AI Now 2017 Report

Authors

Alex Campolo, New York University
Madelyn Sanfilippo, New York University
Meredith Whittaker, Google Open Research, New York University, and AI Now
Kate Crawford, Microsoft Research, New York University, and AI Now

Editors

Andrew Selbst, Yale Information Society Project and Data & Society
Solon Barocas, Cornell University

Table of Contents

Recommendations

These recommendations reflect the views and research of the AI Now Institute at New York University. We thank the experts who contributed to the *AI Now 2017 Symposium and Workshop* for informing these perspectives, and our research team for helping shape the *AI Now 2017 Report.*

1. **Core public agencies, such as those responsible for criminal justice, healthcare, welfare, and education (e.g "high stakes" domains) should no longer use "black box" AI and algorithmic systems.** This includes the unreviewed or unvalidated use of pre-trained models, AI systems licensed from third party vendors, and algorithmic processes created in-house. The use of such systems by public agencies raises serious due process concerns, and at a minimum they should be available for public auditing, testing, and review, and subject to accountability standards.

2. **Before releasing an AI system, companies should run rigorous pre-release trials to ensure that they will not amplify biases and errors due to any issues with the training data, algorithms, or other elements of system design.** As this is a rapidly changing field, the methods and assumptions by which such testing is conducted, along with the results, should be openly documented and publicly available, with clear versioning to accommodate updates and new findings.

3. **After releasing an AI system, companies should continue to monitor its use across different contexts and communities.** The methods and outcomes of monitoring should be defined through open, academically rigorous processes, and should be accountable to the public. Particularly in high stakes decision-making contexts, the views and experiences of traditionally marginalized communities should be prioritized.

4. **More research and policy making is needed on the use of AI systems in workplace management and monitoring, including hiring and HR.** This research will complement the existing focus on worker replacement via automation. Specific attention should be given to the potential impact on labor rights and practices, and should focus especially on the potential for behavioral manipulation and the unintended reinforcement of bias in hiring and promotion.

5. **Develop standards to track the provenance, development, and use of training datasets throughout their life cycle.** This is necessary to better understand and monitor issues of bias and representational skews. In addition to developing better records for how a training dataset was created and maintained, social scientists and measurement researchers within the AI bias research field should continue to examine existing training datasets, and work to understand potential blind spots and biases that may already be at work.

6. **Expand AI bias research and mitigation strategies beyond a narrowly technical approach.** Bias issues are long term and structural, and contending with them necessitates deep interdisciplinary research. Technical approaches that look for a one-time "fix" for fairness risk oversimplifying the complexity of social systems. Within each domain – such as education, healthcare or criminal justice – legacies of bias and movements toward equality have their own histories and practices. Legacies of bias cannot be "solved" without drawing on domain expertise. Addressing fairness meaningfully will require interdisciplinary collaboration and methods of listening across different disciplines.

7. **Strong standards for auditing and understanding the use of AI systems "in the wild" are urgently needed.** Creating such standards will require the perspectives of diverse disciplines and coalitions. The process by which such standards are developed should be publicly accountable, academically rigorous and subject to periodic review and revision.

8. **Companies, universities, conferences and other stakeholders in the AI field should release data on the participation of women, minorities and other marginalized groups within AI research and development.** Many now recognize that the current lack of diversity in AI is a serious issue, yet there is insufficiently granular data on the scope of the problem, which is needed to measure progress. Beyond this, we need a deeper assessment of workplace cultures in the technology industry, which requires going beyond simply hiring more women and minorities, toward building more genuinely inclusive workplaces.

9. **The AI industry should hire experts from disciplines beyond computer science and engineering and ensure they have decision making power.** As AI moves into diverse social and institutional domains, influencing increasingly high stakes decisions, efforts must be made to integrate social scientists, legal scholars, and others with domain expertise that can guide the creation and integration of AI into long-standing systems with established practices and norms.

10. **Ethical codes meant to steer the AI field should be accompanied by strong oversight and accountability mechanisms.** More work is needed on how to substantively connect high level ethical principles and guidelines for best practices to everyday development processes, promotion and product release cycles.

Executive Summary

Artificial intelligence (AI) technologies are in a phase of rapid development, and are being adopted widely. While the concept of artificial intelligence has existed for over sixty years, real-world applications have only accelerated in the last decade due to three concurrent developments: better algorithms, increases in networked computing power and the tech industry's ability to capture and store massive amounts of data.

AI systems are already integrated in everyday technologies like smartphones and personal assistants, making predictions and determinations that help personalize experiences and advertise products. Beyond the familiar, these systems are also being introduced in critical areas like law, finance, policing and the workplace, where they are increasingly used to predict everything from our taste in music to our likelihood of committing a crime to our fitness for a job or an educational opportunity.

AI companies promise that the technologies they create can automate the toil of repetitive work, identify subtle behavioral patterns and much more. However, the analysis and understanding of artificial intelligence should not be limited to its technical capabilities. The design and implementation of this next generation of computational tools presents deep normative and ethical challenges for our existing social, economic and political relationships and institutions, and these changes are already underway. Simply put, AI does not exist in a vacuum. We must also ask how broader phenomena like widening inequality, an intensification of concentrated geopolitical power and populist political movements will shape and be shaped by the development and application of AI technologies.

Building on the inaugural 2016 report, *The AI Now 2017 Report* addresses the most recent scholarly literature in order to raise critical social questions that will shape our present and near future. A year is a long time in AI research, and this report focuses on new developments in four areas: labor and automation, bias and inclusion, rights and liberties, and ethics and governance. We identify emerging challenges in each of these areas and make recommendations to ensure that the benefits of AI will be shared broadly, and that risks can be identified and mitigated.

Labor and automation: Popular media narratives have emphasized the prospect of mass job loss due to automation and the widescale adoption of robots. Such serious scenarios deserve sustained empirical attention, but some of the best recent work on AI and labor has focused instead on specific sectors and tasks. While few jobs will be completely automated in the near term, researchers estimate that about a third of workplace tasks can be automated for the majority of workers. New policies such as the Universal Basic Income (UBI) are being designed to address concerns about job loss, but these need much more study.

An underexplored area that needs urgent attention is how AI and related algorithmic systems are already changing the balance of workplace power. Machine learning techniques are quickly being integrated into management and hiring

decisions, including in the so-called gig economy where technical systems match workers with jobs, but also across more traditional white collar industries. New systems make promises of flexibility and efficiency, but they also intensify the surveillance of workers, who often do not know when and how they are being tracked and evaluated, or why they are hired or fired. Furthermore, AI-assisted forms of management may replace more democratic forms of bargaining between workers and employers, increasing owner power under the guise of technical neutrality.

Bias and inclusion: One of the most active areas of critical AI research in the past year has been the study of bias, both in its more formal statistical sense and in the wider legal and normative senses. At their best, AI systems can be used to augment human judgement and reduce both our conscious and unconscious biases. However, training data, algorithms, and other design choices that shape AI systems may reflect and amplify existing cultural assumptions and inequalities. For example, natural language processing techniques trained on a corpus of internet writing from the 1990s may reflect stereotypical and dated word associations—the word "female" might be associated with "receptionist." If these models are used to make educational or hiring decisions, they may reinforce existing inequalities, regardless of the intentions or even knowledge of system's designers.

Those researching, designing and developing AI systems tend to be male, highly educated and very well paid. Yet their systems are working to predict and understand the behaviors and preferences of diverse populations with very different life experiences. More diversity within the fields building these systems will help ensure that they reflect a broader variety of viewpoints.

Rights and liberties: The application of AI systems in public and civil institutions is challenging existing political arrangements, especially in a global political context shaped by events such as the election of Donald Trump in the United States. A number of governmental agencies are already partnering with private corporations to deploy AI systems in ways that challenge civil rights and liberties. For example, police body camera footage is being used to train machine vision algorithms for law enforcement, raising privacy and accountability concerns. AI technologies are also being deployed in the very legal institutions designed to safeguard our rights and liberties, with proprietary risk assessment algorithms already being used to help judges make sentencing and bail decisions, potentially amplifying and naturalizing longstanding biases, and rendering them more opaque to oversight and scrutiny.

Privacy rights represent a particularly sensitive challenge for current AI applications, especially in domains like healthcare, where AI is being used to help make diagnoses. For AI to deliver on its promises, it requires large amounts of data, which likely means an increase in data collection, both its scale and granularity. Without contextual knowledge, informed consent, and due processes mechanisms, these systems can create risks that threaten and expose already vulnerable populations.

Ethics and governance: The areas of ethics and governance attempt to address many of the challenges and opportunities identified above. We track the growing

interest in ethical codes of conduct and principles, while noting that these need to be tied more closely to everyday AI design and development. The military use of artificial intelligence takes on a special urgency in the case of lethal autonomous weapons systems.

There are multiple signs of progress in the development of professional and legal ethical codes to govern the design and application of AI technologies. However, in the face of rapid, distributed, and often proprietary AI development and implementation, such forms of soft governance face real challenges. Among these are problems of coordination among different ethical codes, as well as questions around enforcement mechanisms that would go beyond voluntary cooperation by individuals working in research and industry. New ethical frameworks for AI need to move beyond individual responsibility to hold powerful industrial, governmental and military interests accountable as they design and employ AI.

The following report develops these themes in detail, and reflects on the latest academic research. AI is already with us, and we are now faced with important choices on how it will be designed and applied. Most promisingly, the approaches described in this report demonstrate that there is growing interest in developing AI that is attuned to underlying issues of fairness and equality.

Introduction

In July of 2016, Kate Crawford and Meredith Whittaker co-chaired the first *AI Now Symposium* in collaboration with the Obama White House's Office of Science and Technology Policy and the National Economic Council. The event brought together experts and members of the public to discuss the near-term social and economic impacts of artificial intelligence (AI).[1] AI systems are already being integrated in social, political and economic domains, and the implications can be complex and unpredictable. The now-annual *AI Now Symposium* focuses on AI's core social implications, bringing together leading experts from across sectors and disciplines with the aim of better understanding how AI systems are already working in the world.

The *AI Now 2016 Symposium* identified instances where AI challenged current thinking about professional responsibilities, decision-making and accountability. Following this, *The AI Now 2016 Report* reflected expert discussion and provided recommendations for future research and policy interventions.[2]

The *AI Now 2017 Symposium* deepened this examination of the near-term social and economic implications of AI, and the accompanying report provides an overview of the key issues that the *2017 Symposium* addressed. These are: 1) *Labor and Automation*, 2) *Bias and Inclusion*, *3) Rights and Liberties* and 4) *Ethics and Governance*. In selecting these four themes, we are building on the 2016 report[3] and introducing new areas of concern, with close attention to developments that have occurred in the last 12 months.

The first section on *Labor and Automation* considers the need for a more granular, skills-based, and sectoral approach to understanding AI and automation's impacts on labor practices. While big questions about what implications automation and AI have for labor overall are still wide open, there are also important questions about the distinct roles that automation and AI will play within specific industries, sectors and tasks - particularly how it will be used as a tool of employee hiring, firing and management. The second section focuses on *Bias and Inclusion*, a growing concern among those looking at the design and social implications of AI decision-making systems. Here, we address the problem of diversity and inclusion within the AI industry itself. We also share new technical advances

---

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. Stuart J. Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, Englewood Cliffs, NJ: Prentice Hall, 1995: 27

2 AI Now, "The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term," (2016) https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3_RpmwKHu.pdf.

3 Ibid.

that help to better understand and mitigate biases that AI systems may perpetuate and even amplify due to biased training data, faulty algorithms or other factors. The third section, on *Rights and Liberties*, begins by recognizing the recent rise of political authoritarianism, and asks about the role of AI systems in either supporting or eroding citizens' rights and liberties in areas like criminal justice, law enforcement, housing, hiring, lending and other domains. The last section, on *Ethics and Governance*, connects AI as we see it today with the history of AI research and development. It also looks at *whose* concerns are ultimately reflected in the ethics of AI, and how ethical codes and other strategies could be developed in a time of political volatility.

We are in the early stages of a long-term discussion, and accordingly, there are as many new questions as there are answers to the old ones. We hope this report provides a productive grounding in the extraordinary challenges and opportunities of the current moment, and helps spur research and inquiry into the social and economic implications of the turn to AI. .

Labor and Automation

The editors of *Nature* have argued that we need to match technical AI research funding with "solid, well-funded research to anticipate the scenarios [AI] could bring about, and to study possible political and economic reforms that will allow those usurped by machinery to contribute to society."[4] The *AI Now Labor Primer* described how forms of automation based on machine learning and robotics have the potential to both increase the productivity of labor and to exacerbate existing inequalities in the distribution of wealth.[5] In an economic context characterized by both low productivity growth and historically high levels of inequality, it will be important to find ways to use AI to promote equality and shared prosperity.[6]

While there is still considerable attention focused on large, structural changes in labor markets and on the economy as a whole, new research has been focusing on specific industries and the impact of AI systems on particular tasks within a profession. This section describes new developments in AI's application within various labor sectors, and suggests directions that research could productively explore in the future.

Research by Sector and Task

At the beginning of 2017, the McKinsey Global Institute (MGI) released a report looking at specific workplace *tasks* and whether they were more or less susceptible to automation, specifically those involving "predictable physical" activities and those involving data

---

4 "Anticipating Artificial Intelligence," *Nature* 532, no. 7600 (April 28, 2016): 413, doi:10.1038/532413a.

5 "Labor and AI" (New York, NY: *AI Now*, July 7, 2016), https://artificialintelligencenow.com/media/documents/AI_NOW_LABOR_PRIMER.pdf.

6 Jason Furman, "Is This Time Different? The Opportunities and Challenges of Artificial Intelligence," expanded remarks from the *AI Now* expert workshop, July 7, 2016, New York University, https://obamawhitehouse.archives.gov/sites/default/files/page/files/20160707_cea_ai_furman.pdf.

collection or processing. While relatively few current jobs can be totally automated with today's technology, MGI estimates that 60 percent of all occupations have the potential for about a third of their activities to be automated.[7] In a similar vein, analysts in Deloitte's Human Capital division predict a future where human skills will be "augmented" through "collaboration" with machines capable of performing routine tasks.[8]

To prepare for these changes, it will be essential that policymakers have access to robust data on how advances in machine learning, robotics and the automation of perceptual tasks are changing the nature and organization of work, and how these changes manifest across different roles and different sectors. This data will be necessary for any robust policy proposal. However, a recent report from the National Academies of Sciences, Engineering, and Medicine identifies a lack of such data, finding existing federal statistical data limited in its capacity to answer these questions. The report recommends new multidisciplinary and qualitative research methods to capture present and future transformations in work.[9]

A series of economic studies have begun to investigate the effects of robots on labor markets from an empirical perspective. A 2015 paper by George Graetz and Guy Michaels used new data from the International Federation of Robots to estimate changes in productivity and employment due to robot adoption, finding increases in productivity and slightly lowered working hours for low and middle-skilled workers.[10] Using the same data, Daron Acemoglu and Pascual Restrepo analyzed developments in labor markets across the United States from 1990 to 2007. They estimated that the number of jobs lost due to robots during this period ranged from 360,000 to 670,000, and that this trend could accelerate with a more intensive adoption of automation across sectors.[11] Model assumptions play an important role in these empirical analyses[12] and will need to be continually tested against employment data. To this end, Management Professor and former Senior Economist at the White House Council of Economic Advisers Robert Seamans argues that even more fine-grained, company-level data will be necessary to understand whether AI and automation systems are replacing or complementing human workers.[13]

---

7 Ibid., 5-6.

8 Jeff Schwartz, Laurence Collins, Heather Stockton, Darryl Wagner and Brett Walsh, "The Future of Work: The Augmented Workforce," (Deloitte Human Capital, February 28, 2017), https://dupress.deloitte.com/dup-us-en/focus/human-capital-trends/2017/future-workforce-changing-nature-of-work.html

9 National Academies of Sciences, Engineering and Medicine, "Information Technology and the U.S. Workforce: Where Are We and Where Do We Go from Here?," (Washington, DC: The National Academies Press, 2017), https://www.nap.edu/read/24649/.

10 Georg Graetz and Guy Michaels, "Robots at Work," IZA Discussion Paper (Institute for the Study of Labor (IZA), March 2015), http://econpapers.repec.org/paper/izaizadps/dp8938.htm.

11 Daron Acemoglu and Pascual Restrepo, "Robots and Jobs: Evidence from US Labor Markets," Working Paper (Cambridge MA: National Bureau of Economic Research, March 2017), doi:10.3386/w23285.

12 For instance, economists at the Economic Policy Institute argue that Restrepo and Acemoglu's estimates of unemployment were localized and that the media distorted their conclusions regarding job loss while also ignoring productivity increases. See: Lawrence Mishel and Bivens, "The Zombie Robot Argument Lurches on: There Is No Evidence That Automation Leads to Joblessness or Inequality" (Washington, DC: Economic Policy Institute, May 24, 2017), http://www.epi.org/publication/the-zombie-robot-argument-lurches-on-there-is-no-evidence-that-automation-leads-to-joblessness-or-inequality/.

13 Robert Seamans, "We Won't Even Know If A Robot Takes Your Job," *Forbes*, January 11, 2017, http://www.forbes.com/sites/washingtonbytes/2017/01/11/we-wont-even-know-if-a-robot-takes-your-job/.

AI and the Nature of Work

While the displacement of entire occupations, such as taxi or truck drivers,[14] is clearly an important concern, AI is also transforming a wide range of occupations and roles. Across sectors, automated management and hiring technologies are being introduced, promising to increase worker productivity and flexibility, but also exposing workers to new forms of monitoring, manipulation and control. This changes labor processes and power relations. Further research on this topic is needed to address how AI is transforming the nature of work itself, and how these transformations are manifesting for specific occupations within specific sectors.

Luke Stark and Alex Rosenblat's research with Uber drivers suggests one model for this approach. By listening to drivers, they identified algorithmic forms of management used by the company.[15] While its driver platform, which acts as a kind of remote management console, helps make more efficient use of driver time in this digital "matching market,"[16] the platform also exposes fundamental informational asymmetries between worker and platform owner. For example, drivers have about 15 seconds to accept ride requests via the platform, and are not shown the rider's destination. With drivers in the dark, they don't know when they will accept short, unprofitable fares. Meanwhile, Uber furthers its own goal of providing near-instantaneous service to all prospective riders.[17] Because Uber designs the platform and can change it at will, conflicts of interest between worker and platform owner are systematically settled in favor of Uber via the platform itself, not collective bargaining or other processes that allow for worker participation. This flatly contradicts any argument that the platform is "neutral." It will be interesting to see what comes of the recent New York administrative law judge's ruling, which classified Uber drivers as "employees" under New York law, contrary to Uber's claims otherwise.[18]

Of course, asymmetrical forms of workplace management and control long predate AI.[19] The task for researchers is to determine specifically what makes AI-powered asymmetries different from other forms of monitoring, such as Taylorist scientific management[20] and the audit culture of total quality control.[21] One clear difference is AI's reliance on workplace surveillance and the data it produces, and thus the normalization of workplace surveillance

---

14 Truckers, like ride-sharing drivers, are also subject to data-driven forms of surveillance and control. e.g. Karen E. C. Levy, "The Contexts of Control: Information, Power, and Truck-Driving Work," *The Information Society* 31, No. 2 (March 15, 2015): 160–74, doi:10.1080/01972243.2015.998105.

15 Alex Rosenblat and Luke Stark, "Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers," International Journal of Communication 10 (July 27, 2016): 3758-3784, http://ijoc.org/index.php/ijoc/article/view/4892/1739.

16 Eduardo M. Azevedo and E. Glen Weyl, "Matching Markets in the Digital Age," Science 352, no. 6289 (May 27, 2016): 1056–57, http://science.sciencemag.org/content/352/6289/1056.

17 Rosenblat and Stark, "Algorithmic Labor and Information Asymmetries," 3762.

18 Dana Rubenstein, "State Labor Judge Finds Uber an 'employer'," *Politico*, May 13, 2017, http://www.politico.com/states/new-york/albany/story/2017/06/13/state-labor-court-finds-uber-an-employer-112733.

19 Ifeoma Ajunwa, Kate Crawford and Jason Schultz, "Limitless Worker Surveillance,"*California Law Review* 105, No. 3 , 2017.

20 Hugh G. J Aitken, *Taylorism at Watertown Arsenal; Scientific Management in Action*, 1908-1915. (Cambridge: Harvard University Press, 1960).

21 Marilyn Strathern, *Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy* (London: Routledge, 2000).

practices. Such systems provide employers with expansive and often invasive data about the workplace behaviors of their employees. It is this data that AI-powered management systems rely on to generate insights. As AI-driven management becomes more common, so will the data collection and worker surveillance practices on which it relies. Worryingly, this employee monitoring is not necessarily limited to the workplace, and can spill into private life, such as with fitness trackers, ubiquitous productivity apps, or company-issued smartphones equipped with monitoring features.

While we might assume this would be held in check by privacy laws and existing policy, Ifeoma Ajunwa, Kate Crawford and Jason Schultz published a study of existing legal frameworks, assessing if there are any meaningful limits on workplace surveillance. They found very few, some of which are already under threat from the Trump administration.[22] This degree of 24/7 surveillance has the potential to transform key features of prior management systems, potentially in ways workers won't be aware of or have a say in. Employers could easily use machine learning techniques to identify behavioral patterns both during and outside of work hours, and then exploit these one-sided insights to increase profits and manipulate behaviors, with potentially negative effects for workers.

Uber's platform demonstrates how workers are directly and indirectly manipulated in service of instant customer gratification. The company wants to keep up the number of available cars, even during times of low demand when drivers make less money. To address this, the ride-sharing company drew on behavioral economic research about the psychological tendency of taxi workers to set round earnings goals and stop working when they reach them.[23] Uber, with access to vast real-time data about driver activities, can quickly test such theories, using machine learning to identify exploitable behavioral patterns, even at an individual level. Uber discovered that drivers quickly abandon mental income targets in favor of working at times of high demand. To combat this tendency, Uber sent tailored nudge messages[24] to drivers indicating when they are close to revenue target during times when it was advantageous for Uber to keep its drivers on the road.[25] Until a recent feature in *The New York Times*, drivers were unaware that they were subjects in a large behavioral experiment that sought to modify their actions to benefit the company's goals. Given the opacity of these systems, there may be many more such experiments that

---

22 Ifeoma Ajunwa, Kate Crawford and Jason Schultz, "Limitless Worker Surveillance,"*California Law Review* 105, No. 3 (June 1, 2017).

23 Colin Camerer, Linda Babcock, George Loewenstein and Richard Thaler, "Labor Supply of New York City Cab Drivers: One Day at a Time," The Quarterly Journal of Economics 112, No. 2 (May 1, 1997): 407–41, doi:10.1162/003355397555244.

24 The use of "nudge" as a more technical, policy-oriented term has emerged out of work in the decision and choice sciences, most influentially that of behavioral economist Richard Thaler and the legal scholar Cass Sunstein, who headed the Obama administration's Office of Information and Regulatory Affairs. They, in turn, draw on psychological studies of how people make decisions under conditions of uncertainty and avoid errors due to heuristics—like an earnings goal—and biases. These were first identified by the influential psychologists Amos Tversky and Daniel Kahneman. V.:Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (New York: Penguin Books, 2009); Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, No. 4157 (September 27, 1974): 1124–31, doi:10.1126/science.185.4157.1124.

25 Noam Scheiber, "How Uber Uses Psychological Tricks to Push Its Drivers' Buttons," *The New York Times*, April 2, 2017, https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html?_r=0.

workers and the public will never know about.

This case illustrates how AI management might differ from past forms of incentive-based control. As companies gather more data on their workers, they no longer need to rely on generalized psychological theories or human-to-human assessments of merit. They can instead exploit information asymmetries to identify behavioral patterns at the *individual level* and nudge people toward the most profitable activities for the platform owners, even when these operate against the best interests of workers themselves. By selectively exploiting workers' behavior, often without workers' consent or even knowledge, these technologies have the potential to make workers complicit in their own exploitation. To address these emerging imbalances of workplace power, it will likely be necessary for unions, labor rights advocates and individual workers to participate in the design of worker platforms. It will also likely be necessary to give workers a democratic voice in shaping both whether and how they are monitored and how machine learning techniques will be used to process such data. This is a rich area of research and design for the technical architects of AI management systems, labor organizers and advocates to explore.

AI management systems also provide new and invasive methods for evaluating employees and making retention decisions. For example, the employee monitoring firm Veriato captures information from nearly any task a worker performs on a computer, from browsing history to email and chat, even taking periodic screenshots of workers' monitor displays. The firm's software aggregates this information, then uses machine learning to detect anomalous behaviors. The program can then send warning messages to employees who deviate from the norm.[26] What the consequences of such deviance are for workers is up to the employer. And this isn't all. Veriato's software also offers features to score email and chats for sentiment using natural language processing. Language that their program determines to be "negative" is interpreted by the company as an indication of a productivity risk, or of an employee who is getting ready to leave the company. Similarly, another company, Workday, assigns employees individualized risk score based on 60 factors.[27] Many employees who use a work-issued computer or mobile are already subject to this type of monitoring and software-driven ranking and assessment. Additionally, many of them likely have no idea that their value as an employee is being determined in part by software systems scoring everything from the emotional content of their emails to their frequency of accepting meeting requests.

Beyond employee surveillance, the combination of customer surveillance and AI has the potential to turn previously stable employment in sectors like food service and retail into a form of gig work. So-called scheduling software has allowed retailers to switch from standard shifts to a more "on call" model, based on algorithmic predictions about whether customers will be in a store at a given time. While the use of such software can cut an employer's costs by reducing staff during off-peak customer hours, as Solon Barocas and

[26] Ted Greenwald, "How AI Is Transforming the Workplace," Wall Street Journal, March 10, 2017, sec. Business, https://www.wsj.com/articles/how-ai-is-transforming-the-workplace-1489371060.

[27] Ibid.

Karen Levy have observed, it is "highly destabilizing" for workers who never know ahead of time whether or not they will be called in for work.[28] The use of predictive scheduling software, whether by gig employers like Uber or more traditional employers, collapses work-life boundaries. It also puts workers at risk of over- or underwork, gives workers little to no control over shift times, and provides them with little ability to predict income flows or to plan ahead for things like child care or a second job. Recognizing the negative impacts that such precarious schedules can have on workers, the Oregon state Senate and House recently passed a bill mandating that large employers in retail, manufacturing and hospitality provide workers a written estimate of their schedule at least 7 days before the start of the work week.[29] Barring a veto from the state's Governor, Oregon will join New York, San Francisco and Seattle, who have also passed laws mandating predictable scheduling.

The increasing role of AI and automation within various labor sectors has the potential to revise our understanding of labor and our expectations of goods and services. As consumers grow accustomed to dealing with automated systems, there is a potential to ignore or devalue the human labor that remains essential in many instances. The *AI Now 2016 Labor Primer* emphasized that AI often demands "human caretakers"[30] — these vary, from workers who maintain and repair data centers to moderators who check the results of even the most sophisticated computer vision algorithms.[31] Since the *AI Now 2016 Labor Primer*, Facebook has announced the hiring of 3,000 workers to monitor its live video streaming services for violence, exploitation and hate speech.[32] This is both an acknowledgement that AI systems don't always do the work as intended, and an example of how essential human work happening behind the scenes of complex systems is often invisible. Not surprisingly, this work tends to be outsourced to countries where wages are very low. How will such maintenance and repair work be valued by consumers who have been led to believe that such services are entirely automated? How will companies that promote themselves as fully automated "AI magic" treat and recognize workers within these systems? Additionally, how will this lack of visibility impact workers' ability to organize and shape their own working conditions?

Managers too, will need to rethink how they formulate goals and use data, while acknowledging the limits and risks of automated systems. Michael Luca, Jon Kleinberg, and Sendhil Mullainathan argue that these systems can miss contextual details and may not

---

28 Solon Barocas and Karen Levy, "What Customer Data Collection Could Mean for Workers," *Harvard Business Review*, August 31, 2016, https://hbr.org/2016/08/the-unintended-consequence-of-customer-data-collection.

29 Hillary Borrud, "Oregon on way to become first state to guarantee predictable work schedules," Oregonian, June 29, 2017, sec. Oregon Live, http://www.oregonlive.com/politics/index.ssf/2017/06/oregon_on_way_to_become_first.html

30 "AI's human caretakers" in the *2016 AI Now Labor and Automation Primer*, "Labor and AI," https://artificialintelligencenow.com/media/documents/AI_NOW_ETHICS_PRIMER_T1yYKVR.pdf.

31 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), 147–60.

32 Kathleen Chaykowski, "Facebook Is Hiring 3,000 Moderators In Push To Curb Violent Videos," *Forbes*, accessed May 10, 2017, http://www.forbes.com/sites/kathleenchaykowski/2017/05/03/facebook-is-hiring-3000-moderators-in-push-to-curb-violent-videos/.

provide clear reasoning for decisions. They advise managers to ask employees and stakeholders to articulate concerns with such systems; more democratic input can often improve performance. Similarly they recommend that diverse data-inputs be used in pursuit of long-term goals and values, instead of focusing too narrowly on low-hanging fruit, which can often produce unintended consequences, like clickbait in search of social media engagement.[33]

Inequality and Redistribution

What happens to workers after their jobs have been automated? The potential for AI systems to exacerbate inequality has been widely acknowledged. To address what to do about it, some are turning to models of resource redistribution, and to the idea of a universal basic income (UBI). The past year has seen a number of high-profile experiments in redistributive social welfare, based on assumptions that AI and automation will require resource distribution not explicitly tied to the sale of individual labor. Some of the most visible efforts have come from governments and private actors running small trials where people receive direct cash transfers in the form of a basic income stipend. It bears noting that payments made as a part of these experiments cannot be considered "universal" insofar as they are provided to a limited number of people. Thus, while these experiments can gather informative data that tells us about individual reactions to the receipt of such funds, they cannot account for the society-wide impact of a universal payment. For example, in April of 2017, the government of Ontario began a UBI pilot research program with 4,000 participants that will provide up to C$16,989 per year for a single person and C$24,027 per year for a couple, less 50 percent of any earned income.[34] Y Combinator, a Silicon Valley-based startup incubator, began a one year UBI pilot study in Oakland in which one hundred families will receive $1,000 to $2,000 per month over the course of a year.[35] Y Combinator president (and OpenAI co-chairman) Sam Altman explicitly references job displacement due to technology as a motivating factor for UBI research.[36] While UBI remains a politically contentious idea with significant variations in approach and implementation, it is currently one of the most commonly proposed policy responses to AI-driven job losses, and as such deserves close assessment.

Bias and Inclusion

The word "bias" has multiple meanings that intersect with AI applications in ways that can overlap and occasionally contradict each other. This can add unnecessary confusion to what is a critically needed domain of research. In statistics—used in many machine learning

---

[33] Michael Luca, Jon Kleinberg, and Sendhil Mullainathan, "Algorithms Need Managers, Too," *Harvard Business Review*, January 1, 2016, https://hbr.org/2016/01/algorithms-need-managers-too.

[34] Ministry of Community and Social Services, "Ontario's Basic Income Pilot," *News.ontario.ca*, April 24, 2017, https://news.ontario.ca/mcss/en/2017/04/ontarios-basic-income-pilot.html.

[35] Michael J. Coren, "Y Combinator Is Running a Basic Income Experiment with 100 Oakland Families," *Quartz*, June 1, 2017, https://qz.com/696377/y-combinator-is-running-a-basic-income-experiment-with-100-oakland-families/.

[36] Sam Altman, "Moving Forward on Basic Income," *Y Combinator*, May 31, 2016, https://blog.ycombinator.com/moving-forward-on-basic-income/.

applications—"bias" has a specific meaning that differs from the popular and social scientific definitions. For example, the idea of "selection bias" refers to errors in estimation that result when some members of a population are more likely to be sampled than others. So when a machine learning program trained to recognize, say, faces of a particular racial group is applied to larger or more diverse populations, it may produce biased results in the sense of having a lower measure of accuracy.

The word "bias" also has normative meanings in both colloquial and legal language, where it refers to judgement based on preconceived notions or prejudices, as opposed to the impartial evaluation of facts. Impartiality is a core value of many legal systems and governs many legal processes, from juror selection to the limitations placed on judges. For example, in the United States the Sixth Amendment to the Constitution mandates a right to an impartial jury and the Fourteenth mandates equal protection under the law. This sense of the word bias is closely linked to normative and ethical perspectives on fairness, and the idea that different groups should be treated equally.

When examining technical systems, there can be a temptation to, or vested interest in, limiting discussion of bias to the first more 'neutral' statistical sense of the term. However, in practice there is rarely a clear demarcation between the statistical and the normative definitions: biased models or learning algorithms, as defined statistically, can lead to unequal and unfair treatments and outcomes for different social or racial groups.

The danger of bias increases when these systems are applied, often in non-transparent ways, to critical institutions like criminal justice and healthcare. The social sciences and critical humanities have decades of research on bias within social systems that have much to offer the current debate on bias in AI and algorithmic systems.[37] Since *AI Now* is deeply interested in the social and political implications of AI, this report will use the word "bias" in its broader, normative sense in the following section, while acknowledging its close relationship with statistical usages.

While the potential impact of such biases are extremely worrying, solutions are complicated. This is in part because biased AI can result from a number of factors, alone or in combination, such as who develops systems, what goals system developers have in mind during development, what training data they use, and whether the the systems work well for different parts of the population.[38] This section addresses the latest research on bias in AI and discusses some of the emerging strategies being used to address it.

Where Bias Comes From

AI systems are taught what they "know" from training data. Training data can be

[37] Barocas, Crawford, Shapiro and Wallach, "The Problem with Bias: Allocative versus Representational Harms in Machine Learning," SIGCIS conference, October 2017.

[38] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," California Law Review 104, No. 3 (June 1, 2016): 671, doi:10.15779/Z38BG31; Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz and Hanna Wallach, "Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI," (2016).

incomplete,[39] biased or otherwise skewed,[40] often drawing on limited and non-representative samples that are poorly defined before use.[41] Such problems with training data may not be obvious, as datasets may be constructed in non-transparent ways.[42] Additionally, given that humans must label much of the training data by hand, human biases and cultural assumptions are transmitted by classification choices.[43] Exclusion of certain data can, in turn, mean exclusion of sub-populations from what AI is able to "see" and "know."[44] While pernicious, these biases are difficult to find and understand, especially when systems are proprietary, treated as black boxes or taken at face value.[45] Computer scientists have noted that the complexity of machine learning systems not only must face difficulties in interpreting opaque, unsupervised models, but may also take on "technical debt" that makes maintenance and improvement costly—leading to situations where bias may be difficult to identify *and* mitigate.[46]

Non-representative collection of data can also produce bias. Data is expensive, and data at scale is hard to come by. Thus, those who want to train an AI system are drawn to the use of easily available data,[47] often crowd-sourced, scraped, or otherwise gathered from existing user-facing apps and properties. This type of data can easily privilege socioeconomically advantaged populations, those with greater access to connected devices and online services. These same types of bias can also exist when data is collected from particular groups and not others.[48] A recent example comes from an experiment by OpenAI in which a year's worth of messages from the discussion forum Reddit were used as data to train an AI model to "speak."[49] Reddit is itself a skewed sub-population of internet users, and this experiment can give us a sense of the types of bias that can occur when a small,

---

39 David J. Beymer, Karen W. Brannon, Ting Chen, Moritz AW Hardt, Ritwik K. Kumar and Tanveer F. Syeda-Mahmoo, "Machine learning with incomplete data sets," U.S. Patent 9,349,105, issued May 24, 2016.

40 Lisa Gitelman, *Raw data is an oxymoron,* (MIT Press: 2013).

41 Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell and Ross Girshick, "Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels," (In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2930-2939, 2016).

42 Josh Attenberg, Prem Melville, Foster Provost and Maytal Saar-Tsechansky, "Selective data acquisition for machine learning," In *Cost-sensitive machine learning*. (CRC Press: 2011), pp. 101-155; Christian Beyer, Georg Krempl and Vincent Lemaire, "How to select information that matters: a comparative study on active learning strategies for classification," In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, p. 2. ACM, 2015.

43 Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou and Mary Wootters, "Strategic classification." (In Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, pp. 111-122, 2016).

44 Matthew Zook, Solon Barocas, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson and Frank Pasquale, "Ten simple rules for responsible big data research," PLoS Computational Biology 13, No. 3 (2017): e1005399.

45 Frank Pasquale, *The black box society: The secret algorithms that control money and information*, (Harvard University Press, 2015).

46 D. Sculley et al., "Machine Learning: The High Interest Credit Card of Technical Debt," SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop), 2014, https://research.google.com/pubs/pub43146.html.

47 Amanda Levendowski, "How copyright law creates biased artificial intelligence," http://www.werobot2017.com/wp-content/uploads/2017/03/Levendowski-How-Copyright-Law-Creates-Biased-Artificial-Intelligence-Abstract-and-Introduction-1.pdf.

48 Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, Chris Parnin and Jon Stallings, *Gender differences and bias in open source: Pull request acceptance of women versus men*, No. e1733v2. PeerJ Preprints, 2016.

49 Ananya Bhattacharya, "Elon Musk's OpenAI is Using Reddit to Teach AI to Speak Like Humans," *Quartz*, October 12, 2016, https://qz.com/806321/open-ai-reddit-human-conversation.

nonrepresentative group is used as a stand-in for the whole.

Problems may also result from the disconnect between the context in which an AI system is used and the assumptions built into the AI system when it was designed. A group of researchers recently assessed how AI-based mapping apps often provide indirect routes to some users as a way to accomplish traffic load-balancing. The system will not be able to tell when the person asking for directions is driving to the hospital in an emergency. Such decontextualized assumptions can put non-consenting and unaware populations at risk while providing little opportunity for direct input.[50]

While widely acknowledged as a problem, bias within and beyond AI is difficult to measure. Unintended consequences[51] and inequalities are by nature collective, relative and contextual, making measurement and baseline comparisons difficult.[52] Information biases in particular are difficult to measure, given the many possible reference points in context: content, users, ranking and access.[53] There is potential for both over- and under-counting biases in measurement of distributions given the limits on observable circumstances for individuals, problems with population gaps and possible measurement errors.[54]

Given the difficulty (and sometimes even technical impossibility) of understanding exactly how AI systems have reached a given decision,[55] bias is often only revealed by demonstrating an inequality in outcomes, post-hoc. Examples of this are familiar from recent news stories. Julia Angwin's ProPublica piece on Northpointe's racially-biased COMPAS system, used to make sentencing decisions in courts across the United States, is an exemplar of the genre.[56] Similarly, Bloomberg found that Amazon's same-day delivery service was bypassing ZIP codes that are predominantly black. This decision may have been made for many reasons, but its result was racial bias.[57]

The AI Field is Not Diverse

Bias can also emerge in AI systems because of the very narrow subset of the population that design them. AI developers are mostly male, generally highly paid, and similarly

---

50 Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz and Hanna Wallach, "Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI," *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2846909 (2016).

51 Marco J Haenssgen and Proochista Ariana, "The Social Implications of Technology Diffusion: Uncovering the Unintended Consequences of People's Health-Related Mobile Phone Use in Rural India and China," *World Development* 94 (2017): 286-304.

52 Frank Cowell, *Measuring inequality*, (Oxford University Press, 2011).

53 Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul and Gerhard Weikum, "On Measuring Bias in Online Information," *arXiv preprint arXiv:1704.05730* (2017).

54 Ashton Anderson, Jon Kleinberg and Sendhil Mullainathan, "Assessing Human Error Against a Benchmark of Perfection," arXiv preprint arXiv:1606.04956 (2016).

55 Jenna Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society* 3, No. 1 (2016): DOI: https://doi.org/10.1177/2053951715622512.

56 Angwin, Larson and 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

57 David Ingold and Spencer Soper, "Amazon Doesn't Consider the Race of Its Customers. Should It?," Bloomberg, April 21, 2016, https://www.bloomberg.com/graphics/2016-amazon-same-day/.

technically educated. Their interests, needs, and life experiences will necessarily be reflected in the AI they create. Bias, whether conscious or unconscious,[58] reflects problems of inclusion and representation. The lack of women and minorities in tech fields, and artificial intelligence in particular, is well known.[59] But this was not always the case. Early programming and data entry work was characterized as secretarial, and was female-dominated. These women were themselves called "computers," and they were often undercompensated and rarely credited.[60] All the while, they were responsible for things like maintaining sophisticated systems that targeted bomb strikes in World War II[61] and tabulating decades of census data.[62]

The history of AI reflects this pattern of gender exclusion. The 1956 Dartmouth Summer Research Project on Artificial Intelligence, which initiated the concept of artificial intelligence,[63] was exclusively attended by men. Pioneering work in natural language processing and computational linguistics, key to contemporary AI systems, has been credited to male colleagues and students rather than to Margaret Masterman, who founded the Cambridge Language Research Unit and was one of the leaders in the field.[64] Intentional exclusion and unintentional "like-me" bias is responsible for a continued lack of demographic representation within the AI field and within the tech industry for women, Hispanics, and African Americans.[65]

Gender and racial disparities among developer cohorts in tech companies are even more skewed than the demographics of students or academics. In the United States, women make up about 18 percent of computer science (CS) graduates, yet only 11 percent of computer engineers are female. African Americans and Hispanics represent only 11 percent of total technology sector employees although they comprise 27 percent of the overall population.[66] Representation in the U.S. context has wide reaching implications, given that 33 percent of knowledge and technology intensive (KTI) jobs worldwide are U.S. based and those firms contribute 29 percent of global GDP, of which 39 percent are U.S. based.[67] Efforts to address gender biases in Google Ad Settings, revealed in 2015,[68] have failed to

---

58 Cathy O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, (New York: Crown Publishing Group, 2016).

59 Kate Crawford, "Artificial Intelligence's White Guy Problem," *The New York Times*, June 25, 2016.

60 Ellen Van Oost, "Making the Computer Masculine," In *Women, Work and Computerization* (2000), pp. 9-16.

61 Nathan Ensmenger, "Making programming masculine," In *Gender codes: Why women are leaving computing* (2010): 115-42.

62 Margaret Ann Boden, *Mind as machine: A history of cognitive science*, (Clarendon Press, 2006).

63 Ronald Kline, "Cybernetics, automata studies, and the Dartmouth conference on artificial intelligence," *IEEE Annals of the History of Computing* 33, No. 4 (2011): 5-16.

64 Margaret Masterman, "1 Personal background," *Early Years in Machine Translation: Memoirs and Biographies of Pioneers* 97 (2000): 279; William Williams and Frank Knowles, "Margaret Masterman: In memoriam," *Computers and translation* 2, No. 4 (1987): 197-203.

65 Google, Inc. and Gallup, Inc., "Diversity Gaps in Computer Science: Exploring the Underrepresentation of Girls, Blacks, and Hispanics," Retrieved from http://goo.gl/PG34aH. Additional reports from Google's Computer Science Education Research are available at https://edu.google.com/resources/computerscience/research.

66 National Science Foundation, "Science and Engineering Indicators," 2016, Chapter 2, https://nsf.gov/statistics/2016/nsb20161/#/report/chapter-2.

67 National Science Foundation, "Science and Engineering Indicators," 2016, Chapter 6, https://nsf.gov/statistics/2016/nsb20161/#/report/chapter-6.

68 Amit Datta, Michael Carl Tschantz and Anupam Datta, "Automated experiments on ad privacy settings," *Proceedings on*

stop inequality in presentation of STEM job ads, even when language in ads are controlled for gender-neutral language.[69]

AI is not impartial or neutral. Technologies are as much products of the context in which they are created as they are potential agents for change.[70] Machine predictions and performance are constrained by human decisions and values,[71] and those who design, develop, and maintain AI systems will shape such systems within their own understanding of the world.[72] Many of the biases embedded in AI systems are products of a complex history with respect to diversity and equality.

Recent Developments in Bias Research

In the year since the *AI Now 2016 Symposium,* there has been a bumper crop of new research on bias in machine learning. One promising development is that many of these studies have reflexively used AI techniques to understand the the ways by which AI systems introduce or perpetuate unequal treatment.

New research on word embeddings has shown the ways in which language, as it is used within our complex and often biased social contexts, reflects bias.[73] Word embeddings are set of natural language processing techniques that map the semantic relationship between words, creating a model that predicts which words are likely to be associated with which. Researchers looking at word embeddings showed that predictable gendered associations between words, such as “female” and “queen” are reflected in the models, as are stereotypes, such as “female” and “receptionist,” while “man” and typically masculine names are associated with programming, engineering and other STEM professions.[74]

Such biases have daily, real-world impacts. Recent analysis of search results and advertisements similarly reveals persistent gendered, racial and cultural biases.[75]

Privacy Enhancing Technologies 2015, No. 1 (2015): 92-112.

69 Anja Lambrecht and Catherine E. Tucker, “Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads,” October 13, 2016. Available at SSRN: https://ssrn.com/abstract=2852260 or http://dx.doi.org/10.2139/ssrn.2852260.

70 Zdenek Smutny, "Social informatics as a concept: Widening the discourse," *Journal of Information Science* 42, No. 5 (2016): 681-710.

71 Kenneth A. Bamberger and Deirdre Mulligan, "Public Values, Private Infrastructure and the Internet of Things: the Case of Automobile," *Journal of Law & Economic Regulation* 9 (2016): 7-44; Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan, “Human decisions and machine predictions,” No. w23180. National Bureau of Economic Research, 2017.

72 Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter and Luciano Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society* 3, No. 2 (2016): 2053951716679679.

73 Aylin Caliskan, Joanna J. Bryson and Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science* 356, No. 6334 (2017): 183-186; Anthony G. Greenwald, "An AI stereotype catcher," *Science* 356, No. 6334 (2017): 133-134.

74 Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama and Adam Kalai, "Quantifying and reducing stereotypes in word embeddings," *arXiv preprint arXiv:1606.06121* (2016); Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama and Adam T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," In *Advances in Neural Information Processing Systems*, pp. 4349-4357, 2016.

75 Datta, Tschantz, and Datta, 2015; Tarleton Gillespie, “Algorithmically recognizable: Santorum’s Google problem, and Google’s Santorum problem,” *Information, Communication & Society* 20, No. 1 (2017): 63-80; Safiya Umoja Noble, “Algorithms of Oppression: How Search Engines Enforce Racism,” (NYU Press, forthcoming 2018).

New work has also highlighted the way in which AI poses risks of significant bias-driven impacts in the educational context, where K-12 educators subject children to treatment, discipline and tracking decisions based on AI-determined characterizations of their abilities and behaviors.[76] Analysis of large data sets reflecting STEM education in K-12 classrooms reveals racial disparities in disciplinary actions and recommendations for advanced coursework. These data, along with the biases they reflect, are very likely to be used to train these educational AI systems, which would then reproduce and further normalize these biases.

In a study that examined the potential for bias, the Human Rights Data Analysis Group demonstrated how commonly used predictive policing system PredPol, were it used in Oakland, CA, would reinforce racially-biased police practices by recommending increased police deployment in neighborhoods of color.[77] Decades of policing research has shown that foot patrols and community rapport decrease policing biases, while studies of "driving while black" and "hot spots" illustrate biases in routine strategies.[78] New technologies appear to prevent the former and amplify the latter, reproducing the most extreme racial stereotyping.[79]

Legal scholarship has also explored the applications of machine testimony at criminal trials,[80] among many possible instances identified in which these skewed systems and biased data could negatively impact human lives due to reproducing stereotypes, with the added challenge that the systems are poorly understood and proprietary.

When bias is embedded in AI health applications, it can have an incredibly high cost. Worryingly, data sets used to train health-related AI often rely on clinical trial data, which are historically skewed toward white men, even when the health conditions studied primarily affect people of color or women.[81] Even without AI amplifying such biases, African Americans with sickle cell anemia are overdiagnosed and unnecessarily treated for diabetes based on insights from studies that excluded them.[82] The prevalence of biases when combined with opacity and inscrutability leads to a lack of trust in AI currently being

---

[76] Benjamin Herold, "Algorithmic Bias as a Rising Concern for Ed-Tech Field, RAND Researchers Say," Education Week, April 11, 2017, http://blogs.edweek.org/edweek/DigitalEducation/2017/04/algorithmic_bias_edtech_RAND.html.

[77] Kristian Lum and William Isaac, "To predict and serve?," *Significance* 13, No. 5 (2016): 14-19.

[78] Prashan Ranasinghe, "Rethinking the Place of Crime in Police Patrol: A Re-Reading of Classic Police Ethnographies," *British Journal of Criminology* (2016): azw028; Patricia Warren, Donald Tomaskovic-Devey, William Smith, Matthew Zingraff and Marcinda Mason, "Driving while black: Bias processes and racial disparity in police stops," *Criminology* 44, No. 3 (2006): 709-738; David Weisburd, "Does Hot Spots Policing Inevitably Lead to Unfair and Abusive Police Practices, or Can We Maximize Both Fairness and Effectiveness in the New Proactive Policing," *University of Chicago Legal Forum* (2016): 661-689.

[79] Andrew Guthrie Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, (NYU Press, forthcoming 2017).

[80] Andrea Roth, "Machine Testimony," Yale Law Journal, forthcoming 2017.

[81] Anita Kurt, Lauren Semler, Jeanne L. Jacoby, Melanie B. Johnson, Beth A. Careyva, Brian Stello, Timothy Friel, Mark C. Knouse, Hope Kincaid and John C. Smulian, "Racial Differences Among Factors Associated with Participation in Clinical Research Trials," *Journal of Racial and Ethnic Health Disparities* (2016): 1-10.

[82] Mary E Lacy, Gregory A. Wellenius, Anne E. Sumner, Adolfo Correa, Mercedes R. Carnethon, Robert I. Liem, James G. Wilson, David B. Saks, David R. Jacobs Jr., April Carson, Xi Luo, Annie Gjelsvik, Alexander P. Reiner, Rhaki Naik, Simin Liu, Solomon K. Musani, Charles B. Eaton and Wen-Chih Wu, "Association of sickle cell trait with hemoglobin A1c in African Americans," *Jama* 317, No. 5 (2017): 507-515.

developed for neuroscience and mental health applications.[83] The prospect of misdiagnosis or improper treatment leading to patient death motivates some to avoid AI systems entirely in the health context.[84]

Emerging Strategies to Address Bias

There is an urgent need to expand cultural, disciplinary and ethnic diversity within the AI field in order to diminish groupthink, mitigate bias and broaden intellectual frames of reference beyond the purely technical. While some have suggested that AI systems can be used to address diversity problems at companies,[85] if AI development is not inclusive, the success of such a bootstrapped approach is doubtful. There have been positive developments prompting inclusion within the AI community, such as Fei-Fei Li's SAILORS summer camp, a program that helps high school girls acquire comfort and experience with AI.[86] Similarly, the Association of Computing Machinery (ACM) increasingly recognizes the need to address algorithmic bias and emphasize diversity.[87] Various conferences have also sought to explore accountability and transparency issues surrounding AI and algorithmic systems as a way to better understand and evaluate biases.[88] Among conferences, the Fairness, Accountability, and Transparency in Machine Learning (FAT/ML and now FAT*) Conferences are notable for a focus on technical research and experimentation dedicated to making AI more inclusive, legible and representative.[89]

While steps are being made to understand and combat bias in some sectors, bias can also be profitable. Insurance and financial lending have long discriminated for their financial advantage, choosing to serve the least risky and, sometimes, leaving the most vulnerable behind.[90] AI systems are now being used to make credit and lending decisions. When underwriting decisions are made by AI systems trained on data that reflects past biased practices and calibrated to detect nuanced signals of "risk," creditors will be able to make more profitable loans while leaving those in precarious situations behind. Due to misaligned interests and the information asymmetry that AI exacerbates in these industries, new incentives for fairness and new methods for validating fair practices need

---

83 Andreas Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," *Brain Informatics* 3, No. 2 (2016): 119-131.

84 Rich Caruana, "Intelligible Machine Learning for Critical Applications Such As Health Care," *2017 AAAS Annual Meeting (February 16-20, 2017)*, AAAS, 2017.

85 Ji-A Min, "Ten ways HR tech leaders can make the most of artificial intelligence," Personnel Today, April 26, 2017, http://www.personneltoday.com/hr/ten-ways-hr-tech-leaders-can-make-artificial-intelligence/.

86 Marie E Vachovsky, Grace Wu, Sorathan Chaturapruek, Olga Russakovsky, Richard Sommer and Li Fei-Fei, "Toward More Gender Diversity in CS through an Artificial Intelligence Summer Program for High School Girls," (In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pp. 303-308), ACM, 2016.

87 Kieth Kirkpatrick, "Battling algorithmic bias: how do we ensure algorithms treat us fairly?," *Communications of the ACM* 59, No. 10 (2016): 16-17.

88 Algorithms and Explanations, Information Law Insitute, New York University, April 27-28, 2017, http://www.law.nyu.edu/centers/ili/events/algorithms-and-explanations.

89 3rd Workshop on Fairness, Accountability, and Transparency in Machine Learning, New York, November 18, 2016, http://www.fatml.org/.

90 JM Schumacher, "Linear Versus Nonlinear Allocation Rules in Risk Sharing Under Financial Fairness," (March 2, 2017), http://dx.doi.org/10.2139/ssrn.2892760.

to be developed.[91]

Part of the fundamental difficulty in defining, understanding and measuring bias stems from the contentious and conceptually difficult task of defining fairness. Tradeoffs are inherent in the adoption of particular fairness definitions, possibly perpetuating particular biases in the service of addressing others.[92] Recent efforts have sought to implement fairness by mathematically specifying social norms and values, then using those specifications as constraints when training AI systems.[93] While these are hopeful developments, none of these methods cleanly solve the problem of bias. Understanding AI not as a purely technical implementation, but as a contextually-specific combination of norms, technical systems and strategic interests is an important step toward addressing bias in AI.[94] There continues to be a deep need for interdisciplinary, socially aware work that integrates the long history of bias research from the social sciences and humanities into the field of AI research.

Rights and Liberties

In the period since the *AI Now 2016 Symposium*, the global political landscape has shifted considerably. The election of Donald Trump is part of a larger wave of populist political movements across the globe, and shares with these a number of hallmark traits. In governing, populists seek to delegitimize political opposition—from opposition parties to institutions like the media and the judiciary—and to crack down on perceived threats to the imagined homogeneous people they claim to represent.[95] While regional instantiations vary, they share an opposition to existing political elites and a nationalist, anti-pluralist approach that claims a moral imperative to represent a silent majority.

The election of Emmanuel Macron in France and the gains by Labour in the UK may indicate a coming backlash to the global populist wave, but given the strong showing from Germany's far-right Alternative für Deutschland party in their 2017 elections, this is by no means certain.

It remains necessary to ask how AI systems are likely to be deployed in governing, and how

---

91 Hamid R. Ekbia and Bonnie A. Nardi, *Heteromation, and Other Stories of Computing and Capitalism*, (MIT Press, 2017); Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra and Z. Steven Wu, "Fairness Incentives for Myopic Agents," arXiv preprint arXiv:1705.02321 (2017); Julia Lane, "Perspective: Fix the incentives," *Nature* 537, No. 7618 (2016): S20-S20.

92 Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan, "Inherent trade-offs in the fair determination of risk scores," arXiv preprint arXiv:1609.05807 (2016).

93 Yiling Chen, Arpita Ghosh, Michael Kearns, Tim Roughgarden and Jennifer Wortman Vaughan, "Mathematical foundations for social computing," *Communications of the ACM* 59, No. 12 (2016): 102-108; Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern and Aaron Roth, "Fair Learning in Markovian Environments," arXiv preprint arXiv:1611.03071 (2016); Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel and Aaron Roth, "Rawlsian fairness for machine learning," arXiv preprint arXiv:1610.09559 (2016).

94 Mike Ananny, "Toward an ethics of algorithms: Convening, observation, probability, and timeliness," *Science, Technology, & Human Values* 41, No. 1 (2016): 93-117.

95 Jan-Werner Müller, *What Is Populism?* (Philadelphia: University of Pennsylvania Press, 2016).

they might be used within populist and authoritarian contexts.[96] What effects will these systems have on vulnerable individuals and minorities? How will AI systems be used by law enforcement or national security agencies? How will AI's use in the criminal justice system affect our understanding of due process and the principle of equal justice under the law? How might complex AI systems centralize authority and power? This section examines these questions, describes applications of AI that pose challenges to rights and liberties, and touches on the technical and normative frameworks we might construct to ensure AI can be a force for good in the face of our contemporary political realities.

Population Registries and Computing Power

In political contexts where minorities or opposition points of view are seen as threats to an imagined homogeneous "people," information technology has been used to monitor and control these segments of a population. Such techno-political projects often build on older colonial histories of censuses and population registries,[97] as well as racialized modes of surveillance and control rooted in the Atlantic slave trade and the plantation system. In *Dark Matters*, Simone Browne connects this deep history of surveillance to contemporary biometric techniques of governing black bodies.[98]

The Book of Life registry project in apartheid South Africa is a useful modern example. In that project, which ran from 1967 to 1983, IBM assisted South Africa in classifying its population by racial descent. This system was used to move all so-called 'non-white citizens' from their homes into segregated neighborhoods.[99] The Book of Life was plagued by technical and operational problems and eventually abandoned. However, as Paul Edwards and Gabrielle Hecht note, "technopolitical projects do not need to fully achieve their technical goals in order to 'work' politically... The registries 'worked' to establish racialized personal identities as elements of governance."[100] As Kate Crawford has recently argued, registries like the Book of Life were reinforcing a way of thinking that was itself autocratic.[101]

More recent computerized registries like The National Security Entry-Exit Registration System (NSEERS) proliferated in the United States and among its allies following the attacks of September 11, 2011. NSEERS centralized documentation for non-citizens in the United States who hailed from a list of 25 predominantly Muslim countries that the Bush administration deemed dangerous. As with the Book of Life, NSEERS' effectiveness in its

---

96 See Kate Crawford, "Dark Days: AI and the Rise of Fascism," SXSW featured talk, March 15 2017, https://www.youtube.com/watch?v=Dlr4O1aEJvI.

97 For just one example in colonial India, v.: Radhika Singha, "Settle, Mobilize, Verify: Identification Practices in Colonial India," *Studies in History* 16, No. 2 (August 1, 2000): 151–98, doi:10.1177/025764300001600201.

98 Simone Browne, *Dark Matters: On the Surveillance of Blackness,* (Durham: Duke University Press, 2015).

99 National Action/Research on the Military-Industrial Complex and American Friends Service Committee, *Automating Apartheid: U.S. Computer Exports to South Africa and the Arms Embargo*, (Philadelphia: NARMIC/America Friends Service Committee, 1982), 19.

100 Paul N. Edwards and Gabrielle Hecht, "History and the technopolitics of identity: The case of apartheid South Africa," *Journal of Southern African Studies* 36, No. 3 (2010): 619-639.

101 Kate Crawford, "Dark Days: AI and the Rise of Fascism," SXSW featured talk, March 15 2017, https://www.youtube.com/watch?v=Dlr4O1aEJvI.

stated goal of stopping domestic terrorism was questionable, and it was dismantled in the final days of the Obama administration (although the data collected by the program still exists).[102] Consistent with Edwards' and Hecht's analysis, NSEERS set into motion state projects of Muslim surveillance and deportation.[103]

The history and political efficacy of registries exposes the urgent need for lines of research that can examine the way citizen registries work currently, enhanced by data mining and AI techniques, and how they may work in the future.[104] Contemporary AI systems intensify these longer-standing practices of surveillance and control. Such systems require the collection of massive amounts of data, which is now possible at large scale via the Internet and connected devices. When these practices are carried out by private enterprise in addition to states, as we will discuss in the next section, they introduce new forms of value extraction and population control unregulated and often unacknowledged by current legal frameworks.[105]

Corporate and Government Entanglements

It remains critically important to understand the history of AI and its shifting relationship to the state. In the mid-twentieth century, advanced computing projects tended to be closely associated with the state, and especially the military agencies who funded their fundamental research and development.[106] Although AI emerged from this context, its present is characterized by a more collaborative approach between state agencies and private corporations engaged in AI research and development. As Gary Marchant and Wendell Wallach argue, governance has expanded far beyond both governmental institutions and legal codes to include a wide range of industry standards and practices that will shape how AI systems are implemented.[107]

Palantir—co-founded by Trump supporter and advisor Peter Thiel with seed money from the CIA's venture capital fund In-Q-Tel—typifies this dynamic.[108] Gotham, Palantir's national security and government software, allows analysts to easily combine, query and visualize structured and unstructured data at large scales.[109] AI can now be used in Palantir products for activities such as lead generation, including a bank's ability to identify

---

[102] J. David Goodman and Ron Nixon, "Obama to Dismantle Visitor Registry Before Trump Can Revive It," *The New York Times*, December 22, 2016, https://www.nytimes.com/2016/12/22/nyregion/obama-to-dismantle-visitor-registry-before-trump-can-revive-it.html.

[103] "Raza v. City of New York - Legal Challenge to NYPD Muslim Surveillance Program," American Civil Liberties Union, March 6, 2017, https://www.aclu.org/cases/raza-v-city-new-york-legal-challenge-nypd-muslim-surveillance-program.

[104] Kate Crawford, "Letter to Silicon Valley," *Harper's Magazin*e, Feb 9, 2017, https://harpers.org/archive/2017/02/trump-a-resisters-guide/11/.

[105] Julie E. Cohen, "The Biopolitical Public Domain: The Legal Construction of the Surveillance Economy," Philosophy & Technology, March 28, 2017, 1–21, doi:10.1007/s13347-017-0258-2.

[106] Paul N. Edwards, The Closed World: *Computers and the Politics of Discourse in Cold War America* (Cambridge, MA: The MIT Press, 1996).

[107] Gary E. Marchant and Wendell Wallach, "Coordinating Technology Governance | Issues in Science and Technology," Issues in Science and Technology XXXI, no. 4 (Summer 2015), http://issues.org/31-4/coordinating-technology-governance/.

[108] Sam Biddle, "How Peter Thiel's Palantir Helped the NSA Spy on the Whole World," *The Intercept*, February 22, 2017, https://theintercept.com/2017/02/22/how-peter-thiels-palantir-helped-the-nsa-spy-on-the-whole-world/.

[109] Ibid.

anomalous credit card activity for fraud protection. More advanced capabilities are available to national security clients as well. How rights and liberties need to be understood and reconfigured in the face of opaque public-private AI systems is still an open question.

Immigration and law enforcement are critical within this debate. In the United States, Immigration and Customs Enforcement (ICE) is expanding its technological reach through tools like Investigative Case Management (ICM), a platform that allows agents to access a wide variety of previously separate databases, including information on a suspect's "schooling, family relationships, employment information, phone records, immigration history, foreign exchange program status, personal connections, biometric traits, criminal records and home and work addresses."[110] This is another Palantir system, first procured by the Obama administration in 2014 and scheduled to become operational late in 2017.

Other law enforcement agencies are currently integrating AI and related algorithmic decision-support systems from the private sector into their existing arsenals. Axon (formerly Taser International) is a publicly traded maker of law enforcement products, including their famous electroshock weapon. The company has now shifted toward body camera technologies, recently offering them for free to any police department in the U.S.[111] In 2017, Axon started an AI division following their acquisition of two machine vision companies. Among their goals is to more efficiently analyze the over 5.2 petabytes of data that they have already acquired from their existing camera systems. Video expands Axon's existing Digital Evidence Management System, signaling a larger shift beyond machine learning and natural language processing of textual sources.[112] Axon CEO Rick Smith has argued that the vast scale of existing law enforcement data could help drive research in machine vision as a whole: "We've got all of this law enforcement information with these videos, which is one of the richest treasure troves you could imagine for machine learning."[113] There are real concerns about the forms of bias embedded in these data sets, and how they would subsequently function as training data for an AI system.

There are some who argue in favor of body camera and machine vision systems for supporting civil liberties, including enhanced law enforcement transparency and accountability.[114] Axon promises that its AI techniques will reduce the time officers currently spend on report-writing and data entry.[115] However, Axon's new focus on

---

[110] Spencer Woodma, "Palantir Provides the Engine for Donald Trump's Deportation Machine," *The Intercept*, March 2, 2017, https://theintercept.com/2017/03/02/palantir-provides-the-engine-for-donald-trumps-deportation-machine/.

[111] Laurel Wamsley, "Taser Changes Its Name To Axon And Offers Free Body Cameras For Police," *NPR*, Arpil 7, 2017, http://www.npr.org/sections/thetwo-way/2017/04/07/522878573/we-re-more-than-stun-guns-says-taser-as-it-changes-company-name.

[112] "TASER Makes Two Acquisitions to Create 'Axon AI,'" *Police Magazine*, February 9, 2017, http://www.policemag.com/channel/technology/news/2017/02/09/taser-makes-two-acquisitions-to-create-axon-ai.aspx.

[113] Doug Wyllie, "What TASER's Acquisition of 2 AI Companies Means for the Future of Policing," *PoliceOne*, February 10, 2017, https://www.policeone.com/police-products/less-lethal/TASER/articles/289203006-What-TASERs-acquisition-of-2-AI-companies-means-for-the-future-of-policing/.

[114] Jay Stanley, "Police Body-Mounted Cameras: With Right Policies in Place, a Win for All," (ACLU, 2013), http://www.urbanaillinois.us/sites/default/files/attachments/police-body-mounted-cameras-stanley.pdf.

[115] Alex Pasternack, "Police Body Cameras Will Do More Than Just Record You," *Fast Company*, March 3, 2017, https://www.fastcompany.com/3061935/police-body-cameras-livestreaming-face-recognition-and-ai.

predictive methods of policing—inspired by Wal-Mart's and Google's embrace of deep learning to increase sales—raises new civil liberties concerns. Instead of purchasing patterns, these systems will be looking for much more vague, context-dependent targets, like "suspicious activity." Behind appearances of technical neutrality, these systems rely on deeply subjective assumptions about what constitutes suspicious behavior or who counts as a suspicious person.[116]

Unsurprisingly, machine vision techniques may reproduce and present as objective existing forms of racial bias.[117] Researchers affiliated with Google's Machine Intelligence Group and Columbia University make a compelling comparison between machine learning systems designed to predict criminality from facial photos and discredited theories of physiognomy—both of which problematically claim to be able to predict character or behavioral traits simply by examining physical features.[118] More generally, Cathy O'Neil identifies the potential for advanced AI systems in law enforcement to create a "pernicious feedback loop"—if these systems are built on top of racially-biased policing practices, then their training data will reflect these existing biases, and integrate such bias into the logic of decision making and prediction.[119]

Ethical questions of bias and accountability will become even more urgent in the context of rights and liberties as AI systems capable of violent force against humans are developed and deployed in law enforcement and military contexts. Robotic police officers, for example, recently debuted in Dubai.[120] If these were to carry weapons, new questions would arise about how to determine when the use of force is appropriate. Drawing on analysis of the Black Lives Matter movement, Peter Asaro has pointed to difficult ethical issues involving how lethal autonomous weapons systems (LAWS) will detect threats or gestures of cooperation, especially involving vulnerable populations. He concludes that AI and robotics researchers should adopt ethical and legal standards that maintain human control and accountability over these systems.[121]

Similar questions apply in the military use of LAWS. Heather Roff argues that fully autonomous systems would violate current legal definitions of war that require human judgment in the proportionate use of force, and guard against targeting of civilians. Furthermore, she argues that AI learning systems may make it difficult for commanders to even know how their weapons will respond in battle situations.[122] Given these legal, ethical

---

[116] Ava Kofman, "Taser Will Use Police Body Camera Videos 'to Anticipate Criminal Activity,'" *The Intercept*, April 30, 2017, https://theintercept.com/2017/04/30/taser-will-use-police-body-camera-videos-to-anticipate-criminal-activity/.

[117] Clare Garvie and Jonathan Frankle, "Facial-Recognition Software Might Have a Racial Bias Problem," *The Atlantic*, April 7, 2016, https://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/.

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

[119] Kofman, "Taser Will Use Police Body Camera Videos 'to Anticipate Criminal Activity.'"

[120] Rory Cellan-Jones, "Robot Police Officer Goes on Duty in Dubai," *BBC News*, May 24, 2017, sec. Technology, http://www.bbc.com/news/technology-40026940.

[121] Peter Asaro, "'Hands Up, Don't Shoot!' HRI and the Automation of Police Use of Force," *Journal of Human-Robot Interaction* 5, No. 3 (December 14, 2016): 55–69.

[122] Heather M. Roff, "Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous

and design concerns, both researchers call for strict limitations on the use of AI in weapons systems.

While predictive policing and the use of force have always been important issues, they take on new salience in populist or authoritarian contexts. As AI systems promise new forms of technical efficiency in the service of safety, we may need to confront a fundamental tension between technological efficiency and a commitment to ideals of justice.

AI and the Legal System

The legal system is the institution tasked with defending civil rights and liberties. Thus, there are two separate questions to consider regarding AI and the legal system: 1) Can the legal system serve the rights-protection functions it is expected to when an AI system produces an unfair result? And, 2) How and where (if at all) should the legal system incorporate AI?

Scholars like Kate Crawford and Jason Schultz have identified a series of conflicts between AI techniques and constitutional due process requirements,[123] such as how AI techniques affect procedural considerations and equal justice under the law. The proliferation of predictive systems demands new regulatory techniques to protect legal rights. Danielle Citron and Frank Pasquale argue that safeguards to rights should be introduced at all stages of the implementation of an AI system, from safeguarding privacy rights in data collection to public audits of scoring systems that critically affect the public in areas like employment and healthcare.[124]

In a similar vein, Andrew Selbst has argued that an impact assessment requirement can force those building and buying AI systems to make explicit the normative choices they are making before implementing them.[125] And as Lilian Edwards and Michael Veale[126] have pointed out, the new EU General Data Protection Regulation (GDPR) includes a requirement for data protection impact assessments, the import of which is unclear as yet. There is also a rapidly emerging scholarly debate about the value of requiring an explanation or interpretation of AI and machine learning systems as a regulatory technique to ensure individual rights,[127] how to operationalize such a requirement,[128] whether such a

---

Weapons" (Geneva: Review Conference of the Convention on Certain Conventional Weapons, December 2016), https://globalsecurity.asu.edu/sites/default/files/files/Control-or-Judgment-Understanding-the-Scope.pdf.

123 Kate Crawford and Jason Schultz, "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms," *Boston College Law Review* 55, No. 1 (January 29, 2014): 93.

124 Danielle Keats Citron and Frank A. Pasquale, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* 89, No. 1 (2014): 1–33.

125 Andrew D. Selbst, "Disparate Impact in Big Data Policing", Georgia Law Review, forthcoming 2017. SSRN preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2819182.

126 Lilian Edwards and Michael Veale, "Slave to the Algorithm? Why a 'Right to Explanation' is Probably Not the Remedy You are Looking for", SSRN preprint, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855.

127 Ibid; Kiel Brennan-Marquez, "'Plausible Cause': Explanatory Standards in the Age of Powerful Machines, *Vanderbilt Law Review* (2017) vol. 70, p. 1249; Andrew D. Selbst, "A Mild Defense of Our New Machine Overlords," *Vanderbilt Law Review En Banc* (2017) vol. 70, p. 87; Katherine Jo Strandburg, "Decisionmaking, machine learning and the value of explanation," (The Human Use of Machine Learning: An Interdisciplinary Workshop, 16 December 2016), http://www.dsi.unive.it/HUML2016/assets/Slides/Talk%202.pdf.

requirement presently exists under the GDPR[129] and more generally how competing interpretations or explanations might be technically formulated and understood by different stakeholders.[130]

The criminal justice system's implementation of risk assessment algorithms provides an example of the legal system's use of AI and its attendant risks.[131] Proponents of risk-based sentencing argue that evidence-based machine learning techniques can be used in concert with the expertise of judges to improve the accuracy of prior statistical and actuarial methods for risk forecasting, such as regression analysis.[132] Along these lines, a recent study by computer scientist Jon Kleinberg, Sendhil Mullainathan, and their co-authors showed that a predictive machine learning algorithm could be used by judges to reduce the number of defendants held in jail as they await trial by making more accurate predictions of future crimes.[133]

While algorithmic decision-making tools show promise, many of these researchers caution against misleading performance measures for emerging AI-assisted legal techniques.[134] For example, the value of recidivism as a means to evaluate the correctness of an algorithmically-assigned risk score is questionable because judges make decisions about risk in sentencing, which, in turn, influences recidivism – or, those assessed as "low risk" and subsequently released are the only ones who will have an opportunity to re-offend, making it difficult to measure the accuracy of such scoring. Meanwhile, Rebecca Wexler has documented the disturbing trend of trade secret doctrine being expressly adopted in courts to prevent criminal defendants from asserting their rights at trial.[135]

Sandra Mayson has recently written on risk assessment in the bail reform movement. Well-intentioned proponents of bail reform argue that risk assessment can be used to spare poor, low-risk defendants from onerous bail requirements or pretrial incarceration. Such arguments tend to miss the potential of risk assessment to "legitimize and entrench" problematic reliance on statistical correlation, and to "[lend such assessments] the aura of scientific reliability."[136] Mayson argues that we also need to ask deeper questions about

---

[128] Andrew D. Selbst and Solon Barocas, "Regulating Inscrutable Systems," in progress.

[129] Bryce Goodman and Seth Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" ICML Workshop on Human Interpretability in Machine Learning, *arXiv preprint*: arXiv:1606.08813 (v3) (2016); forthcoming, *AI Magazine* (2017); Sandra Wachter, Brent Mittelstadt and Luciano Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Protection Law* (2017), https://doi.org/10.1093/idpl/ipx005.

[130] Zachary C. Lipton, "The Mythos of Model Interpretability," arXiv preprint [Cs, Stat], June 10, 2016, http://arxiv.org/abs/1606.03490.

[131] Richard Berk and Jordan Hyatt, "Machine Learning Forecasts of Risk to Inform Sentencing Decisions," *Federal Sentencing Reporter* 27, No. 4 (April 1, 2015): 222–28, doi:10.1525/fsr.2015.27.4.222.

[132] Berk and Hyatt, "Machine Learning Forecasts of Risk to Inform Sentencing Decisions," 222.

[133] Jon Kleinberg et al., "Human Decisions and Machine Predictions," Working Paper (National Bureau of Economic Research, February 2017), doi:10.3386/w23180. http://nber.org/papers/w23180.

[134] Jon Kleinberg, Jens Ludwig and Sendhil Mullainathan, "A Guide to Solving Social Problems with Machine Learning," *Harvard Business Review*, December 8, 2016, https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning.

[135] Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System," SSRN preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2920883.

[136] Sandra G. Mayson, "Bail Reform and Restraint for Dangerousness: Are Defendants a Special Case?",*SSRN Scholarly Paper*

how pretrial restraints are justified in the first place. In other words, policymakers who hope to employ risk assessment in bail reform and pretrial forms of detention need to publicly specify what types of risks can justify these such restraints on liberty, as defendants receiving these scores have not been convicted of anything and these restraints are not imposed on dangerous individuals in the rest of society.

Separately, criminologist Richard Berk and his colleagues argue that there are intractable tradeoffs between accuracy and fairness—the occurrence of false positives and negatives—in populations where base rates (the percentage of a given population that fall into a specific category) vary between different social groups.[137] Difficult decisions need to be made about how we value fairness and accuracy in risk assessment. It is not merely a technical problem, but one that involves important value judgments about how society should work. Left unchecked, the legal system is thus as susceptible to perpetuating AI-driven harm as any other institution.

Finally, machine learning and data analysis techniques are also being used to identify and explain the abuses of rights. Working with human rights advocates in Mexico, the Human Rights Data Analysis Group created a machine learning model that can help guide the search for mass graves.[138]

AI and Privacy

AI challenges current understandings of privacy and strains the laws and regulations we have in place to protect personal information. Established approaches to privacy have become less and less effective because they are focused on previous metaphors of computing, ones where adversaries were primarily human. AI systems' intelligence, as such, depends on ingesting as much training data as possible. This primary objective is adverse to the goals of privacy. AI thus poses significant challenges to traditional efforts to minimize data collection and to reform government and industry surveillance practices.

Of course, privacy as a "right" has always been unevenly distributed. Rights-based discourses are regularly critiqued as being disproportionately beneficial to the privileged while leaving many vulnerable populations partially or entirely exposed. Yet what is different with AI and privacy is that while individualistic and rights-based conceptualizations of privacy remain important to some of the systems at work today, computational systems are now operating outside of the data collection metaphors that privacy law is built on. We are in new terrain, and one that 20th century models of privacy are not designed to contend with.

For example, privacy discourse has not sufficiently accounted for the growing power asymmetries between the institutions that accumulate data and the people who generate

(Rochester, NY: Social Science Research Network, August 15, 2016), https://papers.ssrn.com/abstract=2826600, 2.

[137] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns and Aaron Roth, "Fairness in Criminal Justice Risk Assessments: The State of the Art," arXiv:1703.09207, March 27, 2017.

[138] J. M. Porup, "Hunting for Mexico's Mass Graves with Machine Learning," Ars Technica UK, April 17, 2017, https://arstechnica.co.uk/information-technology/2017/04/hunting-for-mexicos-mass-graves-with-machine-learning/.

that data, even as they are approaching threshold levels which may make these asymmetries very hard to reverse. Models of privacy based on data as a tradable good fail to contend with this power difference. People cannot trade effectively with systems they do not understand, particularly when the system understands them all too well and knows how to manipulate their preferences. Additionally, adaptive algorithms are changing constantly, such that even the designers who created them cannot fully explain the results they generate. In this new model of computational privacy adversaries, both power and knowledge gaps will continue to widen. We must ask how 'notice and consent' is possible or what it would mean to have 'access to your data' or to 'control your data' when so much is unknown or in flux.

There has also been a shift in the quality of the data used for AI. In order to help develop sophisticated diagnostic models, designers often seek to use inputs that are extremely sensitive in nature. For example, in the case of DeepMind's partnership with the UK's National Health Service, the company acquired large amounts of very sensitive public health data. Even though this data may have been required for some of the project's goals, the resulting backlash and government censure[139] illustrate the emerging tensions related to the AI industry's use of such data and the current limits of democratic processes to address questions of agency, accountability and oversight for these endeavors.

The expansion of AI into diverse realms like urban planning also raises privacy concerns over the deployment of IoT devices and sensors, arrayed throughout our daily lives, tracking human movements, preferences and environments.[140] These devices and sensors collect the data AI requires to function in these realms. Not only does this expansion significantly increase the amount and type of data being gathered on individuals, it also raises significant questions around security and accuracy as IoT devices are notoriously insecure, and often difficult to update and maintain.[141]

AI's capacity for prediction and inference also adds to the set of privacy concerns. Much of the value that AI offers is the ability to predict or "imagine" information about individuals and groups that is otherwise difficult to collect, compute or distribute. As more AI systems are deployed and focus on ever-more granular levels of detail, such "predictive privacy harms" will become greater concerns, especially if there are few or no due process constraints on how such information impacts vulnerable individuals.[142] Part of the promise of predictive techniques is to make accurate, often intimate deductions based on a seemingly-unrelated pieces of data or information, such as detecting substance abusers from Facebook posts[143], or identifying gang members based on Twitter data.[144] Significant

---

[139] Information Commissioner's Office, "Royal Free - Google DeepMind trial failed to comply with data protection law," July 3 (2017) https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/

[140] See, for example, Nvida's Deep Learning for Smart Cities: https://www.nvidia.com/en-us/deep-learning-ai/industries/ai-cities/

[141] There have been several serious attacks on network infrastructure and logistics from hacked IoT networked devices in the last 12 months. For example: Andrew Peterson, "Internet of Things compounded Friday's hack of major web sites," *Washington Post,* October 21 (2016)

[142] Kate Crawford and Jason Schultz, "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms," *Boston College Law Review* 55, No. 1 (January 29, 2014): 93.

[143] Tao Ding, Warren Bickel and Shimei Pan, "Social Media-based Substance Use Prediction," May 16 (2017)

shifts are needed in the legal and regulatory approaches to privacy if they are to keep pace with the emerging capacities of AI systems.

Ethics and Governance

So far, this report has addressed issues of power, markets, bias, fairness and rights and liberties – all subjects closely tied to ethics. This section presents a distinct discussion of ethics in the uses, deployment and creation of AI.[145]

Ethical questions surrounding AI systems are wide-ranging, spanning creation, uses and outcomes. There are important questions about which set of values and interests are reflected in AI, as well as how machines can recognize values and ethical paradigms. An important distinction in this area is between what is called 'machine ethics' and the wider domain of the ethics of AI. Machine ethics is more narrowly and explicitly concerned with the ethics of artificially intelligent beings and systems; Isaac Asimov's laws of robotics are one example that captured the popular imagination. AI ethics concerns wider social concerns about the effects of AI systems and the choices made by their designers and users. Here, we are mostly concerned with the latter approach.

AI is certainly not unique among emerging technologies in creating ethical quandaries, and, similar to other computational technologies, AI ethics have roots in the complex history of military influence on computing development and the more recent commercialization and corporate dominance of networked technologies. Yet ethical questions in AI research and development present unique challenges in that they ask us to consider whether, when and how machines should to make decisions about human lives - and whose values should guide those decisions.

Ethical Concerns in AI

Articulating ethical values for AI systems has never been simple. In the 1960s, AI pioneer Joseph Weizenbaum created the early chatbot system ELIZA as a technical demonstration of a system capable of maintaining an interrogative "conversation" with a human counterpart. Rudimentary as it was by today's standards, some psychologists adopted it as a tool for treatment, much to the creator's concern and dismay. In response, Weizenbaum raised ethical concerns around our reflexive reliance and trust in automated systems that may appear to be objective and "intelligent," but are ultimately simplistic and prone to error.[146]

---

https://arxiv.org/abs/1705.05633

144 See Lakshika Balasuriya et al., "Finding Street Gang Members on Twitter," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016) http://knoesis.wright.edu/sites/default/files/ASONAM2016_GANG_MEMBER_IDENTIFICATION_LAKSHIKA.pdf

145 Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng and Max Kramer, "Moral Decision Making Frameworks for Artificial Intelligence," (Association for the Advancement of Artificial Intelligence, 2017).

146 Hans Pruijt, "Social Interaction With Computers: An Interpretation of Weizenbaum's ELIZA and Her Heritage," *Social science computer review* 24, No. 4 (2006): 516-523; Joseph Weizenbaum, *Computer power and human reason: From judgement to calculation,* Harmondsworth, UK: Penguin, 1984.

Currently there are heated debates about whether AI systems should be used in sensitive or high-stakes contexts, who gets to make these important decisions , and what the proper degree of human involvement should be in various types of decision-making.[147] These are ethical questions with a longstanding history. In examining these questions, we must also look at the power dynamics of current AI development and deployment – and the way in which decision-making, both by AI systems and the people who build them, is often obscured from public view and accountability practices.

Just in the last year, we've learned how Facebook mines user data to reveal teenagers' emotional state for advertisers, specifically targeting depressed teens.[148] Cambridge Analytica, a controversial data analytics firm that claims to be able to shift election results through micro-targeting, has been reported to have expansive *individual* profiles on 220 million adult Americans,[149] and fake news has been instrumented to gain traction within algorithmically filtered news feeds and search rankings in order to influence elections.[150] There are now multiple approaches for using machine learning techniques to synthesize audio- and video-realistic representations of public figures and news events.[151] Each of these examples shows how the interests of those deploying advanced data systems can overshadow the public interest, acting in ways contrary to individual autonomy and collective welfare, often without this being visible at all to those affected.[152]

AI Reflects Its Origins

The U.S. military has been one of the single most influential institutions in shaping modern AI, with DARPA's funding of AI being among the most visible.[153] Indeed, AI has historically been shaped largely by military goals, with its capabilities and incentives defined by military objectives and desires.[154] AI development continues to be supported by DARPA and other national defense agencies, particularly in the area of lethal autonomous weapons systems, as discussed above.

However, current research into AI technology is highly industry-driven, with proprietary systems supplementing military-funded classified systems and AI research increasingly

---

147 Sue Newell and Marco Marabelli, "Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'," *The Journal of Strategic Information Systems* 24, No. 1 (2015): 3-14.

148 Darren Davidson, "Facebook targets 'insecure' young people," The Australian, May 1, 2017.

149 Hannes Grassegger and Mikael Krogerus, "TheData That Turned the World Upside Down," *Motherboard*, 2017, https://publicpolicy.stanford.edu/news/data-turned-world-upside-down.

150 Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini and Filippo Menczer, "Hoaxy: A platform for tracking online misinformation," In *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 745-750. International World Wide Web Conferences Steering Committee, 2016.

151 Simon Adler, "Breaking News," Radiolab, July 27, 2017, http://www.radiolab.org/story/breaking-news/.

152 Kate Crawford and Meredith Whittaker, "Artificial Intelligence is Hard to See," *Medium*, September 11, 2016, : https://medium.com/@katecrawford/artificial-intelligence-is-hard-to-see-a71e74f386db.

153 Sidney G Reed, Richard H. Van Atta and Seymour J. Dietman, "DARPA Technical Accomplishments: An Historical Review of Selected DARPA Projects," Defense Advanced Research Projects Agency, Vol. 1, 1990.
Sidney G Reed, Richard H. Van Atta and Seymour J. Dietman, "DARPA Technical Accomplishments: An Historical Review of Selected DARPA Projects," Defense Advanced Research Projects Agency, Vol. 2, 1991.

154 Alex Roland and Philip Shiman, *Strategic computing: DARPA and the quest for machine intelligence, 1983-1993*, (MIT Press, 2002).

taking place in closed-door industry settings, often without peer review or oversight. Accordingly, user consent, privacy and transparency are often overlooked in favor of frictionless functionality that supports profit-driven business models based on aggregated data profiles.[155] While there are those advocating for clearer laws and policies, the ambiguous space in which information rights are governed does not clearly regulate in favor of individual control over personal technologies or online services.[156]

The make up of AI researchers – what is and is not considered "AI research" – also has a history which influences the current state of AI and its ethical parameters. Beginning with the Dartmouth Conference in 1956, AI researchers established a male-dominated, narrowly-defined community. The boundaries of participation in the AI community were relatively closed, and privileged mathematics, computer science and engineering over perspectives that would provide for a more rigorous discussion of AI's ethical implications.[157] Producing technologies that work within complex social realities and existing systems requires understanding social, legal and ethical contexts, which can only be done by incorporating diverse perspectives and disciplinary expertise.

Ethical Codes

While decades of AI research have cited Asimov's three laws of robotics,[158] and some applied AI systems have been designed to comply with biomedical ethics,[159] the tools that have been available to developers to contend with social and ethical questions have been relatively limited. Ethical codes are gradually being developed in the AI research space, as we discuss below, but they are necessarily incomplete: they will always need to evolve in ways that are sensitive to the rapidly changing contexts and conditions in which AI systems are deployed. These codes constitute one form of soft governance, where industry standards and technical practices serve as alternatives to more traditional "hard" forms of government regulation and legal oversight of AI. As AI systems are woven through a growing number of domains, the needs for such a contextually-anchored approach to ethics and governance only grows.[160]

Two related problems have emerged: there is no tracking of adherence to ethical guidelines or soft governance standards in the AI industry, and we have not developed ways to link the adherence to ethical guidelines to the ultimate impact of an AI systems in

---

[155] Elaine Sedenberg and Ann Lauren Hoffmann, "Recovering the History of Informed Consent for Data Science and Internet Industry Research Ethics" (September 12, 2016). Available at SSRN: https://ssrn.com/abstract=2837585.

[156] United States (2016) Executive Office of the President and Jason Furman, John P. Holdren, Cecilia Muñoz, Megan Smith and Jeffery Zients, "Artificial Intelligence, Automation, and the Economy," Technical report, National Science and Technology Council, Washington D.C. 20502, October 2016, https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF.

[157] Nils J Nilsson, *The quest for artificial intelligence*, (Cambridge University Press, 2009).

[158] Susan Leigh Anderson, "Asimov's "three laws of robotics" and machine metaethics," *Ai & Society* 22, No. 4 (2008): 477-493.

[159] Raymond Heatherly, "Privacy and security within biobanking: The role of information technology," *The Journal of Law, Medicine & Ethics* 44, No. 1 (2016): 156-160.

[160] Robert Rosenberger, "Phenomenological Approaches to Technological Ethics," *The Ethics of Technology: Methods and Approaches* (2017): 67.

the world.

Examples of intertwined practice and ethics can be found in the biomedical uses of AI. Bioethics already offers a series of standards, values and procedures,[161] along with enforcement and accountability mechanisms. But how these should apply to medical AI systems is often unclear, and researchers have been tracking the disparities.[162] This is also true of privacy requirements, which, given modern AI's capability to make very personal inferences given only limited data, are increasingly insufficient.[163] Where ethical standards aimed at protecting patient privacy have been proposed, some biomedical researchers have rejected them, seeing them as an impediment to innovation.[164]

A more intentional approach to ethics is needed, and some are working toward this. Teaching ethics to practitioners is one such example.[165] The Blue Sky Agenda for AI Education, a collection of ideas for ethics education in AI, seeks democratization of AI education and emphasizes inclusiveness in development toward the goal of respecting the values and rights of diverse populations.[166] But education is not enough. Opportunities must open up for ethics to be integrated in early stage design, and incentives for designing and implementing AI ethically must be built into the companies and institutions currently driving development.

Ethical values and norms around accountability,[167] social and political responsibility, inclusion and connectivity,[168] legibility and security and privacy[169] are embedded in every system via their default settings, whether intentionally or not.[170] Often, these invisibly-embedded values reflect the status quo, the context and interests of their developers, and matters of convenience and profit. Once set, these implicit values are hard to change for a variety of reasons,[171] even as they tend to shape the capabilities and roles of systems within various lived contexts.[172] Ethical codes should work to ensure that these

---

161 Stephen Brotherton, Audiey Kao, and B. J. Crigger, "Professing the values of medicine: the modernized AMA Code of Medical Ethics," *JAMA* 316, No. 10 (2016): 1041-1042.

162 Jake Metcalf and Kate Crawford, "Where are human subjects in big data research? The emerging ethics divide," *Big Data & Society* 3.1 (2016): http://journals.sagepub.com/doi/pdf/10.1177/2053951716650211

163 Stephen J Tipton, Sara Forkey and Young B Choi, "Toward Proper Authentication Methods in Electronic Medical Record Access Compliant to HIPAA and CIA Triangle," *Journal of medical systems* 40, No. 4 (2016): 1-8.

164 Wendy Lipworth and Renata Axler, "Towards a bioethics of innovation," *Journal of medical ethics* (2016): medethics-2015.

165 Judy Goldsmith and Emanuelle Burton, "Why Teaching Ethics to AI Practitioners Is Important," (2017).

166 Eric Eaton, Sven Koenig, Claudia Schulz, Francesco Maurelli, John Lee, Joshua Eckroth, Mark Crowley, Richard G. Freedman, Rogelio E. Cardona-Rivera, Tiago Machado and Tom Williams, "Blue Sky Ideas in Artificial Intelligence Education from the EAAI 2017 New and Future AI Educator Program," *arXiv preprint arXiv:1702.00137* (2017).

167 Wendell Wallach, *A Dangerous Master: How to keep technology from slipping beyond our control*, (Basic Books, 2014).

168 Anna Lauren Hoffmann, Nicholas Proferes and Michael Zimmer, "'Making the world more open and connected': Mark Zuckerberg and the discursive construction of Facebook and its users." *new media & society* (2016): 1461444816660784.

169 John Wilbanks, "Public domain, copyright licenses and the freedom to integrate science," *Public Communication of Science and Technology* 25 (2016): 11.

170 Ian Kerr, "The Devil is in the Defaults," *Critical Analysis of Law* 4, No. 1 (2017).

171 Dylan Wesley Mulvin, "Embedded dangers: The history of the year 2000 problem and the politics of technological repair," A*oIR Selected Papers of Internet Research* 6 (2017).

172 Narendra Kumar, Nidhi Kharkwal, Rashi Kohli and Shakeeluddin Choudhary, "Ethical aspects and future of artificial intelligence," (In *Innovation and Challenges in Cyber Security (ICICCS-INBUSH), 2016 International Conference on*, pp. 111-114), IEEE, 2016.

values are expressly designed into AI systems through processes of open and well-documented decision-making that center the populations who will be most affected.

While nascent, efforts to address these concerns have emerged in recent years. A series of White House reports under President Obama examined tensions between social interests and ethical values on one hand, and business and industry objectives on the other.[173] Recent soft governance efforts from IEEE,[174] The Future of LIfe Institute,[175] the ACM[176] and the Oxford Internet Institute[177] have produced principles and codes of ethics for AI. Perspectives from diverse industry and intellectual leaders are often reflected in these documents. While these are positive steps, they have real limitations. Key among these is that they share an assumption that industry will voluntarily begin to adopt their approaches. They rarely mention the power asymmetries that complicate and underlie terms like "social good," and the means by which such a term would be defined and measured. The codes are necessarily limited in what they address, how much insider information they have access to and what mechanisms would be used for monitoring and enforcement.[178] While these efforts set moral precedents and start conversations,[179] they provide little to help practitioners in navigating daily ethical problems in practice[180] or in diagnosing ethical harms,[181] and do little to directly change ethics in the design and use of AI.[182]

Challenges and Concerns Going Forward

Current framings of AI ethics are failing partly because they rely on individual responsibility, placing the onus of appropriate information flow with users and concentrating decision-making power in individual AI developers and designers.[183] In order to achieve ethical AI systems in which their wider implications are addressed, there must be

---

[173] United States (2016) Executive Office of the President and M. Holden, J.P. and Smith. "Preparing for the future of artificial intelligence," Technical report, National Science and Technology Council, Washington D.C. 20502, October 2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

[174] IEEE, "Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems," The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, December 13, 2016, http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf.

[175] "Asilomar AI Principles," 2017, https://futureoflife.org/ai-principles/.

[176] ACM, "Code 2018," version 2, 2017, https://ethics.acm.org/2018-code-draft-2/.

[177] Mike Wooldridge, Peter Millican, and Paula Boddington, "Towards a Code of Ethics for Artificial Intelligence Research," Oxford, 2017 https://www.cs.ox.ac.uk/efai/towards-a-code-of-ethics-for-artificial-intelligence/.

[178] Bo Brinkman, Catherine Flick, Don Gotterbarn, Keith Miller, Kate Vazansky and Marty J. Wolf, "Listening to professional voices: draft 2 of the ACM code of ethics and professional conduct," *Communications of the ACM* 60, No. 5 (2017): 105-111.

[179] Eugene Schlossberger, "Engineering Codes of Ethics and the Duty to Set a Moral Precedent," *Science and engineering ethics* 22, No. 5 (2016): 1333-1344.

[180] Stuart Ferguson, Clare Thornley and Forbes Gibb, "Beyond codes of ethics: how library and information professionals navigate ethical dilemmas in a complex and dynamic information environment," *International Journal of Information Management* 36, No. 4 (2016): 543-556.

[181] Christian Sandvig, Kevin Hamilton, Karrie Karahalios and Cedric Langbort, "Automation, Algorithms, and Politics| When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software," *International Journal of Communication* 10 (2016): 19.

[182] Mike Ananny, "Toward an ethics of algorithms: Convening, observation, probability, and timeliness," *Science, Technology, & Human Values* 41, No. 1 (2016): 93-117.

[183] Florencia Marotta-Wurgler, "Self-Regulation and Competition in Privacy Policies." *The Journal of Legal Studies* 45, No. S2 (2016): S13-S39.

institutional changes to hold power accountable.[184] Yet, there are obvious challenges to this approach, such as disagreement about the risks of AI,[185] the potential for greenwashing ethical AI as a superficial marketing strategy rather than a substantive commitment,[186] the practical challenges of stopping unethical AI research designed to privilege the interests of a few over the many[187] and the current economic system within which the incentives driving AI development are embedded. In addition, the effective invisibility of many of these systems to the people on whom they act, the obscurity of their algorithmic mechanisms, the ambiguity of their origins and their inescapable pervasiveness make public discourse difficult and opting-out impossible.[188] The responsibility to strive for better outcomes thus falls squarely on creators and regulators, who are only beginning to establish dialogue[189] even as there are few incentives for change and significant tension between ethics and "compliance."[190]

This brings us to the wider political landscape in which AI is being created in the U.S.: how will the Trump administration affect the use of these technologies? Prior to the election, over 100 technology sector leaders articulated their priorities: "freedom of expression, openness to newcomers, equality of opportunity, public investments in research and infrastructure and respect for the rule of law. We embrace an optimistic vision for a more inclusive country where American innovation continues to fuel opportunity, prosperity and leadership."[191] President Trump's policies do not reflect these priorities. Rather, there has been significant defunding of research, an increase in deportations, and heightened screening of personal communications and social media at national borders, among many other concerning policy shifts. Simply put, it does not appear that the current administration can be counted on to support the creation and adoption of ethical frameworks for AI.

---

184 Ira S. Rubinstein, "The Future of Self-Regulation is Co-Regulation" (October 5, 2016). The Cambridge Handbook of Consumer Privacy, From Cambridge University Press (Forthcoming). Available at SSRN: https://ssrn.com/abstract=2848513.

185 Vincent C Müller, ed. *Risks of artificial intelligence*. CRC Press, 2016.

186 Michael Stocker, "Decision-making: Be wary of 'ethical' artificial intelligence," *Nature* 540, No. 7634 (2016): 525-525.

187 Federico Pistono and Roman V. Yampolskiy, "Unethical Research: How to Create a Malevolent Artificial Intelligence," *arXiv preprint arXiv:1605.02817* (2016).

188 Jatin Borana, "Applications of Artificial Intelligence & Associated Technologies." *Proceeding of International Conference on Emerging Technologies in Engineering, Biomedical, Management and Science [ETEBMS-2016]* 5 (2016): 64-67; Ira S. Rubinstein, "Big Data: The End of Privacy or a New Beginning?." *International Data Privacy Law* (2013): ips036; Fred Turner. "Can we write a cultural history of the Internet? If so, how?." Internet Histories 1, No. 1-2 (2017): 39-46.

189 Kate Darling and Ryan Calo, "Introduction to Journal of Human-Robot Interaction Special Issue on Law and Policy," Journal of Human-Robot Interaction 5, No. 3 (2016): 1-2.

190 Kate Crawford and Ryan Calo, "There is a blind spot in AI research," *Nature* 538 (2016): 311-313; Gary E Merchant and Wendell Wallach, *Emerging Technologies: Ethics, Law and Governance,* (Ashgate Publishing, 2017).

191 Marvin Ammori, Adrian Aoun, Greg Badros, Clayton Banks, Phin Barnes, Niti Bashambu, et al., "An open letter from technology sector leaders on Donald Trump's candidacy for President," 2016, https://shift.newco.co/an-open-letter-from-technology-sector-leaders-on-donald-trumps-candidacy-for-president-5bf734c159e4.

Conclusion

AI systems are now being adopted across multiple sectors, and the social effects are already being felt: so far, the benefits and risks are unevenly distributed. Too often, those effects simply happen, without public understanding or deliberation, led by technology companies and governments that are yet to understand the broader implications of their technologies once they are released into complex social systems. We urgently need rigorous research that incorporates diverse disciplines and perspectives to help us measure and understand the short and long-term effects of AI across our core social and economic institutions.

Fortunately, more researchers are turning to these tasks all the time. But research is just the beginning. Advocates, members of affected communities and those with practical domain expertise should be included at the center of decision making around how AI is deployed, assessed and governed. Processes must be developed to accommodate and act on these perspectives, which are traditionally far removed from engineering and product development practices. There is a pressing need now to understand these technologies in the context of existing social systems, to connect technological development to social and political concerns, to develop ethical codes with force and accountability, to diversify the field of AI and to integrate diverse social scientific and humanistic research practices into the core of AI development. Only then can the AI industry ensure that its decisions and practices are sensitive to the complex social domains into which these technologies are rapidly moving.