanding and remedial accountability for these systems, even in situations where government agencies have attempted to disclaim ownership, understanding, or control. Combined with tools such as AI Now's Algorithmic Impact Assessment framework, alongside robust regulatory oversight regimes, we can begin to identify, measure, and, when necessary, intervene in efforts to use AI and automated systems in ways that produce harm.[213] However, in order to continue to build on recent progress, lawyers and community activists who represent individuals in such suits need greater funding and support, as well as networks of domain experts that they can draw on to help advise strategy and audit systems.

3.7 Research and Organizing: An Emergent Coalition

The rapid deployment of AI and related systems in everyday life is not a concern for the future—it is already here, with no signs of slowing down. Recognizing this, a set of strategies have emerged, drawing on long-standing traditions of activism and organizing to demand structural changes for greater accountability.

Social activism by technologists is nothing new. In the early 1980s, Computer Professionals for Social Responsibility formed to oppose the use of computers in warfare.[214] More recently, the 2016 "Never Again" technology pledge rallied thousands of workers in various technology sectors to sign a promise not to build databases or conduct data collection that could be used to target religious minorities or facilitate mass deportations.[215] While 2018's organizing and activism draws from a long tradition, its scale is new to the technology sector. Technology workers are joining forces with civil society organizations and researchers in opposition to their employers' technical and business decisions.

Google employees kicked off publicly visible organizing in 2018, opposing Project Maven, a Pentagon effort to apply Google's machine vision AI capabilities to Department of Defense drone surveillance.[216] Researchers and human rights organizations joined the cause, and in June,

Google announced it would abandon the project.[217] At Amazon, Salesforce, and Microsoft, employees petitioned their leadership to end contracts with Immigrations and Customs Enforcement (ICE), supported by immigration and advocacy organizations.[218] Amazon employees also joined the ACLU in petitioning the company to stop selling facial recognition to law enforcement, responding to the ACLU's work exposing existing contracts.[219] Following Maven, Google employees again rose up against Project Dragonfly, a version of the Google search engine enabling government-directed censorship and surveillance, planned for the Chinese market.[220] In response to media reports that disclosed the secretive effort, employees requested ethical oversight and accountability, and over 700 of them joined Amnesty International in a call to cancel the project, signing their name publicly to an open letter which coincided with Amnesty International protests[221]

The biggest moment occurred in early November, when 20,000 Google workers walked out around the globe in an action called Walkout for Real Change.[222] The walkout characterized Google as a company at which "abuse of power, systemic racism, and unaccountable decision-making are the norm."[223] Organizers called on leadership to meet five demands, including ending pay and opportunity inequity, eliminating forced arbitration in cases of sexual harassment and discrimination, and adding an employee representative to the board of directors. A week after the walkout, Google met a small portion of these demands, agreeing to end forced arbitration in cases of sexual harassment (but notably ignoring discrimination).[224] This move was quickly replicated throughout the industry, with Facebook, Square, eBay, and Airbnb following suit.[225]

By joining forces with researchers and civil society groups, this new wave of labor organizing mirrors calls for greater diversity and openness within the AI research domain.[226] These movements are incorporating diverse perspectives across class, sector, and discipline, working to ensure they are capable of understanding the true costs of company practices, including the impact of the systems they build. The Google workers who participated in the walkout expanded their coalition across class and sector, emphasizing contract workers in their demands, and situating themselves within a growing movement "not just in tech, but across the country, including teachers, fast-food workers and others who are using their strength in numbers to make real change."[227]

The recent surge in activism has largely been driven by whistleblowers within technology companies, who have disclosed information about secretive projects to journalists.[228] These disclosures have helped educate the public, which is traditionally excluded from such access, and helped external researchers and advocates provide more informed analysis. By establishing shared ground truth, whistleblowing has helped build the broad coalitions that characterize these movements. The critical role of ethical whistleblowing over the last year has also highlighted both its social importance, and the lack of protections for those who make such disclosures.

The broad coalition of technology worker organizers, researchers, and civil society is playing an increasing role in the push for accountability in the technology sector. Many engineering employees have considerable bargaining power and are uniquely positioned to demand change

from their employers.[229] Applying this power to push for greater accountability presents a hopeful model for labor organizing in the public interest, especially given the current lack of government regulation, external oversight, and other meaningful levers capable of reviewing and steering technology company decision making.

CONCLUSION

This year saw AI systems rapidly introduced into more social domains, leaving increasing numbers of people at risk. While AI techniques still offer considerable promise, rapid deployment of systems without appropriate assessment, accountability, and oversight can create serious hazards. We urgently need to regulate AI systems sector-by-sector, with particular attention paid to facial and affect recognition, and to inform those policies with rigorous research.

But regulation can only be effective if the legal and technological barriers that prevent auditing, understanding, and intervening in these systems are removed. Back in 2016, we recommended in the first AI Now report that the Computer Fraud and Abuse Act (CFAA) and the Digital Millennium Copyright Act (DMCA) should not be used to restrict research into AI accountability and auditing.[230] This year, we go further: AI companies should waive trade secrecy and other legal claims that would prevent algorithmic accountability in the public sector. Governments and public institutions must be able to understand and explain how and why decisions are made, particularly when people's access to healthcare, housing, and employment is on the line.

The question is no longer whether there are harms and biases in AI systems. That debate has been settled: the evidence has mounted beyond doubt in the last year. The next task now is addressing these harms. This is particularly urgent given the scale at which these systems are deployed, the way they function to centralize power and insight in the hands of the few, and the increasingly uneven distribution of costs and benefits that accompanies this centralization. We need deeper analyses of the "full stack supply chain" behind AI systems, to track their development and deployment across the product life cycle, and to take into account their true environmental and labor costs.[231]

Furthermore, it is long overdue for technology companies to directly address the cultures of exclusion and discrimination in the workplace. The lack of diversity and ongoing tactics of harassment, exclusion, and unequal pay are not only deeply harmful to employees in these companies but also impacts the AI products they release, producing tools that perpetuate bias and discrimination.[232]

The current structure within which AI development and deployment occurs works against meaningfully addressing these pressing issues. Those in a position to profit are incentivized to accelerate the development and application of systems without taking the time to build diverse teams, create safety guardrails, or test for disparate impacts. Those most exposed to harm from

these systems commonly lack the financial means and access to accountability mechanisms that would allow for redress or legal appeals.[233] This is why we are arguing for greater funding for public litigation, labor organizing, and community participation as more AI and algorithmic systems shift the balance of power across many institutions and workplaces.

It is imperative that the balance of power shifts back in the public's favor. This will require significant structural change that goes well beyond a focus on technical systems, including a willingness to alter the standard operational assumptions that govern the modern AI industry players. The current focus on discrete technical fixes to systems should expand to draw on socially-engaged disciplines, histories, and strategies capable of providing a deeper understanding of the various social contexts that shape the development and use of AI systems.

As more universities turn their focus to the study of AI's social implications, computer science and engineering can no longer be the unquestioned center, but should collaborate more equally with social and humanistic disciplines, as well as with civil society organizations and affected communities.

Fortunately, we are beginning to see new coalitions form between researchers, activists, lawyers, concerned technology workers, and civil society organizations to support the oversight, accountability, and ongoing monitoring of AI systems. For these important connections to grow, more protections are needed, including a commitment from technology companies to provide protections for conscientious objectors who do not want to work on military or policing contracts, along with protections for employees involved in labor organizing and ethical whistleblowers.[234]

The last year revealed many of the hardest challenges for accountability and justice as AI systems moved deeper into the social world. Yet there have been extraordinary moments of potential, as well as significant public debates and hopeful forms of protest, that may ultimately illuminate the pathways for consequential and positive change.

ENDNOTES

1. As AI pioneers Stuart Russell and Peter Norvig point out, the history of artificial intelligence has not produced a clear definition of AI, but can be seen as variously emphasizing four possible goals: "systems that think like humans, systems that act like humans, systems that think rationally, systems that act rationally." In this report we use the term AI to refer to a broad assemblage of technologies, from early rule-based algorithmic systems to deep neural networks, all of which rely on an array of data and computational infrastructures. These technologies span speech recognition, language translation, image recognition, predictions, and determinations—tasks that have traditionally relied on human capacities across the four goals Russell and Norvig identify. While AI is not new, recent developments in the ability to collect and store large quantities of data, combined with advances in computational power have led to significant breakthroughs in the field over the last ten years, along with a strong push to commercialize these technologies and apply them across core social domains. See: Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, (Englewood Cliffs, NJ: Prentice Hall, 1995), 2.

2. Carole Cadwalladr and Emma Graham-Harrison, "Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach," *The Guardian*, March 17, 2018, https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

3. Guy Rosen, "Security Update," *Facebook Newsroom*, September 28, 2018, https://newsroom.fb.com/news/2018/09/security-update/.

4. Josh Eidelson, "Facebook Tools Are Used to Screen Out Older Job Seekers, Lawsuit Claims," Bloomberg, May 29, 2018, https://www.bloomberg.com/news/articles/2018-05-29/facebook-tools-are-used-to-screen-out-older-job-seekers-lawsuit-claims.

5. Bloomberg Editorial Board, "Think the U.S. Has a Facebook Problem? Look to Asia," *Bloomberg*, October 22, 2017, https://www.bloomberg.com/opinion/articles/2017-10-22/facebook-has-a-bigger-problem-than-washington.

6. Andrew Liptak, "The US Government Alleges Facebook Enabled Housing Ad Discrimination," *The Verge*, August 19, 2018, https://www.theverge.com/2018/8/19/17757108/us-department-of-housing-and-urban-development-facebook-complaint-race-gender-discrimination.

7. Elizabeth Weise, "Russian Fake Accounts Showed Posts to 126 Million Facebook Users," *USA TODAY*, October 30, 2017, https://www.usatoday.com/story/tech/2017/10/30/russian-fake-accounts-showed-posts-126-million-facebook-users/815342001/.

8. Hamza Shaban, Craig Timberg, and Elizabeth Dwoskin, "Facebook, Google and Twitter Testified on Capitol Hill. Here's What They Said," *Washington Post*, October 31, 2017, https://www.washingtonpost.com/news/the-switch/wp/2017/10/31/facebook-google-and-twitter-are-set-to-testify-on-capitol-hill-heres-what-to-expect/; Casey Newton, "Mark Zuckerberg's Appearance before European Parliament Yields an Empty Spectacle," *The Verge*, May 22, 2018, https://www.theverge.com/2018/5/22/17381250/mark-zuckerberg-european-parliament-facebook.

9. Drew Harwell, "AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how," *Washington Post*, April 11, 2018, https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/.

10. Kate Conger and Dell Cameron, "Google Is Helping the Pentagon Build AI for Drones," *Gizmodo*, March 6, 2018, https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533.

11. Rick Paulas, "A New Kind of Labor Movement in Silicon Valley," *The Atlantic*, September 4, 2018, https://www.theatlantic.com/technology/archive/2018/09/tech-labor-movement/567808/.

12. Hamza Shaban, "Amazon Employees Demand Company Cut Ties with ICE," *Washington Post*, June 22, 2018, https://www.washingtonpost.com/news/the-switch/wp/2018/06/22/amazon-employees-demand-company-cut-ties-with-ice/; Jacob Kastrenakes, "Salesforce Employees Ask CEO to 'Re-Examine' Contract with Border Protection Agency," *The Verge*, June 25, 2018, https://www.theverge.com/2018/6/25/17504154/salesforce-employee-letter-border-protection-ice-immigration-cbp; Colin Lecher, "The Employee Letter Denouncing Microsoft's ICE Contract Now Has over 300 Signatures," *The Verge*, June 21, 2018, https://www.theverge.com/2018/6/21/17488328/microsoft-ice-employees-signatures-protest.

13. Nikhil Sonnad, "US Border Agents Hacked Their "Risk Assessment" System to Recommend Detention 100% of the Time," *Quartz*, June 26, 2018, https://qz.com/1314749/us-border-agents-hacked-their-risk-assessment-system-to-recommend-immigrant-detention-every-time/.

14. Daisuke Wakabayashi, "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam," *The New York Times*, July 30, 2018, https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

15. Nikhil Sonnad, "A Flawed Algorithm Led the UK to Deport Thousands of Students," *Quartz*, May 3, 2018, https://qz.com/1268231/a-toeic-test-led-the-uk-to-deport-thousands-of-students/.

16. Casey Ross and Ike Swetlitz, "IBM's Watson Recommended 'unsafe and Incorrect' Cancer Treatments," *STAT*, July 25, 2018, https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/.

17. George Joseph and Kenneth Lipp, "IBM Used NYPD Surveillance Footage to Develop Technology That Lets Police Search by Skin Color," *The Intercept*, September 6, 2018, https://theintercept.com/2018/09/06/nypd-surveillance-camera-skin-tone-search/.

18. To see a large-scale timeline of events in 2018, see: Kate Crawford and Meredith Whittaker, "AI in 2018: A Year in Review," *Medium*, October 14, 2018, https://medium.com/@AINowInstitute/ai-in-2018-a-year-in-review-8b161ead2b4e.

19. Jon Evans, "The Techlash," *TechCrunch*, June 17, 2018,

20. "Microsoft Calls for Facial Recognition Technology Rules given 'Potential for Abuse,'" *The Guardian*, July 14, 2018, https://www.theguardian.com/technology/2018/jul/14/microsoft-facial-recognition-technology-rules-potential-for-abuse.

21. Natalie Ram, "Innovating Criminal Justice," *Northwestern University Law Review* 112, no. 4 (February 1, 2018): 659–724, https://scholarlycommons.law.northwestern.edu/nulr/vol112/iss4/2; Rebecca Wexler, "Life, Liberty, and Trade Secrets," *Stanford Law Review* 70, no. 5 (May 2018): 1343–1429, https://www.stanfordlawreview.org/print/article/life-liberty-and-trade-secrets/; Danielle Keats Citron and Frank A. Pasquale, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* 89, no. 1 (2014): 1–33, https://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1318/89WLR0001.pdf.

22. See: Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information,* (Cambridge: Harvard University Press, 2015).

23. D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi, "Winner's Curse? On Pace, Progress and Empirical Rigor," *6th International Conference on Learning Representations* (ICLR), (Vancouver, 2018), https://openreview.net/pdf?id=rJWF0Fywf.

24. Kate Crawford, "The Test We Can—and Should—Run on Facebook," *The Atlantic*, July 2, 2014, https://www.theatlantic.com/technology/archive/2014/07/the-test-we-canand-shouldrun-on-facebook/373819/; Molly Jackman and Lauri Kanerva, "Evolving the IRB: Building Robust Review for Industry Research," *Washington & Lee Law Review Online* 72, no. 8 (June 14, 2016): 442–457; Zoltan Boka, "Facebook's Research Ethics Board Needs to Stay Far Away from Facebook," *Wired*, June 23, 2016, https://www.wired.com/2016/06/facebooks-research-ethics-board-needs-stay-far-away-facebook/.

25. "Sandvig v. Sessions — Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online," September 12, 2017, *American Civil Liberties Union*, https://www.aclu.org/cases/sandvig-v-sessions-challenge-cfaa-prohibition-uncovering-racial-discrimination-online.

26. See: Simone Browne, *Dark Matters: On the Surveillance of Blackness* (Durham: Duke University Press, 2015); Alvaro M. Bedoya, "What the FBI's Surveillance of Martin Luther King Tells Us About the Modern Spy Era," *Slate*, January 18, 2016, https://slate.com/technology/2016/01/what-the-fbis-surveillance-of-martin-luther-king-says-about-modern-spying.html; James Ball, Julian Borger, and Glenn Greenwald, "Revealed: How US and UK Spy Agencies Defeat Internet Privacy and Security," *The Guardian*, September 6, 2013, https://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security; Shoshana Zuboff, "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization," *Journal of Information Technology* 30, no. 1 (March 1, 2015): 75–89, https://doi.org/10.1057/jit.2015.5.

27. Alice Shen, "Facial Recognition Tech Comes to Hong Kong-Shenzhen Border," *South China Morning Post*, July 24, 2018, https://www.scmp.com/news/china/society/article/2156510/china-uses-facial-recognition-system-deter-tax-free-traders-hong.

28. Stephen Chen, "China's Robotic Spy Birds Take Surveillance to New Heights," *South China Morning Post*, June 24, 2018, https://www.scmp.com/news/china/society/article/2152027/china-takes-surveillance-new-heights-flock-robotic-doves-do-they.

29. David Z. Morris, "China Will Block Travel for Those With Bad 'Social Credit,'" *Fortune*, March 18, 2018, http://fortune.com/2018/03/18/china-travel-ban-social-credit/.

30. Nathan Vanderklippe, "Chinese Blacklist an Early Glimpse of Sweeping New Social-Credit Control," *The Globe and Mail*, January 3, 2018, https://www.theglobeandmail.com/news/world/chinese-blacklist-an-early-glimpse-of-sweeping-new-social-credit-control/article37493300/.

31. "China Has Turned Xinjiang into a Police State like No Other," T*he Economist*, May 31, 2018, https://www.economist.com/briefing/2018/05/31/china-has-turned-xinjiang-into-a-police-state-like-no-other.

32. Emily Feng and Louise Lucas, "Inside China's Surveillance State," *Financial Times*, July 20, 2018, https://www.ft.com/content/2182eebe-8a17-11e8-bf9e-8771d5404543.

33. Angus Berwick, "A New Venezuelan ID, Created with China's ZTE, Tracks Citizen Behavior," *Reuters*, November 14, 2018, https://www.reuters.com/investigates/special-report/venezuela-zte/.

34. Nafeez Ahmed, "Pentagon Wants to Predict Anti-Trump Protests Using Social Media Surveillance," *Motherboard*, October 30, 2018, https://motherboard.vice.com/en_us/article/7x3g4x/pentagon-wants-to-predict-anti-trump-protests-using-social-media-surveillance.

35. Karen Hao, "Amazon Is the Invisible Backbone behind ICE's Immigration Crackdown," *MIT Technology Review*, October 22, 2018, https://www.technologyreview.com/s/612335/amazon-is-the-invisible-backbone-behind-ices-immigration-crackdown/.

36. "Who's behind Ice?" (Empower LLC, Mijente, The National Immigration Project, and the Immigrant Defense Project, October 23, 2018), https://mijente.net/wp-content/uploads/2018/10/WHO%E2%80%99S-BEHIND-ICE_-The-Tech-and-Data-Companies-Fueling-Deportations_v3-.pdf.

37. Brendan Shillingford et al., "Large-Scale Visual Speech Recognition," *arXiv preprint [Cs]*, arXiv:1807.05162, July 13, 2018.

38. "Machine Vision Algorithm Learns to Recognize Hidden Facial Expressions," *MIT Technology Review*, November 13, 2015, https://www.technologyreview.com/s/543501/machine-vision-algorithm-learns-to-recognize-hidden-facial-expressions/.

39. Richard T. Gray, *About Face: German Physiognomic Thought from Lavater to Auschwitz* (Detroit: Wayne State University Press, 2004); Sharrona Pearl, *About Faces: Physiognomy in Nineteenth-Century Britain* (Cambridge, MA: Harvard University Press, 2010).

40. Blaise Aguera y Arcas, Margaret Mitchell, and Alexander Todorov, "Physiognomy's New Clothes," *Medium*, May 7, 2017, https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

41. Ruth Leys, "How Did Fear Become a Scientific Object and What Kind of Object Is It?," *Representations* 110, no. 1 (2010): 66–104, https://doi.org/10.1525/rep.2010.110.1.66. Leys has offered a number of critiques of Ekman's research program, most recently in Ruth Leys, *The Ascent of Affect: Genealogy and Critique* (Chicago: University of Chicago Press, 2017).

42. Alan J. Fridlund, *Human Facial Expression: An Evolutionary View* (San Diego: Academic Press, 1994).

43. Lisa Feldman Barrett, "Are Emotions Natural Kinds?," *Perspectives on Psychological Science* 1, no. 1 (March 2006): 28–58, https://doi.org/10.1111/j.1745-6916.2006.00003.x; Erika H. Siegel, Molly K. Sands, Wim Van den Noortgate, Paul Condon, Yale Chang, Jennifer Dy, Karen S. Quigley, and Lisa Feldman Barrett. 2018. "Emotion Fingerprints or Emotion Populations? A Meta-Analytic Investigation of Autonomic Features of Emotion Categories." *Psychological Bulletin* 144 (4): 343–93, https://doi.org/10.1037/bul0000128.

44. For example, despite criticism by the U.S. Government Accountability Office, the Transportation Security Administration invested over one billion dollars in its SPOT program, aimed at identifying potential terrorists based on these behavioral indicators. See: "Aviation Security: TSA Should Limit Future Funding for Behavior Detection Activities" (Washington, DC: U.S. Government Accountability Office, November 13, 2013), https://www.gao.gov/products/GAO-14-159.

45. Jonathan Metzl, *The Protest Psychosis: How Schizophrenia Became a Black Disease* (Boston: Beacon Press, 2009).

46. Mark Lieberman, "Sentiment Analysis Allows Instructors to Shape Course Content around Students' Emotions," *Inside Higher Education*, February 20, 2018, https://www.insidehighered.com/digital-learning/article/2018/02/20/sentiment-analysis-allows-instructors-shape-course-content.

47. Will Knight, "Emotional Intelligence Might Be a Virtual Assistant's Secret Weapon," *MIT Technology Review*, June 13, 2016, https://www.technologyreview.com/s/601654/amazon-working-on-making-alexa-recognize-your-emotions;

48. "Affectiva Automotive AI," *Affectiva*, accessed November 18, 2018, http://go.affectiva.com/auto.

49. Jeff Weiner, "ACLU: Amazon's Face-Recognition Software Matched Members of Congress with Mugshots," *Orlando Sentinel*, July 26, 2018, https://www.orlandosentinel.com/news/politics/political-pulse/os-amazon-rekognition-face-matching-software-congress-20180726-story.html; "Amazon Rekognition Announces Real-Time Face Recognition, Text in Image Recognition, and Improved Face Detection," *Amazon Web Services*, November 21, 2017, https://aws.amazon.com/about-aws/whats-new/2017/11/amazon-rekognition-announces-real-time-face-recognition-text-in-image-recognition-and-improved-face-detection/.

50. Chris Adzima, "Using Amazon Rekognition to Identify Persons of Interest for Law Enforcement," *Amazon Web Services*, June 15, 2017, https://aws.amazon.com/blogs/machine-learning/using-amazon-rekognition-to-identify-persons-of-interest-for-law-enforcement/.

51. Ranju Das, "Image & Video Rekognition Based on AWS" (Amazon Web Services Summit, Seoul, 2018), https://youtu.be/sUzuJc-xBEE?t=1889.

52. Jacob Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots," *American Civil Liberties Union*, July 26, 2018, https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

53. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability and Transparency* (New York, 2018), 77–91, http://proceedings.mlr.press/v81/buolamwini18a.html.

54. Matt Wood, "Thoughts On Machine Learning Accuracy," *AWS News Blog*, July 27, 2018, https://aws.amazon.com/blogs/aws/thoughts-on-machine-learning-accuracy/.

55. Sidney Fussell, "Amazon Accidentally Makes Rock-Solid Case for Not Giving Its Face Recognition Tech to Police," *Gizmodo*, July 27, 2018, https://gizmodo.com/amazon-accidentally-makes-rock-solid-case-for-not-givin-1827934703.

56. Bryan Menegus, "Amazon Breaks Silence on Aiding Law Enforcement Following Employee Backlash," *Gizmodo*, August 11, 2018, https://gizmodo.com/amazon-breaks-silence-on-aiding-law-enforcement-followi-1830321057.

57. Joseph and Lipp, "IBM Used NYPD Surveillance Footage to Develop Technology That Lets Police Search by Skin Color."

58. Jenna Bitar and Jay Stanley, "Are Stores You Shop at Secretly Using Face Recognition on You?," *American Civil Liberties Union*, March 26, 2018, https://www.aclu.org/blog/privacy-technology/surveillance-technologies/are-stores-you-shop-secretly-using-face.

59. John Brandon, "Walmart Will Scan for Unhappy Shoppers Using Facial Recognition (Cue the Apocalypse)," *VentureBeat*, August 9, 2017, https://venturebeat.com/2017/08/09/walmart-will-scan-for-unhappy-shoppers-using-facial-recognition-cue-the-apocalypse/.

60. AG and Legal Workforce Act, H.R. 6417, 115th Cong. (2017-2018), https://www.congress.gov/115/bills/hr6417/BILLS-115hr6417ih.pdf; Securing America's Future Act of 2018, H.R. 4760, 115th Cong. (2017-2018), https://www.congress.gov/115/bills/hr4760/BILLS-115hr4760ih.pdf; Strong Visa Integrity Secures America Act," H.R. 2626, 115th Cong. (2017-2018), https://www.congress.gov/115/bills/hr2626/BILLS-115hr2626ih.pdf; H.R. Security and Immigration Reform Act of 2018, 6136, 115th Cong. (2017-2018), https://www.congress.gov/115/bills/hr6136/BILLS-115hr6136ih.pdf.

61. Biometric Information Privacy Act, 740 ILCS 14/1 (Statutes current through the end of the 2018 Regular Session of the 100th General Assembly), https://advance-lexis-com.proxy.library.nyu.edu/api/document?collection=statutes-legislation&id=urn:contentItem:5C66-0WY1-6YS3-D06V-00000-00&context=1516831.

62. Brian Hofer, "BART Board Approves Surveillance Ordinance, Lake Merritt Development," *KTVU*, September 13, 2018, http://www.ktvu.com/news/bart-board-approves-surveillance-ordinance-lake-merritt-development.

63. "Letter from Nationwide Coalition to Amazon CEO Jeff Bezos Regarding Rekognition," *American Civil Liberties Union*, June 18, 2018, https://www.aclu.org/letter-nationwide-coalition-amazon-ceo-jeff-bezos-regarding-rekognition.

64. Abrar Al-Heeti, "Congress Still Wants Answers from Amazon about Its Facial Recognition Tech," *CNET*, November 29, 2018, https://www.cnet.com/news/congress-still-wants-answers-from-amazon-about-its-facial-recognition-tech/.

65. Woodrow Hartzog and Evan Selinger, "Facial Recognition Is the Perfect Tool for Oppression," *Medium*, August 2, 2018, https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66.

66. Yilun Wang and Michael Kosinski, "Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images," *Journal of Personality and Social Psychology* 114, no. 2 (2018): 246–57, http://dx.doi.org/10.1037/pspa0000098.

67. Frank Pasquale, "When Machine Learning Is Facially Invalid," *Communications of the ACM* 61, no. 9 (August 2018): 25–27, https://doi.org/10.1145/3241367.

68. Kade Crockford, "Massachusetts Should Ban Face Recognition Technology," *WBUR*, August 1, 2018, http://www.wbur.org/cognoscenti/2018/08/01/kade-crockford-face-surveillance-technology-ban. See also: Hartzog and Selinger, "Facial Recognition Is the Perfect Tool for Oppression."

69. Brad Smith, "Facial Recognition Technology: The Need for Public Regulation and Corporate Responsibility," *Microsoft on the Issues*, July 13, 2018, https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/; Sidney Fussell, "Axon CEO Says Face Recognition Isn't Accurate Enough for Body Cams Yet," *Gizmodo*, August 8, 2018, https://gizmodo.com/axon-ceo-says-face-recognition-isnt-accurate-enough-for-1828205723.

70. "Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems" (New York: AI Now Institute, September 2018), https://ainowinstitute.org/litigatingalgorithms.pdf; Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability" (New York: AI Now Institute, April 2018), https://ainowinstitute.org/aiareport2018.pdf; Micah Altman, Alexandra Wood, and Effy Vayena, "A Harm-Reduction Framework for Algorithmic Fairness," *IEEE Security Privacy* 16, no. 3 (May 2018): 34–45, https://doi.org/10.1109/MSP.2018.2701149.

71. Lecher, "A Healthcare Algorithm Started Cutting Care, and No One Knew Why."

72. Ibid.

73. Carlos Gradín, "World Income Inequality Database (WIID3.4)," *United Nations University*, October 27, 2015, https://www.wider.unu.edu/project/wiid-world-income-inequality-database.

74. Serena Lei, "Nine Charts about Wealth Inequality in America," *The Urban Institute*, 2015, http://urbn.is/wealthcharts; Lisa J. Dettling, Joanne W. Hsu, Lindsay Jacobs, Kevin B. Moore, and Jeffrey P. Thompson, "Recent Trends in Wealth-Holding by Race and Ethnicity: Evidence from the Survey of Consumer Finances," (Washington: Board of Governors of the Federal Reserve System, September 27, 2017), https://doi.org/10.17016/2380-7172.2083.

75. Andrew Restuccia, Sarah Ferris, and Helena Bottemiller Evich, "Behind Trump's Plan to Target the Federal Safety Net," *POLITICO*, December 11, 2017, http://politi.co/2AJfCzL.

76. Brakkton Booker, "White House Budget Calls For Deep Cuts To HUD," *NPR.org*, February 13, 2018, https://www.npr.org/2018/02/13/585255697/white-house-budget-calls-for-deep-cuts-to-hud.

77. Alison Kodjak, "Federal Judge Blocks Medicaid Work Requirements In Kentucky," *NPR.org*, June 29, 2018, https://www.npr.org/sections/health-shots/2018/06/29/624807533/federal-judge-blocks-medicaid-work-requirements-in-kentucky; Rachel Garfield et al., "Implications of Work Requirements in Medicaid: What Does the Data Say?," *The Henry J. Kaiser Family Foundation*, June 12, 2018, https://www.kff.org/medicaid/issue-brief/implications-of-work-requirements-in-medicaid-what-does-the-data-say/.

78. Stacy Dean, "President's Budget Would Shift Substantial Costs to States and Cut Food Assistance for Millions," *Center on Budget and Policy Priorities*, May 23, 2017, https://www.cbpp.org/research/food-assistance/presidents-budget-would-shift-substantial-costs-to-states-and-cut-food.

79. David Pegg and Niamh McIntyre, "Child Abuse Algorithms: From Science Fiction to Cost-Cutting Reality," *The Guardian*, September 16, 2018, https://www.theguardian.com/society/2018/sep/16/child-abuse-algorithms-from-science-fiction-to-cost-cutting-reality.

80. David Scharfenberg, "Computers Can Solve Your Problem. You May Not like the Answer," *The Boston Globe*, September 21, 2018, https://apps.bostonglobe.com/ideas/graphics/2018/09/equity-machine; Joe Flood, *The Fires: How a Computer Formula, Big Ideas, and the Best of Intentions Burned Down New York City–and Determined the Future of Cities* (New York: Riverhead Books, 2011).

81. Virginia Eubanks, "We Created Poverty. Algorithms Won't Make That Go Away," *The Guardian*, May 13, 2018, https://www.theguardian.com/commentisfree/2018/may/13/we-created-poverty-algorithms-wont-make-that-go-away;

82. Leo Morales, "Federal Court Rules Against Idaho Department of Health and Welfare in Medicaid Class Action," *ACLU of Idaho*, March 30, 2016, https://www.acluidaho.org/en/news/federal-court-rules-against-idaho-department-health-and-welfare-medicaid-class-action.

83. "Litigating Algorithms."

84. Ibid., 10.

85. Michael Nash, "Examination of Using Structured Decision Making and Predictive Analytics in Assessing Safety and Risk in Child Welfare" (Los Angeles: County of Los Angeles Office of Child Protection, May 4, 2017), http://file.lacounty.gov/SDSInter/bos/bc/1023048_05.04.17OCPReportonRiskAssessmentTools_SDMandPredictiveAnalytics_.pdf.

86. Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018).

87. "Vulnerable Children Predictive Modelling" (Wellington: New Zealand Ministry of Social Development), accessed November 18, 2018, https://www.msd.govt.nz/about-msd-and-our-work/publications-resources/research/predicitve-modelling/.

88. Eubanks, *Automating Inequality*.

89. Margaret Talbot, "The Case Against Cash Bail," *The New Yorker*, August 25, 2015, https://www.newyorker.com/news/news-desk/the-case-against-cash-bail.

90. Sam Levin, "Imprisoned by Algorithms: The Dark Side of California Ending Cash Bail," *The Guardian*, September 7, 2018, https://www.theguardian.com/us-news/2018/sep/07/imprisoned-by-algorithms-the-dark-side-of-california-ending-cash-bail; Maddie Hanna, "What Happened When New Jersey Stopped Relying on Cash Bail," *The Philadelphia Inquirer*, February 16, 2018, http://www2.philly.com/philly/news/new_jersey/new-jersey-cash-bail-risk-assessment-20180216.html.

91. Colleen O'Dea, "Civil Rights Coalition Calls for End to Core Element of NJ Bail Reform," *NJ Spotlight*, July 31, 2018, http://www.njspotlight.com/stories/18/07/30/civil-rights-groups-call-for-end-to-core-element-of-nj-bail-reform/.

92. Jeremy B. White, "California Ended Cash Bail. Why Are So Many Reformers Unhappy About It?," *POLITICO Magazine*, August 29, 2018, https://politi.co/2PPLnyz.

93. 93. Rose Luckin, "Towards Artificial Intelligence-Based Assessment Systems," *Nature Human Behaviour* 1 (March 1, 2017): 1–3, https://doi.org/10.1038/s41562-016-0028.

94. Brent Bridgeman, Catherine Trapani, and Yigal Attali, "Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country," *Applied Measurement in Education* 25, no. 1 (January 2012): 27–40, https://doi.org/10.1080/08957347.2012.635502.

95. Nitin Madnani et al., "Building Better Open-Source Tools to Support Fairness in Automated Scoring," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (Valencia: Association for Computational Linguistics, 2017), 41–52, https://doi.org/10.18653/v1/W17-1605.

96. Daniel T. O'Brien et al., "An Evaluation of Equity in the Boston Public Schools' Home-Based Assignment Policy" (Boston: Boston Area Research Initiative, July 2018), https://news.northeastern.edu/wp-content/uploads/2018/07/BPSHBAP.pdf.

97. Ibid.

98. Scharfenberg, "Computers Can Solve Your Problem. You May Not like the Answer."

99. Ibid.

100. Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability" (New York: AI Now Institute, April 2018), https://ainowinstitute.org/aiareport2018.pdf.

101. "Litigating Algorithms."

102. "Algorithmic Accountability Policy Toolkit" (New York: AI Now Institute, October 2018), https://ainowinstitute.org/aap-toolkit.pdf.

103. Reisman et al., "Algorithmic Impact Assessments."

104. Wakabayashi, "Self-Driving Uber Car Kills Pedestrian in Arizona."

105. Jack Stewart, "Tesla's Self-Driving Autopilot Involved in Another Deadly Crash," *Wired*, March 31, 2018, https://www.wired.com/story/tesla-autopilot-self-driving-crash-california/.

106. "Uber Autonomous-SUV Driver Streamed 'The Voice' Just before Deadly Arizona Crash, Report Says," *Los Angeles Times*, June 22, 2018, https://www.latimes.com/business/autos/la-fi-hy-uber-self-driving-death-20180622-story.html.

107. "Early Rider Program," *Waymo*, accessed November 16, 2018, https://waymo.com/apply/.

108. Aarian Marshall, "Wanna Save Lots of Lives? Put (Imperfect) Self-Driving Cars on the Road, ASAP," *Wired*, November 7, 2017, https://www.wired.com/story/self-driving-cars-rand-report/.

109. Lindsay Moore, "Autonomous Vehicles Continue to Drive Without Rules," *Phoenix New Times*, November 9, 2017, https://www.phoenixnewtimes.com/news/autonomous-vehicles-continue-to-drive-without-rules-9853743.

110. "Uber Autonomous-SUV Driver Streamed 'The Voice' Just before Deadly Arizona Crash, Report Says."

111. Ross and Swetlitz, "IBM's Watson Recommended 'unsafe and Incorrect' Cancer Treatments."

112. Angela Chen, "What the Apple Watch's FDA Clearance Actually Means," *The Verge*, September 13, 2018, https://www.theverge.com/2018/9/13/17855006/apple-watch-series-4-ekg-fda-approved-vs-cleared-meaning-safe.

113. Mark Vermette, "Apple Watch Approval Marks Shift In Device Development And Approvals," *Med Device Online*, October 24, 2018, https://www.meddeviceonline.com/doc/apple-watch-approval-marks-shift-in-device-development-and-approvals-0001; Rajiv Leventhal, "4.4M Patient Records Breached in Q3 2018, Protenus Finds," *Healthcare Informatics Magazine*, November 7, 2018, https://www.healthcare-informatics.com/news-item/cybersecurity/44m-patient-records-breached-q3-2018-protenus-finds.

114. Alex Hern, "Google 'betrays Patient Trust' with DeepMind Health Move," *The Guardian*, November 14, 2018, https://www.theguardian.com/technology/2018/nov/14/google-betrays-patient-trust-deepmind-healthcare-move.

115. Benjamin Herold, "Pearson Tested 'Social-Psychological' Messages in Learning Software, With Mixed Results," *Education Week*, April 17, 2018, http://blogs.edweek.org/edweek/DigitalEducation/2018/04/pearson_growth_mindset_software.html?cmp=SOC-SHR-FB.

116. Rolfe Winkler and Laura Stevens, "New Parents Complain Amazon Baby-Registry Ads Are Deceptive," *Wall Street Journal*, November 28, 2018, https://www.wsj.com/articles/new-parents-complain-amazon-ads-are-deceptive-1543417201.

117. Julia Angwin, Jeff Larson, and Laura Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," *Reuters*, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

118. Shira Mitchell, "Mirror Mirror: Reflections on Quantitative Fairness," 2018, https://shiraamitchell.github.io/fairness/; Arvind Narayanan, *Tutorial: 21 Fairness Definitions and Their Politics*, accessed November 18, 2018, https://www.youtube.com/watch?v=jIXIuYdnyyk.

119. Solon Barocas, Kate Crawford, Aaron Shapiro, Hanna Wallach, 2017, "The Problem with Bias: Allocative Versus Representational Harms in Machine Learning" (Measure, Model, Mix: Computer as Instrument: 9th Annual SIGCIS Conference, Philadelphia, 2017).

120. Kate Crawford, "The Trouble with Bias" (Conference on Neural Information Processing Systems, Long Beach, CA, 2017), https://www.youtube.com/watch?v=fMym_BKWQzk.

121. Ibid.

122. Sam Corbett-Davies and Sharad Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," *arXiv preprint [CS]* arXiv:1808.00023, July 31, 2018.

123. Solon Barocas and Moritz Hardt, "Fairness in Machine Learning" (Conference on Neural Information Processing Systems, Long Beach, CA, 2017), https://mrtz.org/nips17/#/; Narayanan, *21 Fairness Definitions and Their Politics*.

124. Tracy Jan, "Redlining Was Banned 50 Years Ago. It's Still Hurting Minorities Today," *Washington Post*, March 28, 2018, https://www.washingtonpost.com/news/wonk/wp/2018/03/28/redlining-was-banned-50-years-ago-its-still-hurting-minorities-today/.

125. Cynthia Dwork, Nicole Immorlica, Adam T. Kalai, Mark DM Leiserson, "Decoupled classifiers for group-fair and efficient machine learning", Conference on Fairness, Accountability and Transparency (January 21, 2018), 119-133, http://proceedings.mlr.press/v81/dwork18a/dwork18a.pdf.

126. Jon Kleinberg, "Inherent Trade-Offs in Algorithmic Fairness," in *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '18 (New York: ACM, 2018), 40–40, https://doi.org/10.1145/3219617.3219634; Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *arXiv preprint [Cs, Stat]* arXiv:1610.07524, October 24, 2016.

127. Frank Pasquale, "Odd Numbers," *Real Life*, August 20, 2018, https://reallifemag.com/odd-numbers/.

128. Sebastian Benthall, "Critical Reflections on FAT* 2018: A Historical Idealist Perspective," *Dataactive*, April 11, 2018, https://data-activism.net/2018/04/critical-reflections-on-fat-2018-a-historical-idealist-perspective/.

129. Mitchell, "Mirror Mirror: Reflections on Quantitative Fairness."

130. Ben Hutchinson and Margaret Mitchell, "50 Years of Test (Un)Fairness: Lessons for Machine Learning," *arXiv preprint [CS]*, arXiv:1811.10104, November 25, 2018.

131. Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli, "A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics," *arXiv preprint [Cs, Math, Stat]*, arXiv:1807.00553, July 2, 2018.

132. Batya Friedman and Helen Nissenbaum, "Discerning Bias in Computer Systems," in *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*, (New York: ACM, 1993), 141–142, https://doi.org/10.1145/259964.260152.

133. Batya Friedman and Helen Nissenbaum, "Bias in Computer Systems," *ACM Transactions on Information Systems* 14, no. 3 (July 1996): 330–347, https://doi.org/10.1145/230538.230561.

134. Kate Crawford and Ryan Calo, "There Is a Blind Spot in AI Research," *Nature* 538, no. 7625 (October 20, 2016): 311, https://doi.org/10.1038/538311a.

135. See: Meredith Whittaker, "Data Genesis: AI's Primordial Soup" (Eyeo Festival, Minneapolis, 2018), https://vimeo.com/287094149.

136. Timnit Gebru et al., "Datasheets for Datasets," *arXiv preprint [Cs]*, arXiv:1803.09010, March 23, 2018.

137. Margaret Mitchell et al., "Model Cards for Model Reporting," *arXiv preprint [Cs]* arXiv:1810.03993, October 5, 2018.

138. Jevan A. Hutson, Jessie G. Taft, Solon Barocas and Karen Levy,, "Debiasing Desire: Addressing Bias & Discrimination on Intimate Platforms," *Proceedings of the ACM On Human-Computer Interaction (CSCW)* 2 (November 2018): 1–18, https://doi.org/10.1145/3274342.

139. Animesh Singh and Michael Hind, "AI Fairness 360: Attacking Bias from All Angles!," *IBM Developer*, September 19, 2018, https://developer.ibm.com/blogs/2018/09/19/ai-fairness-360-attacking-bias-from-all-angles/.

140. James Wexler, "The What-If Tool: Code-Free Probing of Machine Learning Models," *Google AI Blog*, September 11, 2018, http://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html; James Wexler, "Facets: An Open Source Visualization Tool for Machine Learning Training Data," *Google AI Blog*, June 17, 2017, http://ai.googleblog.com/2017/07/facets-open-source-visualization-tool.html.

141. Alekh Agarwal et al., "A Reductions Approach to Fair Classification," *arXiv preprints [CS]*, arXiv:1803.02453, March 6, 2018.

142. Dave Gershgorn, "Facebook Says It Has a Tool to Detect Bias in Its Artificial Intelligence," *Quartz*, May 3, 2018, https://qz.com/1268520/facebook-says-it-has-a-tool-to-detect-bias-in-its-artificial-intelligence/.

143. Jeremy Kahn, "Accenture Unveils Tool to Help Companies Insure Their AI Is Fair," *Bloomberg*, June 13, 2018, https://www.bloomberg.com/news/articles/2018-06-13/accenture-unveils-tool-to-help-companies-insure-their-ai-is-fair.

144. "The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)," *IEEE Standards Association*, accessed November 19, 2018, https://standards.ieee.org/industry-connections/ecpais.html.

145. "TRUSTe Settles FTC Charges It Deceived Consumers Through Its Privacy Seal Program," *Federal Trade Commission*, November 17, 2014, https://www.ftc.gov/news-events/press-releases/2014/11/truste-settles-ftc-charges-it-deceived-consumers-through-its.

146. "AI Now 2017 Report."

147. Scott Shane, Cade Metz, and Daisuke Wakabayashi, "How a Pentagon Contract Became an Identity Crisis for Google," *The New York Times*, November 2, 2018, https://www.nytimes.com/2018/05/30/technology/google-project-maven-pentagon.html.

148. Sundar Pichai, "AI at Google: Our Principles," *The Keyword*, June 7, 2018, https://www.blog.google/technology/ai/ai-principles/.

149. Alan Boyle, "Microsoft Is Turning down Some Sales over AI Ethics, Top Researcher Eric Horvitz Says," *GeekWire*, April 10, 2018, https://www.geekwire.com/2018/microsoft-cutting-off-sales-ai-ethics-top-researcher-eric-horvitz-says/; Jordan Novet, "Facebook Forms Ethics Team to Prevent Bias in AI Software," *CNBC*, May 3, 2018, https://www.cnbc.com/2018/05/03/facebook-ethics-team-prevents-bias-in-ai-software.html; "Axon AI and Policing Technology Ethics Board," *Axon*, accessed November 19, 2018, https://www.axon.com/info/ai-ethics.

150. "Ethically Aligned Design Version 2: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems" (New York: IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2018), https://ethicsinaction.ieee.org/; Brent Hecht, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Cathy Wu. "It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process," *ACM Future of Computing Blog*, March 29, 2018, https://acm-fca.org/2018/03/29/negativeimpacts/.

151. Alina Tugend, "Colleges Grapple With Teaching the Technology and Ethics of A.I.," *The New York Times*, November 3, 2018, https://www.nytimes.com/2018/11/02/education/learning/colleges-grapple-with-teaching-ai.html.

152. "La Déclaration de Montréal pour un Développement Responsable de l'Intelligence Artificielle," accessed November 19, 2018, https://www.declarationmontreal-iaresponsable.com/.

153. "Microsoft Professional Program for Artificial Intelligence," *Microsoft*, accessed November 29, 2018, https://academy.microsoft.com/en-us/professional-program/tracks/artificial-intelligence/.

154. Lucy Suchman, "Corporate Accountability," *Robot Futures*, June 11, 2018, https://robotfutures.wordpress.com/2018/06/10/corporate-accountability/.

155. Ben Wagner, "Ethics as Escape From Regulation: From Ethics-Washing to Ethics-Shopping?" In *Being Profiling. Cogitas Ergo Sum*, ed. Mireille Hildebrandt. (Amsterdam: Amsterdam University Press, forthcoming 2019), https://www.privacylab.at/wp-content/uploads/2018/07/Ben_Wagner_Ethics-as-an-Escape-from-Regulation_2018_BW9.pdf.

156. Sandy Parakilas, "We Can't Trust Facebook to Regulate Itself," *The New York Times*, January 20, 2018. https://www.nytimes.com/2017/11/19/opinion/facebook-regulation-incentive.html.

157. Dan M. Greene, Anna Laura Hoffman, and Luke Stark, "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning" (Hawaii International Conference on System Sciences, Maui, forthcoming 2019), http://dmgreene.net/wp-content/uploads/2018/09/Greene-Hoffman-Stark-Better-Nicer-Clearer-Fairer-HICSS-Final-Submission.pdf.

158. Ibid.

159. Andrew McNamara, Justin Smith, and Emerson Murphy-Hill, "Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018 (New York: ACM, 2018), 729, https://doi.org/10.1145/3236024.3264833.

160. Pichai, "AI at Google: Our Principles."

161. Ryan Gallagher, "Google Plans to Launch Censored Search Engine in China, Leaked Documents Reveal," *The Intercept*, August 1, 2018, https://theintercept.com/2018/08/01/google-china-search-engine-censorship/.

162. Ronald Deibert, Rebecca Mackinnon, Xiao Qiang, and Lokman Tsui, "Open Letter to Google on Reported Plans to Launch a Censored Search Engine in China," *Amnesty International*, August 28, 2018, https://www.amnesty.org/en/documents/document/?indexNumber=ASA17%2f9001%2f2018&language=en; Kate Conger and Daisuke Wakabayashi, "Google Employees Protest Secret Work on Censored Search Engine for China," *The New York Times*, September 10, 2018, https://www.nytimes.com/2018/08/16/technology/google-employees-protest-search-censored-china.html.

163. Mark Bergen, "Google CEO Tells Staff China Plans Are 'Exploratory' After Backlash," *Bloomberg* August 17, 2018, https://www.bloomberg.com/news/articles/2018-08-17/google-ceo-is-said-to-tell-staff-china-plans-are-exploratory.

164. Matt Phillips, "Facebook's Stock Plunge Shatters Faith in Tech Companies' Invincibility," *The New York Times*, October 17, 2018, https://www.nytimes.com/2018/07/26/business/facebook-stock-earnings-call.html; Rupert Neate, "Twitter Stock Plunges 20% in Wake of 1m User Decline," *The Guardian*, July 27, 2018, https://www.theguardian.com/technology/2018/jul/27/twitter-share-price-tumbles-after-it-loses-1m-users-in-three-months.

165. For a more general description of justice as fairness, see: John Rawls, *Justice as Fairness: A Restatement*, ed. Erin I. Kelly (Cambridge, MA: Harvard University Press, 2001).

166. Brent Hecht et al., "It's Time to Do Something," https://acm-fca.org/2018/03/29/negativeimpacts/.

167. Ben Green, "'Fair' Risk Assessments: A Precarious Approach for Criminal Justice Reform" (5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, 2018), https://scholar.harvard.edu/files/bgreen/files/18-fatml.pdf.

168. Ben Green, "Putting the J(ustice) in FAT," *Berkman Klein Center*, February 26, 2018, https://medium.com/berkman-klein-center/putting-the-j-ustice-in-fat-28da2b8eae6d.

169. Kate Crawford, "Just An Engineer: The Politics of AI," (The Royal Society, London, July 2018), https://www.youtube.com/watch?v=HPopJb5aDyA.

170. Paul N. Edwards, Geoffrey C. Bowker, Steven J. Jackson, and Robin Williams. "Introduction: An Agenda for Infrastructure Studies," Journal of the Association for Information Systems 10, no. 5 (May 28, 2009): 364–74, https://aisel.aisnet.org/jais/vol10/iss5/6.

171. Susan Leigh Star, "The Ethnography of Infrastructure," *American Behavioral Scientist* 43, no. 3 (November 1, 1999): 377–91, https://doi.org/10.1177/00027649921955326; Paul N. Edwards, Thomas J. Misa, and Philip Brey, "Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems," in *Modernity and Technology* (Cambridge, MA: MIT Press, 2003), 185–225; Shannon Mattern, "The Big Data of Ice, Rocks, Soils, and Sediments", *Places,* November 2017, https://placesjournal.org/article/the-big-data-of-ice-rocks-soils-and-sediments/; Jean-Christophe Plantin, Carl Lagoze, Paul N. Edwards and Christian Sandvig. "Infrastructure studies meet platform studies in the age of Google and Facebook," *New Media & Society,* 20, no. 1 (2018): 293-310, https://doi.org/10.1177/1461444816661553.

172. Kate Crawford and Vladan Joler, "Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources," (AI Now Institute and Share Lab, September 7, 2018), https://anatomyof.ai.

173. Steven J. Jackson, "Rethinking Repair", in *Media Technologies: Essays on Communication, Materiality, and Society,* Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot, eds. (Cambridge: MIT Press, 2014), 221-239.

174. Nick Seaver, "What Should An Anthropology of Algorithms Do?" *Cultural Anthropology,* 33, No. 3 (2018): 375-385, https://doi.org/10.14506/ca33.3.04.

175. Li Yuan, "How Cheap Labor Drives China's A.I. Ambitions," *The New York Times*, November 27, 2018, https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html; Mary L. Gray and Siddharth Suri, "The Humans Working Behind the AI Curtain, *Harvard Business Review,* January 9, 2017, https://hbr.org/2017/01/the-humans-working-behind-the-ai-curtain.

176. Lilly Irani, "The Hidden Faces of Automation," *XRDS* 23, no. 2 (December 2016): 34–37, https://doi.org/10.1145/3014390.

177. Ibid.

178. Janine Berg et al., "Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World," (Geneva: International Labour Organization, September 20, 2018), http://www.ilo.org/global/publications/books/WCMS_645337/lang--en/index.htm.

179. Sarah T. Roberts, "Commercial Content Moderation: Digital Laborers' Dirty Work," in *The Intersectional Internet: Race, Sex, Class and Culture Online*, ed. Safiya Umoja Noble and Brendesha M. Tynes (New York: Peter Lang, 2016), 147–159.

180. M. S. Silberman et al., "Responsible Research with Crowds: Pay Crowdworkers at Least Minimum Wage," *Communications of the ACM* 61, no. 3 (February 2018): 39–41, https://doi.org/10.1145/3180492.

181. Arlene Kaplan Daniels, "Invisible Work," *Social Problems* 34, no. 5 (1987): 403–15, https://doi.org/10.2307/800538; Arlie Russell Hochschild, *The Managed Heart: Commercialization of Human Feeling* (Berkeley: University of California Press, 2012).

182. Astra Taylor, "The Automation Charade," *Logic Magazine*, October 2, 2018, https://logicmag.io/05-the-automation-charade/.

183. Jana Kasperkevic, "Ex-McDonald's CEO Suggests Replacing Employees with Robots amid Protests," *The Guardian*, May 25, 2016, https://www.theguardian.com/us-news/2016/may/25/former-mcdonalds-ceo-threatens-replace-employees-robots; Ed Rensi, "Thanks To 'Fight For $15' Minimum Wage, McDonald's Unveils Job-Replacing Self-Service Kiosks Nationwide," *Forbes*, November 29, 2016, https://www.forbes.com/sites/realspin/2016/11/29/thanks-to-fight-for-15-minimum-wage-mcdonalds-unveils-job-replacing-self-service-kiosks-nationwide/#5defa3eb4fbc.

184. Uliana Pavlova, "McDonald's Kiosks Mean More Staff at Chicago Flagship, Not Fewer," *Bloomberg*, August 8, 2018, https://www.bloomberg.com/news/articles/2018-08-08/mcdonald-s-kiosks-mean-more-staff-at-chicago-flagship-not-fewer.

185. Alex Rosenblat, *Uberland: How Algorithms Are Rewriting the Rules of Work* (Berkeley: University of California Press, 2018).

186. Ben Chapman, "Uber Eats and Deliveroo Riders Are Going on Strike This Week," *The Independent*, October 3, 2018, https://www.independent.co.uk/news/business/news/uber-eats-deliveroo-strike-mcdonalds-wetherspoons-mcstrike-industrial-action-a8567286.html; Carolyn Said, "Uber, Lyft Drivers Fear Getting Booted from Work," *San Francisco Chronicle*, October 14, 2018, https://www.sfchronicle.com/business/article/Uber-Lyft-drivers-fear-getting-booted-from-work-13304052.php.

187. Kate Conger, "Google and Facebook's Security Guards Are Fighting to Earn a Living Wage," *Gizmodo*, July 27, 2018, https://gizmodo.com/google-and-facebooks-security-guards-are-fighting-to-ea-1826104897.

188. Mark Bergen and Josh Eidelson, "Inside Google's Shadow Workforce," *Bloomberg*, July 25, 2018, https://www.bloomberg.com/news/articles/2018-07-25/inside-google-s-shadow-workforce.

189. Thuy Ong, "Amazon Patents Wristbands That Track Warehouse Employees' Hands in Real Time," *The Verge*, February 1, 2018, https://www.theverge.com/2018/2/1/16958918/amazon-patents-trackable-wristband-warehouse-employees; Hayley Peterson, "Missing Wages, Grueling Shifts, and Bottles of Urine: The Disturbing Accounts of Amazon Delivery Drivers May Reveal the True Human Cost of 'Free' Shipping," *Business Insider*, September 11, 2018, https://www.businessinsider.com/amazon-delivery-drivers-reveal-claims-of-disturbing-work-conditions-2018-8.

190. Catie Keck, "Amazon Workers Across Europe Protest Black Friday, Citing Grueling Work Conditions," *Gizmodo*, November 23, 2018, https://gizmodo.com/amazon-workers-across-europe-protest-black-friday-citi-1830622250.

191. "AI Now 2017 Report."

192. L. Rafael Reif, "Letter to the MIT Community Regarding the MIT Stephen A. Schwarzman College of Computing," *MIT News*, October 15, 2018, https://news.mit.edu/2018/letter-mit-community-regarding-mit-stephen-schwarzman-college-computing; Steve Lohr, "M.I.T. Plans College for Artificial Intelligence, Backed by $1 Billion," *The New York Times*, October 16, 2018, https://www.nytimes.com/2018/10/15/technology/mit-college-artificial-intelligence.html.

193. Lev Manovich, "Can We Think Without Categories?" *Digital Culture & Society,* 4, no. 1 (2018): 17-28.

194. AI Now Institute, "Gender, Race and Power: Outlining a New AI Research Agenda," *Medium,* November 15, 2018. https://medium.com/@AINowInstitute/gender-race-and-power-5da81dc14b1b.

195. Kristian Lum, "Statistics, we have a problem," *Medium,* December 13, 2017, https://medium.com/@kristianlum/statistics-we-have-a-problem-304638dc5de5.

196. Jennings Brown, "'NIPS' AI Conference Changes Name Following Protests Over Gross Acronym," *Gizmodo,* November 19, 2018. https://gizmodo.com/nips-ai-conference-changes-name-following-protests-ov-1830548185; Corinna Cortes et al., "From the Board: Changing Our Acronym," *NeurIPS* November 16, 2018, https://nips.cc/Conferences/2018/News.

197. "NIPS Name Change," *NeurIPS*, October 17, 2018, https://nips.cc/Conferences/2018/News.

198. Meredith Whittaker, one of the co-founders of AI Now, was one of the organizers of the Google Walkout. See also: Caroline O'Donovan and Ryan Mac, "Google Engineers Are Organizing A Walkout To Protest The Company's Protection Of An Alleged Sexual Harasser", *Buzzfeed,* Oct. 30, 2018, https://www.buzzfeednews.com/article/carolineodonovan/googles-female-engineers-walkout-sexual-harassment.

199. Davey Alba and Caroline O'Donovan, "Square, Airbnb, And eBay Just Said They Would End Forced Arbitration For Sexual Harassment Claims," *Buzzfeed,* Nov. 15, 2018, https://www.buzzfeednews.com/article/daveyalba/tech-companies-end-forced-arbitration-airbnb-ebay.

200. Sharon Florentine, "Alphabet Dismisses Action on Diversity and Inclusion," *CIO*, June 15, 2018, https://www.cio.com/article/3281867/it-industry/alphabet-dismisses-action-on-diversity-and-inclusion.html; Sara Ashley O'Brien, "Apple's Board Calls Diversity Proposal 'Unduly Burdensome and Not Necessary,'" *CNNMoney*, January 15, 2016, https://money.cnn.com/2016/01/15/technology/apple-diversity/index.html.

201. Tom Simonite, "AI Is the Future - But Where Are the Women?," *Wired*, August 17, 2018, https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/.

202. Jackie Snow, "'We're in a Diversity Crisis:' Cofounder of Black in AI on What's Poisoning Algorithms in Our Lives," *MIT Technology Review*, February 14, 2018, https://www.technologyreview.com/s/610192/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/.

203. Paula A. Johnson, Sheila E. Widnall, and Frazier F. Benya, eds., *Sexual Harassment of Women: Climate, Culture, and Consequences in Academic Sciences, Engineering, and Medicine* (Washington, DC: The National Academies Press, 2018), https://www.nap.edu/catalog/24994/sexual-harassment-of-women-climate-culture-and-consequences-in-academic.

204. Caroline Clark Hayes, "Computer Science: The Incredible Shrinking Woman," in *Gender Codes: Why Women Are Leaving Computing*, ed. Thomas J Misa (Hoboken: Wiley, 2010), 25–50.

205. Dale Strok, "Women in AI," *IEEE Expert: Intelligent Systems and Their Applications* 7, no. 4 (August 1992): 7–22, https://doi.org/10.1109/64.153460.

206. "Barriers to Equality in Academia: Women in Computer Science at M.I.T." (Cambridge, MA: The Laboratory of Computer Science and the Artificial Intelligence Laboratory at M.I.T., February 1983), https://homes.cs.washington.edu/~lazowska/mit/Images/title.pdf.

207. Langdon Winner, "Do Artifacts Have Politics?," *Daedalus* 109, no. 1 (1980): 121–36.

208. Stephen Merity, "Bias is not just in our datasets, it's in our conferences and community," *Smerity.com*, December 11, 2017, https://smerity.com/articles/2017/bias_not_just_in_datasets.html.

209. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women."

210. Safiya U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism,* (New York: NYU Press, 2018);

211. Clemens Apprich, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl, *Pattern Discrimination* (Minneapolis: University of Minnesota Press, forthcoming 2019).

212. "Litigating Algorithms."

213. Reisman et al., "Algorithmic Impact Assessments."

214. "CPSR History," *Computer Professionals for Social Responsibility*, June 1, 2005, http://cpsr.org/about/history/.

215. "The Never Again Pledge," accessed November 29, 2018, http://neveragain.tech/

216. Scott Shane and Daisuke Wakabayashi, "'The Business of War:' Google Employees Protest Work for the Pentagon," *The New York Times*, November 2, 2018, https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html.

217. Lucy Suchman, Lilly Irani, and Peter Asaro, "Google's March to the Business of War Must Be Stopped," *The Guardian*, May 16, 2018, https://www.theguardian.com/commentisfree/2018/may/16/google-business-war-project-maven; Mary Wareham, "Letter to Sergey Brin and Sundar Pichai," March 13, 2018, https://www.stopkillerrobots.org/wp-content/uploads/2018/04/KRC_LtrGoogle_12March2018.pdf; Daisuke Wakabayashi and Scott Shane, "Google Will Not Renew Pentagon Contract That Upset Employees," *The New York Times*, November 2, 2018, https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html.

218. Shaban, "Amazon Employees Demand Company Cut Ties with ICE;" "Who's Behind Ice?;" Caroline O'Donovan, "Employees Of Another Major Tech Company Are Petitioning Government Contracts," *BuzzFeed News*, June 26, 2018, https://www.buzzfeednews.com/article/carolineodonovan/salesforce-employees-push-back-against-company-contract; Sheera Frenkel, "Microsoft Employees Question C.E.O. Over Company's Contract With ICE," *The New York Times*, July 27, 2018, https://www.nytimes.com/2018/07/26/technology/microsoft-ice-immigration.html; Peter Kotecki, "Burning Man Protesters Raise Awareness of Palantir, Amazon ICE Ties," *Business Insider*, August 31, 2018, https://www.businessinsider.com/burning-man-protestors-palantir-amazon-ice-2018-8.

219. Greg Sandoval, "Over 100 Amazon Employees Sign Letter Asking Jeff Bezos to Stop Selling Facial-Recognition Software to Police," *Business Insider*, June 22, 2018, https://www.businessinsider.com/over-100-amazon-employees-sign-letter-jeff-bezos-stop-selling-facial-recognition-software-police-2018-6; Kade Crockford, "Over 150,000 People Tell Amazon: Stop Selling Facial Recognition Tech to Police," *American Civil Liberties Union*, June 18, 2018, https://www.aclu.org/blog/privacy-technology/surveillance-technologies/over-150000-people-tell-amazon-stop-selling-facial; Matt Cagle and Nicole Ozer, "Amazon Teams Up With Government to Deploy Dangerous New Facial Recognition Technology," *ACLU Free Future*, May 22, 2018, https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazon-teams-government-deploy-dangerous-new.

220. Ryan Gallagher "Google China Prototype Links Searches to Phone Numbers" *The Intercept*, September 14, 2018, https://theintercept.com/2018/09/14/google-china-prototype-links-searches-to-phone-numbers/.

221. Bryan Menegus, "Here's the Letter 1,400 Google Workers Sent Leadership in Protest of Censored Search Engine for China," *Gizmodo*, August 16, 2018, https://gizmodo.com/heres-the-letter-1-400-google-workers-sent-leadership-i-1828393599; Google Employees Against Dragonfly, "We Are Google Employees. Google Must Drop Dragonfly.," *Medium*, November 27, 2018, https://medium.com/@googlersagainstdragonfly/we-are-google-employees-google-must-drop-dragonfly-4c8a30c5e5eb; "Google Must Not Capitulate to China's Censorship Demands," *Amnesty International*, November 28, 2018, https://www.amnesty.org/en/latest/news/2018/11/google-must-not-capitulate-to-chinas-censorship-demands/.

222. Co-founder of AI Now, Meredith Whittaker, was one of the eight core organizers of the Google Walkout.

223. Google Walkout for Real Change, "Google Employees and Contractors Participate in 'Global Walkout for Real Change,'" *Medium*, November 2, 2018, https://medium.com/@GoogleWalkout/google-employees-and-contractors-participate-in-global-walkout-for-real-change-389c65517843.

224. Richard Waters, "Google Ends Forced Arbitration for Sexual Harassment Claims," *Financial Times*, November 8, 2018, https://www.ft.com/content/ce3c11ec-e37e-11e8-a6e5-792428919cee.

225. Kate Gibson, "Tech Signals End of Forced Arbitration for Sexual Misconduct Claims," *CBS MoneyWatch*, November 16, 2018, https://www.cbsnews.com/news/tech-signals-end-of-forced-arbitration-for-sexual-misconduct-claims/.

226. "AI Now 2017 Report."

227. Google Walkout for Real Change, "#GoogleWalkout Update: Collective Action Works, but We Need to Keep Working.," *Medium*, November 8, 2018, https://medium.com/@GoogleWalkout/googlewalkout-update-collective-action-works-but-we-need-to-keep-working-b17f673ad513; Noam Scheiber, "Google Workers Reject Silicon Valley Individualism in Walkout," *The New York Times*, November 7, 2018, https://www.nytimes.com/2018/11/06/business/google-employee-walkout-labor.html.

228. For example, Google's highly secret Dragonfly project to censor search results in China and link Chinese residents' phone numbers to search logs. See: "We Are Google Employees. Google Must Drop Dragonfly.," *Medium*, November 27, 2018, https://medium.com/@googlersagainstdragonfly/we-are-google-employees-google-must-drop-dragonfly-4c8a30c5e5eb.

229. Cade Metz, "Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent," *The New York Times*, October 22, 2017, https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html.

230. "Recommendations" in "AI Now 2016 Report," (New York: AI Now Institute, 2016), https://ainowinstitute.org/AI_Now_2016_Report.pdf.

231. Crawford and Joler, "Anatomy of an AI System," https://anatomyof.ai.

232. See, as just one of many examples: Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

233. Lecher, "A Healthcare Algorithm Started Cutting Care, and No One Knew Why."

234. Sonia Katyal, "Private Accountability in the Age of the Algorithm," *UCLA Law Review* 66 (forthcoming 2019), https://www.uclalawreview.org/private-accountability-age-algorithm/.