


2021

AI and Inequality

Pauline Kim

Washington University in St. Louis School of Law, kim@wustl.edu

Follow this and additional works at: https://openscholarship.wustl.edu/law_scholarship

 Part of the [Civil Rights and Discrimination Commons](#), [Labor and Employment Law Commons](#), and the [Legal Studies Commons](#)

Repository Citation

Kim, Pauline, "AI and Inequality" (2021). *Scholarship@WashULaw*. 451.
https://openscholarship.wustl.edu/law_scholarship/451

This Article is brought to you for free and open access by the Law School at Washington University Open Scholarship. It has been accepted for inclusion in Scholarship@WashULaw by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

AI and Inequality

(forthcoming in *The Cambridge Handbook on Artificial Intelligence & The Law*,
Kristin Johnson & Carla Reyes, eds. (2022))

Pauline T. Kim

Daniel Noyes Kirby Professor of Law
Washington University School of Law

1. Introduction

Artificial Intelligence (AI) is everywhere in our economic and social lives. While applications like self-driving vehicles have captured popular imagination, predictive AI tools are already being used to make everyday decisions about people, in contexts ranging from criminal sentencing to employment to finance.¹ As these uses proliferate, observers argue that far from being infallible, AI tools pose significant risks of treating individuals unfairly. This chapter focuses on the broader social consequences of those risks—namely, that the use of predictive AI in social domains may contribute to growing inequality.

Many AI applications involve prediction, often concerning observable facts about the physical world—for example, will it rain tomorrow? Is this an image of an animal? When will this part on a machine fail? In contrast, the focus of this chapter is on predictions about human behavior. It examines AI that seeks to predict how specific persons will act in the future in order to make decisions about them. An algorithm might score individuals according to their likelihood of falling behind on future payments in order to decide who gets a loan. Or an AI model might predict which persons accused of a crime are more likely to be involved in criminal activity in the future to determine who should be released on bail. A prediction tool might decide which candidates offer the best hiring prospects for an employer. When thus used in social domains, AI determines who receives resources and opportunities.

Relying on AI tools in these settings is attractive because they seem to offer an evidence-based approach to making difficult decisions. The algorithm promises a more objective and accurate basis for acting—one that can avoid a human decision-maker’s implicit biases or cognitive limitations. Whether these advantages are actually realized, however, depends a great deal upon how exactly a model is constructed. Because of the complexity of prediction tasks, a model’s designers must make a myriad of choices when building it, each of which will shape the results it produces. Poor choices can increase error or cause biased outcomes.

When AI is used in social domains, the possibility of error and bias raises policy concerns. Numerous commenters have called out the risks that people will be unfairly denied access to

¹ See, e.g., Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871 (2016); Matthew T. Bodie, Miriam A. Cherry, Marcia L. McCormick & Jintong Tang, *The Law and Policy of People Analytics*, 88 U. COLO. L. REV. 961 (2017); Kristin Johnson, Frank Pasquale & Jennifer Chapman, *Artificial Intelligence, Machine Learning, and Bias in Finance: Toward Responsible Innovation*, 88 FORDHAM L. REV. 499 (2019).

essential resources like housing, jobs and credit. In addition to the impact on specific individuals, the growing use of predictive AI also threatens broader social harms. In particular, these technologies risk increasing inequality in two distinct ways—first, by reproducing or exacerbating the marginalization of historically disadvantaged groups, and second, by reinforcing the power hierarchies that contribute to economic inequality.

These points about the potential impacts of predictive AI on inequality apply across a number of domains, although many issues will arise that are specific to a particular context. In order to make the discussion more concrete, this Chapter closely analyzes the employment relationship to illustrate the challenges. It begins by offering some foundational definitions to frame the discussion, then explores complications that arise when AI technologies are used to predict human behavior. It next undertakes a closer examination of these issues in the employment context, explaining how firms increasingly rely on AI-powered tools to recruit, hire and manage their workforces, and the risks that these developments will increase inequality along the lines of race, sex and other protected characteristics, as well as worsening class inequalities by concentrating managerial power over workers. Finally, it considers the legal and policy implications of these developments. While the law likely restrains discriminatory applications of AI to some extent, it currently offers few tools for addressing the growing concentration of managerial power.

2. Preliminary Definitions

Because the term “artificial intelligence” is used loosely to refer to many different things, defining the scope of the inquiry here is important. AI is perhaps best understood as an umbrella term that encompasses many different computational techniques applied in a wide range of settings.² These techniques have a number of features in common—for example, the reliance on large amounts of data and rapid computing power—and they generally have the objective of automating or replicating some aspect of human cognition or behavior. The applications, however, are wildly diverse, ranging from self-driving cars and robots, to smart home systems, to online book recommendations, dating platforms and fraud detection systems.

This chapter focuses on predictive AI that is used in social domains—areas like employment, housing, education, financial services, insurance, and criminal law enforcement. These applications entail “disembodied AI, which acquires, processes, and outputs information as data” rather than acting on the physical world the way robots do.³ Some forms of disembodied AI aim to describe or anticipate phenomena that unfold as a result of physical properties, for example, predicting when parts in a machine are at risk of failure, or whether a radiological scan shows signs of abnormality. In social domains, however, AI is used to make predictions about human behavior in order to make decisions about who gets access to certain resources, or how benefits or burdens will be distributed. As examples, predictive AI is currently used to inform decisions about where policing efforts are focused, which criminal defendants are released pretrial, and who is hired or given a loan.

² Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C.DAVIS L. REV. 399, 404 (2017).

³ *Id.* at 407.

Data—in the sense of information about people—has always been used to make predictions about them, so what is distinctive about predictive AI? Drawing a sharp line demarcating AI tools may not be possible,⁴ but they are arguably distinct because they operate on vastly greater amounts of data, across larger populations and in less visible ways than earlier methods of making decisions. These developments are the result of the growing availability of personal data, advances in computational techniques, and greatly increased computing power. Although prior decision-making practices raised concerns about fairness, the enormously expanded scale and efficiency of AI tools raise heightened concerns.

The next section explores the particular challenges involved when making predictions about human behavior. How these challenges are met in turn shapes the outputs of AI tools in ways that can have significant societal consequences.

3. Predicting Human Behavior

To see the challenges involved in making predictions about individual people, consider first a relatively pedestrian, straightforward example of an AI system—the junk email filter. Anyone with an email account is plagued by unwanted emails. Today, much of that spam is filtered out by software that analyzes the characteristics of incoming email to identify unwanted messages. A simple rule-based algorithm cannot succeed at this task, because spammers are constantly changing their tactics and can easily evade any static system. Effective filters must rely on machine learning techniques in order to constantly learn and update their models as new information becomes available. These AI tools leverage vast amounts of data that is constantly being collected about threats posed by phishing schemes and spammers, as well as information about how the individual user responds to particular messages.

Spam filters are highly effective at automating a task to address a specific problem. They replicate human judgments, learn over time, and pose minimal societal risks. The success of AI in this context rests on a number of factors. First, spam filters are asked to perform a relatively straightforward task: decide whether a particular email is junk or not. The decision involves a simple binary choice, and there is a correct outcome. If the algorithm makes a mistake in classifying a message, the error can be ascertained and that information used to retrain and refine the algorithm. Given readily available, unbiased data, the algorithm should be able to adjust to changes in the environment and to improve over time. A user may occasionally have to delete an email that got through the filter or search through a junk folder to retrieve an important message, but these errors impose relatively small costs.

In addition, in the case of commercial email providers, the interests of the entity deploying the algorithm and the persons subject to its application are generally aligned.⁵ Even though the email application is harvesting information about the behavioral patterns of its subscribers, those users—as data subjects—share an interest in reducing the amount of junk mail. Because data

⁴ A paradox is that new technologies are initially considered cutting edge AI, but as their use becomes commonplace they come to be viewed as routine data processing. *See* Preparing for the Future of Artificial Intelligence, Executive Office of the President, National Science and Technology Council, at 7 (2016).

⁵ Of course, spam filtering software is adverse to the interests of spammers and phishers, but their relationships to the email application and its users are expressly adversarial.

about their behavior makes the filtering software more accurate, that use promotes their interests and is consistent with their expectations. To that extent, it satisfies various notions of privacy—such as displaying contextual integrity⁶ and being in accord with trust norms in information relationships.⁷

Note that this need not always be the case. A state-provided email system might use AI to filter out messages that are critical of the government, or a firm might use screening software to analyze whether employees are engaged in union organizing or whistleblowing activities. In those situations, not only are the interests of the provider and the users not aligned, but a power imbalance allows the state or the firm to act in conflict with the interests of the data subjects, a point to which I will return below.

Putting aside nefarious purposes, the spam filter illustrates the promise of AI. It provides accurate, neutral tools that can automate decisions, saving time and human effort. Although these tools are not perfect—some junk email still slips through and relevant messages are occasionally screened out—they offer significant advantages to users with little downside risk.

Applying AI to human behavior, however, is more perilous. Making accurate predictions is difficult, and because they will affect decisions about people’s lives, a great deal more is at stake. The relevant behaviors are both complex and difficult to define precisely. For example, recidivism tools seek to predict whether a criminal defendant poses an ongoing threat to society, but how should that be measured?⁸ Should the target of prediction be subsequent arrests, criminal charges, or only actual guilty pleas and convictions? Should any offense count, even traffic violations, or only violent felonies?

Similarly, job applicants do not neatly fall into a binary category of good employee/bad employee, in the way that email is either junk or not. There are better and worse employees, and they can be arrayed along any number of dimensions (reliability, teamwork, loyalty, skill, creativity, leadership, etc.), each of which represents a continuum, not an either/or quality. None of those characteristics can be directly observed or objectively measured. Instead, some proxy must be chosen to capture the targeted qualities. There is no correct solution regarding how to define the target variable or what factors should be used to measure it, and yet, these choices can have significant distributional consequences.

Even after a choosing a definition for the targeted behavior, it is impossible to know if a prediction is correct or not in any given instance. Unlike junk mail, where it is easy to ascertain whether a correct decision was made, the “true” prediction about a particular person is not directly observable. The only way to know if that individual will actually commit another crime or default on a loan is to wait and observe what actually happens over time.

Because the accuracy of predictions about humans is not easily observed, there is a greater risk of systemic biases. In contexts like the spam filter, errors can be detected and this information

⁶ HELEN NISSENBAUM, *PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE* (2009).

⁷ Neil Richards & Woodrow Hartzog, *Taking Trust Seriously in Privacy Law*, 19 *STAN. TECH. L. REV.* 431 (2016); ARI EZRA WALDMAN, *PRIVACY AS TRUST: INFORMATION PRIVACY FOR AN INFORMATION AGE* (2018).

⁸ *See, e.g.*, Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 *EMORY L.J.* 59, 75–78 (2017).

used to update and retrain the algorithm. In social domains, however, all outcomes will not be observed because decision-makers are relying on the outputs to make choices. Persons rated as posing high risks of re-offending will not be released from custody; those predicted to default will not be given the loan. There will be no opportunity to observe whether those individuals actually avoid further entanglement with law enforcement or pay off their mortgages. False positives will be observed, while false negatives will not. This one-sided censoring of information will limit how the model learns over time.

Of course, the same is true of traditional decision-making methods. Human judges or loan officers might be wrong in their decisions as well, and when they fail to release someone who would not reoffend or refuse a loan to someone who would have repaid, these “false negatives” will not be observed. In this aspect, predictive AI is no different than human judgments about how individuals will behave in the future. However, it also means that one of the claimed benefits of AI—its ability to learn over time—is far more limited when used to make decisions about people than in contexts like the junk mail filter.

Another limitation of predictive AI is that it is generally not theory-driven. Machine learning models do not necessarily rely on characteristics or attributes that are causally related to the predicted outcome. Instead, they simply mine the available data to uncover patterns. This is often touted as one of the advantages of machine learning techniques. By applying advanced computational methods, these tools can analyze immense amounts of data and discover complex relationships that are beyond human capacities to perceive. This type of data-mining makes predictions based on patterns that have been observed in the past. The usefulness of this approach, however, depends upon whether those patterns will hold true in the future. If past behavior was influenced by the environment, the predictions may not be valid when conditions change. These models often treat people as having fixed traits that determine their behavior, rather than paying attention to how the environment can influence and shape how they act.

Another relevant consideration is the impact of an erroneous decision. Mistakenly classified email is annoying and inefficient; mistaken predictions about people may lead to the loss of significant economic opportunities for some individuals. Moreover, unlike spam filters, which face no limit on the number of messages that can be approved, many algorithms applied in social domains operate under conditions of scarcity. For example, employers typically can hire only a fraction of available job-seekers and so a screening or hiring algorithm must make comparative judgments about applicants. As a result, job candidates will be impacted not only if the algorithm misjudges them in the abstract, but also if it makes errors in assessing them relative to others.

Persons who are adversely affected might have cause for complaint, and hence, some legal regimes give individuals the right to disclosures about the process and the data on which is relied, and to contest the decision.⁹ If errors are distributed across individuals somewhat randomly, these types of rights may be sufficient to provide recourse to those who have been

⁹ See, e.g., General Data Protection Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) [hereinafter GDPR].

affected. However, a model may also be systematically wrong in a way that has broad social consequences beyond the individuals affected. Individual data protection rights cannot fully address these concerns. If an AI tool results in mortgage denials concentrated in a certain neighborhood, it is not just the individual borrower, but a whole community, that will suffer. Moreover, when systematic effects occur, human behavior is likely to change in response, which can cause a feedback loop, further exacerbating the negative effects. In a neighborhood suffering from disinvestment, for example, people who have some social or financial capital may choose to exit altogether, causing conditions to further deteriorate.

Finally, in social domains, the interests of the entities that employ AI tools are typically not aligned with the interests of the people who are subject to them. As a result, there is no assurance that the tools will be designed in a way that minimizes potential harms. Designers build screening tools to serve those who will use them, such as courts, banks and employers. Their goals will often be at odds with the interests of the data subjects—the criminal defendants, loan applicants and workers—to whom the algorithms will be applied. As explained above, the definition of the relevant behavior and how it is measured is very much up for grabs and will have different distributional consequences. In making these types of choices, model designers will focus on optimizing the interests of their clients, rather than the humans to whom their models will be applied.

In sum, when predictive AI is used in social domains, the nature of the problem it is trying to solve is quite different from more mechanical applications like the spam filter. The appropriate definition of the targeted behavior is inherently subjective and cannot be objectively measured; not all errors will be observed, limiting the possibility of learning; systematic errors can cause broader social impacts; and the interests of the designers and the data subjects are likely to conflict. These characteristics of the decision environment mean that the consequences of using AI are likely to be much broader than the immediate impact on an individual.

The next section explores these issues in the context of work relationships, illustrating how these tools risk increasing inequality in two senses. First, if poorly designed, these tools can encode discriminatory bias, reducing equality of opportunity across demographic groups. They may reflect and reproduce historical patterns of disadvantage faced by racial minorities, women and other marginalized groups when distributing opportunities. Second, when deployed in the context of an unequal relationship, AI tools are likely to strengthen the power of the dominant party, further reinforcing the inequality of that relationship.

4. The Social Consequences of Predictive AI

a. The Employment Context

To illustrate how predictive AI can contribute to inequality, this section examines in greater depth the use of these technologies in the employment context. Vendors now offer employers an array of tools to recruit, track, sort and screen job applicants, as well as to monitor and evaluate

employees after they are hired.¹⁰ The promise of AI is that in addition to enhancing efficiency, it can provide objective, neutral tools that avoid human biases, resulting in more objective and more accurate personnel decisions. Because of the complexity of predicting human behavior, however, the actual impact of using these technologies in the workplace depends upon how they are developed and deployed.

Miranda Bogen and Aaron Rieke have comprehensively catalogued the ways in which algorithms are currently used in the hiring context.¹¹ They describe the process as a funnel, in which a series of decisions are made that progressively narrow the pool of potential hires, leading ultimately to a job offer or a rejection. The funnel involves several stages—sourcing, screening, interviewing, and selection—with employers currently using predictive algorithms at each stage to eliminate some candidates from the pool. AI may determine who learns about a job opportunity and who actually applies at the sourcing stage. Screening tools analyze candidate profiles in order to eliminate some from further consideration, and as the funnel narrows, some automated tools can inform or even make decisions about who will receive an offer.

A number of vendors now offer AI tools that purport to identify the best candidates for available jobs.¹² These vendors rely on a variety of data types to make their predictions. Some analyze data directly provided by applicants, such as resumes and responses to specific questions. Others harvest additional data about an applicant’s online interactions. One company records video interviews and analyzes applicants’ facial expression and speech patterns in order to draw inferences about their personality traits. Another engages applicants with video games, analyzing their behaviors when performing each task to predict personal characteristics like risk-taking, perseverance and conscientiousness. These algorithms rely on different data, but share an underlying premise—namely, that given enough information about individual applicants, AI can accurately predict their future job performance.

Designing an algorithm to make predictions about job performance requires considerable amounts of information about a large pool of workers to use as training data. Often that training data is collected from current employees. For example, in order to build an algorithm to predict which applicants will be “top performers,” the model builders will first need to have data on a great variety of characteristics of current employees, as well as information about who among them are considered top performers and who are less successful. The algorithm analyzes that data using machine learning techniques to learn patterns, such as which resume items, or facial expressions, or game play, or combinations thereof, distinguish the top performers from the rest. The resulting model is then applied to applicants using data about their attributes to predict who among them will succeed—and screening out the rest.

¹⁰ See Lisa Kresge, *Data and Algorithms in the Workplace: A Primer on New Technologies*, U.C. Berkeley Labor Center (November 2020).

¹¹ MIRANDA BOGEN & AARON RIEKE, *HELP WANTED: AN EXAMINATION OF HIRING ALGORITHMS, EQUITY, AND BIAS* (2018).

¹² Manish Raghavan, Solon Barocas, Jon Kleinberg & Karen Levy, *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 469 (ACM Jan. 2020).

In addition to the use of AI in the screening and selection process, employers are also increasingly relying on these tools to manage their workforces. New technologies allow firms to more closely monitor the activities of workers throughout the day, and sometimes even off-duty. And the data collected through these monitoring tools can be used to automate firms' management functions, by algorithmically directing work efforts and evaluating performance.¹³

b. Race and Sex Inequality

When firms rely on predictive AI to make personnel decisions, these tools can reproduce or worsen existing inequality along the lines of race, sex and other protected attributes.¹⁴ Labor markets have long been plagued by race- and sex-based discrimination, which has resulted in patterns of occupational segregation, lower pay and reduced advancement opportunities for marginalized groups.¹⁵ One of the promises of algorithmic decision-making tools is that it can avoid the human prejudice and bias that produced workplace inequalities in the past. Unfortunately, however, depending upon how it is constructed, predictive AI can also reflect or even exacerbate existing race- and gender-based biases.

This risk of reproducing bias is present in the very structure of today's labor markets.¹⁶ Employers increasingly rely on online platforms like Facebook and Google to advertise job openings and reach potential hires. Other platforms such as ZipRecruiter and Indeed are specially aimed at matching candidates with available jobs. These new labor market intermediaries rely on AI technology to determine who sees what ads and which applicants are recommended for particular job openings. While the algorithms are not public, a number of studies have documented race and gender biases in how information relevant to employment opportunities are distributed online.

Lambrecht and Tucker found that ads promoting STEM (Science, Technology, Engineering and Math) careers were shown to significantly more men than women.¹⁷ Similarly, Ali, et al.¹⁸ and Sapiezynski, et al.¹⁹ have documented that employment ads posted on Facebook were delivered to race- and gender-skewed audiences in ways that reflect stereotyped notions about who does what kinds of jobs. For example, ad for jobs as AI researchers or truck drivers were overwhelmingly targeted at male users, while ads for a janitor position were served to a largely black and female audience. Importantly, the advertisers specifically targeted race and

¹³ Jeremias Adams-Prassl, *What if your boss was an algorithm? Economic Incentives, Legal Challenges, and the Rise of Artificial Intelligence at Work*, 41 COMP. LAB. L. & POL'Y J. 123 (2019).

¹⁴ See Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017).

¹⁵ KEVIN STAINBACK & DONALD TOMASKOVIC-DEVEY, DOCUMENTING DESEGREGATION: RACIAL AND GENDER SEGREGATION IN PRIVATE SECTOR EMPLOYMENT SINCE THE CIVIL RIGHTS ACT (2012).

¹⁶ Pauline T Kim, *Manipulating Opportunity*, 106 VA. L. REV. 69 (2020).

¹⁷ Anja Lambrecht & Catherine Tucker, *Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads*, 65 MANAGEMENT SCIENCE 2966 (2019).

¹⁸ Muhammad Ali et al., *Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes*, 3 PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION 1 (2019).

¹⁹ Piotr Sapiezynski et al., *Algorithms that "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences*, ARXIV:1912.07579 [CS] (2019).

gender neutral audiences when placing these ads. Thus, it was the operation of the platforms' algorithms that produced the demographically skewed delivery. In all likelihood, the algorithms, attempting to predict which users were most likely to respond, ended up reflecting and reproducing existing patterns of occupational segregation.

A study by Sweeney offers another example of how algorithmic bias can worsen racial equality in labor markets.²⁰ Sweeney found that advertisements for criminal background checks appeared more often next to Google searches for African-American associated names, suggesting that such records exist, than searches for Caucasian-associated names. Those results likely reflected past patterns of users' search behavior, rather than discriminatory intent on the part of Google's programmers. Regardless of the reasons, however, the unequal display of the ads could nudge employers to scrutinize African-American applicants more closely than white applicants. And if, as a result of the nudge, an employer conducts more criminal background checks for African-American applicants than white applicants, then it will learn of more instances of involvement with the criminal enforcement system where it makes these inquiries, further reinforcing a cycle of bias.

When employers rely on predictive AI to screen, sort and hire applicants, similar risks arise. In building a predictive model, the AI's designers must make a series of choices,²¹ each of which could affect the distribution of outcomes across groups. The first choice involves defining the target variable—that is, deciding what behavior or characteristics the model will predict. As explained above, this step is complicated, because the true target—identifying those who will be the “best” employees—cannot be easily observed and quantified. When the designers make their target more concrete, they may implicitly build in systematic biases. For example, an employer might aim to identify not just workers with relevant job skills, but also those willing to work long hours on short notice. If the target variable implicitly rests on the notion of an ideal worker built around male norms of constant availability for work,²² the algorithm will tend to screen out those with family care responsibilities—principally female candidates. Similarly, because women and individuals with disabilities are more likely to take breaks from paid employment, they may be disadvantaged under a model that selects for those most likely to continue working without interruption.

In addition to defining the target variable, the designers must decide what data to use to train the model. This involves not just finding an appropriate dataset but ensuring that it contains relevant class labels. Once again, there is a risk that this process could incorporate systematic bias. Consider, for example, an employer that wants to identify applicants who are most similar to its current high-performing employees. In order to build such a model, it needs data that correctly labels which workers fall into the category of “high-performing.” The employer might identify high performers by relying on supervisors' judgments. If, however, those ratings are affected by cognitive or implicit biases against marginalized groups, the labels applied, and, hence, the model's predictions, will reproduce those biases. Class labels may also reflect

²⁰ Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 COMMUNICATIONS OF THE ACM 44 (2013).

²¹ David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn about Machine Learning*, 51 U.C.DAVIS L. REV. 653 (2017).

²² JOAN C. WILLIAMS, UNBENDING GENDER: WHY FAMILY AND WORK CONFLICT AND WHAT TO DO ABOUT IT (2000).

structural inequalities in the workplace.²³ For example, racially or sexually hostile work environments, or limited to access to mentorship, can hamper the job performance of women and workers of color. Predicting future potential based on measures of past performance that are affected by unequal conditions may unfairly reinforce those structural disadvantages.

The selection of the training dataset is also consequential, because the model's predictions turn on the examples to which it has been exposed. If certain groups are under-represented in the training data, the resulting model may be less accurate when applied to members of those groups, or may learn to exclude them altogether. Amazon reportedly faced such a problem when it tried to create an AI tool to identify promising candidates to hire as software developers.²⁴ Because the model was trained using current employees in that position—who were overwhelmingly male—it “learned” to systematically downgrade the resumes of women regardless of their qualifications for the job. Factual errors or missing data may also bias predictions, and some evidence suggests that these data problems are more common for demographic groups that are further from society's mainstream.²⁵

When an AI tool predicts different rates of success along race or gender lines, it is possible that it is capturing real differences between groups that are relevant to their ability to do the job. Lack of equal access to education, for example, could produce genuine differences in acquiring job-relevant skills. However, because predictive AI tools typically entail pattern-finding rather than directly measuring job-related skills and abilities, it often will be impossible to know whether disparate outcomes across groups reflect relevant differences in skills or result from the types of bias discussed above. The risk is that outcomes will be perceived as “neutral” and “accurate” because they reproduce familiar patterns, even if those patterns were shaped by human biases. And, as discussed above, if some individuals capable of doing the job have been erroneously screened out, those errors will not be observed and there will be no opportunity to correct any mistaken assumptions that may underlie the model.

Applicants who are erroneously classified may have cause for complaint. However, any system of prediction or selection will inevitably make some mistakes. Because they can never be entirely eliminated, what matters from an individual perspective is that each person was treated fairly in the decision process. When aggregated, though, these errors can have broader social consequences. In particular, when AI models systematically disfavor groups that have historically been marginalized in the labor market, they reproduce those past patterns of disadvantage and exclusion. Such a result not only runs counter to societal commitments to equal opportunity, it risks further entrenching discriminatory patterns through feedback effects.²⁶ Members of groups that are systematically disadvantaged may perceive fewer opportunities and reduce investments in their own human capital. Reduced investments in turn may further decrease expectations for the group, creating a negative feedback cycle.

²³ See Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 195 (2017).

²⁴ Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*, REUTERS (Oct. 10, 2018), <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

²⁵ Kim, *supra* note 14, at 885–86.

²⁶ *Id.* at 882–83.

It is sometimes argued in defense of algorithmic hiring tools that the alternative—subjective human judgments—are likely to be even more biased. Certainly, a great deal of social science evidence indicates that racial and gender biases affect human decisions,²⁷ and AI tools may well be less discriminatory in some situations. However, there are at least two reasons to be especially concerned about machine bias. First, the technical nature of these tools may convey a false sense of precision and objectivity, lending a sense of inevitability to outcomes that in fact rest of human choices. And second, because of their efficiency, these tools have the potential to scale up their impact more rapidly than biased human managers.

In any case, the fact that humans can be biased is not a reason to overlook or ignore algorithmic bias. Precisely because humans have biases, policy-makers should be attentive to the ways those biases might become embedded in algorithms through the choices made in constructing the model. On the other hand, the possibility that AI models can be discriminatory does not argue for prohibiting the use of the technologies entirely. Rather, the goal should be to create an appropriate legal framework that incentivizes firms to closely scrutinize these tools for bias, but without discouraging equality-promoting uses.

The reminder that humans also make biased decisions is helpful in another way. As awareness has grown that algorithms can be discriminatory, increased effort is now directed at devising technical solutions—strategies that can be written into code to ensure that a model is not biased. The limitation of this approach is that often times the source of bias lies outside the code and cannot be fixed by tweaking it.²⁸ As an example, consider a workplace in which women are routinely excluded from key opportunities for professional development and are subjected to demeaning and harassing behavior by supervisors and co-workers. As a result, their performance suffers, which causes a model to predict that they are not good prospects for promotion. No amount of tweaking of the code will successfully remove the discriminatory effects of those workplace practices. The only effective solution would be to address the source of discrimination directly—by ceasing the harassing behavior and providing development opportunities on an equal basis. The recognition that algorithms can be biased should not cause us to lose sight of discriminatory practices that are occurring in the real world and must be addressed directly.

c. Class Inequality

Separate from the risk of reinforcing race and gender inequalities, reliance on predictive AI in the workplace also threatens to worsen class inequality. A great deal of attention has focused on whether AI-driven robots will displace human labor on an unprecedented scale—a matter of ongoing debate.²⁹ While this question has obvious import given concerns about growing

²⁷ For reviews of portions of this vast literature, see Marianne Bertrand & Esther Duflo, *Field experiments on discrimination*, in 1 HANDBOOK OF FIELD EXPERIMENTS 309 (Abhijit Vinayak Banerjee & Esther Duflo, eds.) (2017); Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CALIF. L. REV. 945 (2006); Jerry Kang & Kristin Lane, *Seeing Through Colorblindness: Implicit Bias and the Law*, 58 UCLA L. REV. 465 (2010); Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN.L.REV. 1161, 1186-87 (1995).

²⁸ Kim, *supra* note 23, at 196.

²⁹ CYNTHIA ESTLUND, *AUTOMATION ANXIETY: WHY AND HOW TO SAVE WORK* (2021); ERIK BRYNJOLFSSON & ANDREW MCAFEE, *THE SECOND MACHINE AGE: WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES* (2014).

inequality, there is a less-noticed consequence of utilizing AI in the workplace—namely, the increase in managerial power. This effect is independent of the risks of race and gender bias. A model might produce positive outcomes at equal rates for different racial and gender groups, but nevertheless weaken workers’ interests vis à vis their employer.

AI tools are deployed at the behest of employers, and so designers focus on optimizing the firms’ interests. As a result, their choices in building the model will generally advance managerial control in ways that reinforce existing power hierarchies. Consider, again, an applicant screening tool. As discussed earlier, a key question is how to define a “good employee.” Because no objective measure exists, a model’s designers will reference the needs of the employer to determine the target of prediction. Perhaps the employer has faced frequent employee turnover and wants to prioritize hiring employees who will stay on the job for a long time. But perhaps those high turnover rates are due to substandard working conditions and low pay. Optimizing “likelihood of staying on the job” may mean the algorithm in effect selects for willingness to work under poor conditions without complaint. It may thus end up screening applicants in a way that undermines workers’ collective interests or even their legal rights.³⁰

This outcome need not be explicit or even conscious. An employer may articulate a goal of reducing workforce turnover. The designer trains the algorithm to identify the personal characteristics of workers who have been on the job the longest and to select applicants who share those characteristics. The pool of workers with the greatest longevity, however does not simply reflect their skills and work performance. It is also shaped by the employers’ past practices, including its hiring and firing decisions, the working conditions it maintained, and how it has responded to worker complaints or collective activities. If the employer has engaged in retaliation, then training the algorithm with this sample of workers may implicitly screen out those who are more likely to exercise their legal rights by organizing co-workers or complaining to a government agency. As a result, the AI tool may indirectly undermine collective labor rights or enforcement of minimum standards regulation. Rather than incentivizing the employer to provide more attractive working conditions to reduce turnover, the availability of AI tools instead allows it to select for more quiescent workers.

Other uses of AI beyond the hiring stage can also undermine conditions for workers. Retail firms have increasingly relied on smart scheduling software to determine their employees’ hours. By collecting data on customer traffic, this software can more precisely predict how many workers are needed at any given time and adjust workers’ schedules accordingly. Because foot traffic in retail tends to be concentrated in certain time blocks—over the lunch period and after working hours, for example—the effect of this software has been to reduce both the hours of work available to individual workers as well as the predictability of their schedules.

This software lowers labor costs for the firm, but at the same time it imposes significant costs on individual workers in the form of reduced hours and pay, and greater scheduling instability. Workers subjected to “just-in-time” scheduling often find it difficult to arrange for child care, or

³⁰ Nathan Newman, *Reengineering Workplace Bargaining: How Big Data Drives Lower Wages and How Reframing Labor Law Can Restore Information Equality in the Workplace*, 85 U. CIN. L. REV. 693 (2017).

to improve their situation by taking classes or supplementing their income with other work.³¹ If workers' interests shaped the goals of scheduling software, it would look quite different. It would instead optimize scheduling predictability in order to provide stable incomes and facilitate balancing family obligations, schooling and other work opportunities. Thus, *who* gets to define the problem that AI tools are supposed to solve matters quite a bit. Because in today's world employers pay for and deploy AI, these tools are built to meet their needs, often at the expense of worker interests.

AI also gives firms more control over when and how work is performed. Companies can utilize a wide variety of technologies, such as GPS locators, activity trackers, socio-metric badges, keystroke monitoring, voice analysis and facial recognition software, to collect highly granular data about workers' whereabouts, activities, and social interactions. The ability to collect immense amounts of data about individual workers has fueled two distinct trends.

First, firms are severing the employment relationship altogether, while relying on AI to maintain control over workers and the value created by their labor. These technologies have enabled the gig economy to flourish—perhaps best exemplified by Uber offering transportation services on a massive scale while employing no drivers.³² The second trend entails increasing levels of scrutiny and control over employees within the firm. For example, Amazon reportedly tracks the movements of its warehouse workers to monitor their productivity, and in some instances has used this information to automate dismissals.³³ The pressure to meet aggressive performance standards can lead workers to skip breaks and risk injury. Other firms are exploring the use of AI to predict workers' emotional states based on voice or facial analysis in order to more finely calibrate work assignments and monitor performance. And constant surveillance of workers' interactions can chill efforts to organize unions or engage in other forms of collective activity.

Whether work is performed within or outside the firm, the overall impact of these technologies is to reinforce the power hierarchies inherent in these relationships by concentrating more control in management. In order to use AI tools effectively, data is required, and the more comprehensive that data, the more accurate the predictions will be. Firms thus have incentives to collect yet more data about their workers in order to further improve these tools. And as monitoring devices increase in sophistication and efficiency, that data collection is likely to become still more all-encompassing, further increasing the informational asymmetry, and thus the power imbalance, between firms and workers.

³¹ See, e.g., Golden, Lonnie, *Irregular Work Scheduling and Its Consequences* (April 9, 2015). Economic Policy Institute Briefing Paper No. 394, Available at SSRN: <https://ssrn.com/abstract=2597172> or <http://dx.doi.org/10.2139/ssrn.2597172>

³² Alex Rosenblat & Luke Stark, *Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers*, 10 INT'L. J. COMM. 3758 (2016).

³³ Colin Lecher, *How Amazon Automatically Tracks and Fires Warehouse Workers for 'Productivity,'* THE VERGE, <https://www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations>.

Note that this dynamic is not intrinsic to the technology. The objectives of AI tools could be defined in reference to workers' interests, and as suggested above, if that were the case, their design and operation, and therefore their social effects, would be quite different. In fact, however, these tools primarily serve the interests of firms. This outcome reflects how power is distributed in labor markets and who has the resources to leverage these new tools to advantage. In a capitalist system, the law grants employers the authority to determine how to run the workplace, and firms, rather than workers, have access to the capital necessary to invest in these new technologies.

In recent decades, class inequality has grown starker, heightening concerns about undermining social cohesion and stability.³⁴ Many factors contribute to this growing inequality, but AI tools that increase employer power over workers will likely only worsen that trend.

* * * * *

The discussion above focused on the employment context, but the risks that predictive AI poses to equality are present across a number of domains. Studies have documented numerous other examples of algorithms that produce racial differences in outcomes, such as when predicting recidivism risk for criminal defendants,³⁵ or pricing mortgages.³⁶ Other work has shown how an algorithm directed greater health care resources to white patients as compared with black patients at the same level of medical need.³⁷

Predictive AI can also worsen economic inequality in domains beyond employment. Banks, for example, can use better predictions about vulnerable borrowers to engage in predatory pricing,³⁸ and firms can leverage large amounts of individual data and AI-driven behavioral insights to extract more consumer surplus.³⁹ Troubling practices like these are not new; however, the efficiency of data-driven tools enables firms to scale them up with potentially much broader consequences. Because the unrestrained use of predictive AI to make socially consequential decisions risks worsening inequality across a number of dimensions, these tools warrant scrutiny by policy-makers.

5. Implications for Law and Policy

This section considers the extent to which existing law applies when predictive AI is used in social domains and how the law might need to be reformed in light of the risks of increasing

³⁴ THOMAS PIKETTY, *CAPITAL IN THE TWENTY-FIRST CENTURY* (2017).

³⁵ Julia Angwin et al., *Machine Bias*, PROPUBLICA (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=Gg58888u2U5db3W3CsuKrD0LD_VQJReQ.

³⁶ See Robert P. Barlett, Adair Morse, Richard H. Stanton & Nancy E. Wallace, *Consumer-Lending Discrimination in the FinTech Era* (Nov. 2019) NBER Working Paper No. w25943, Available at SSRN: <https://ssrn.com/abstract=3491267> (finding racial/ethnic discrimination in pricing of mortgages by fintech algorithms, although to a lesser degree than by face-to-face lenders).

³⁷ Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447 (2019).

³⁸ Johnson et al., *supra* note 1, at 511, 517.

³⁹ See, e.g., Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995 (2014); JULIE E. COHEN, *BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM* (2019).

inequality. A first important point is that any regulation should be domain-specific. The particular risks posed by algorithmic tools and their social impacts will differ depending upon whether they are used to allocate housing, employment, credit, or policing efforts. When building a predictive model, designers often face difficult tradeoffs, and how those tradeoffs should be weighed depends upon the effects of different types of errors in the specific context. Blanket rules that regulate AI generically will either be too blunt to be effective or so specific that they produce unintended harmful consequences in certain situations. Developing appropriate regulation requires not only technical knowledge, but also deep domain expertise to understand the particular setting. With that caveat, this section addresses the application and limits of current law, once again using the employment context as an illustrative example.

Since at least the mid-twentieth century, law in the United States has been concerned with preventing and redressing discrimination because of ascriptive characteristics like race, ethnicity or gender.⁴⁰ More recently, protections have been extended to persons with disabilities who were historically marginalized by structures that limited their access to mainstream society.⁴¹ A number of statutes prohibit discrimination in areas such as housing, employment, and financial services,⁴² and the constitutional guarantee of equal protection likewise bars some forms of discrimination by state actors.⁴³ Although these laws offer protection against egregious forms of bias, they were developed to address a particular set of problems and in many ways are not well suited to address the challenges posed by biased algorithms.

Existing anti-discrimination law certainly applies to some discriminatory uses of AI. If, for example, an employer deliberately builds a tool with the intent of excluding a particular demographic group, it would run afoul of Title VII of the Civil Rights Act of 1964.⁴⁴ Similarly, if it relies on a predictive algorithm that disproportionately screens out members of a protected group, that practice would face legal scrutiny under the disparate impact theory.⁴⁵ And if selection tools, such as online tests or interviews, are inaccessible to persons with disabilities or fail to accurately measure their ability to do the job, then employers are required under the Americans with Disabilities Act to provide reasonable accommodation or an alternative test format.⁴⁶

The law thus provides some leverage for addressing biased algorithms; however, anti-discrimination case law has largely developed through individual suits challenging discrete decisions not to hire or to fire someone. As a result, much of the doctrine focuses on intentional discrimination, on the assumption that the central problem is a human actor with invidious motive. In contrast, algorithmic bias is typically not the product of conscious intent, and it operates on a systemic rather than individual basis, rendering some traditional approaches

⁴⁰ Civil Rights Act of 1964, P.L. 88-352, 78 Stat. 241 (1964).

⁴¹ The Americans with Disabilities Act of 1990, Pub. L. No. 101-336, 104 Stat. 328 (1990), codified as amended at 42 U.S.C. §§12101, et seq.

⁴² See, e.g., Equal Credit Opportunity Act, Pub. L. No. 94-239, 90 Stat. 251 (1976), codified as amended at 15 U.S.C. § 1691 et seq. (credit); Fair Housing Act, Pub. L. No. 90-284, 82 Stat. 81 (1968), codified as amended at 42 U.S.C. § 3601 et seq. (housing); Title VII of the Civil Rights Act of 1964, 42 U.S.C. §2000e et seq. (employment).

⁴³ See, e.g., *United States v. Virginia*, 518 U.S. 515 (1996); *Washington v. Davis*, 426 U.S. 229 (1976).

⁴⁴ See, e.g., *Teamsters v. United States*, 431 U.S. 324 (1977).

⁴⁵ The theory was first articulated by the U.S. Supreme Court in *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971), then codified by Congress in the Civil Rights Act of 1991 at 42 U.S.C. §2000e-2(k).

⁴⁶ 42 U.S.C. §12112(b)(6) and (7).

irrelevant. For example, although blinding a human decision-maker to information about race or gender can prevent bias, prohibiting AI tools from accessing that information will not preclude discrimination from occurring.⁴⁷ Because AI tools are built using rich datasets, information about race or other protected characteristics will be captured by numerous proxies in the data, such that an algorithm could effectively produce the same results, even with the sensitive characteristics removed. Moreover prohibiting access to this information may make it more difficult to audit the operation of the AI, and to detect and correct for unintended bias.⁴⁸

Disparate impact theory is better suited for addressing concerns that predictive AI will undermine equal employment laws, but applying that doctrine to biased algorithms also poses some challenges. Under current law, once a prima facie showing is made that an employer practice has a disparate impact based on a protected characteristic, the law provides the employer with a defense if it can show that the practice is “job related and consistent with business necessity.”⁴⁹ If the challenged practice is a predictive algorithm, when is that standard met? As explained earlier, these tools often rely on pattern-finding rather than explanatory theories, and may rest on wholly unexpected or even inexplicable correlations in the data. In the case of advanced machine learning techniques, the precise reasons behind a particular decision may be difficult to access, depending upon the designers’ choices when building the model.

In order to effectively promote workplace equality, the law should be applied in ways that recognize the nature and sources of algorithmic bias. In the employment context, an employer ought not be permitted to justify use of an AI tool with a disparate impact merely by showing that its predictions have a strong statistical correlation with the employer’s target outcome. Instead, courts should require the employer to show that the tool is statistically valid—in the sense of avoiding obvious sources of bias in the data—as well as substantively meaningful.⁵⁰ By substantively meaningful, I mean that there is some explanation of the decision process and how it relates to job performance. Otherwise, these tools may enable a form of statistical discrimination that would be unlawful if undertaken explicitly. If, for example, an employer refused to hire women of childbearing age because they are more likely to take time off for family reasons, it would clearly be engaged in unlawful discrimination. An algorithm might implicitly accomplish the same thing by selecting for applicants least likely to have a break in service in the near future. Thus, it is crucial that the employer provide sufficient information about the workings of a model with a disparate impact in order to judge whether its use is justified.

Predictive algorithms also raise questions about the scope of anti-discrimination law when multiple parties play a role in the decision-making process. If a discriminatory impact occurs, employers may try to disclaim responsibility because they relied on a facially neutral algorithm created by an outside vendor. The vendor may argue that it merely created the tool and is not responsible for how it is implemented in practice. Similarly, when job opportunities are distributed in ways that exclude certain demographic groups, it is uncertain whether the platform

⁴⁷ See, e.g., Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).

⁴⁸ See Kim, *supra* note 23.

⁴⁹ 42 U.S.C. § 2000e-2(k)(1)(A)(i).

⁵⁰ Kim, *supra* note 14, at 920-23.

intermediaries that channel the flow of this information will be considered employment agencies covered by anti-discrimination law.⁵¹ As predictive AI increasingly shapes how workers and employers interact in labor markets, courts will have to sort out how far liability for algorithmic bias should extend.

Other challenges are more practical. Anti-discrimination law currently relies primarily on individual litigants to identify violations of the law and bring enforcement suits; however, it will be exceedingly difficult for an applicant to determine when a predictive algorithm has a disparate impact. In order to do so, she will need data about outcomes across the applicant pool, not just in her individual case, and access to considerable technical expertise. Litigating such a suit would undoubtedly require significant resources as well. Because the risks of discrimination posed by predictive AI do not fit well with the backward-looking, liability-focused approach of existing law, some have argued that a regulatory approach may be more appropriate.⁵² Such a strategy might entail requiring employers to disclose when and how they use predictive AI. Regulatory agencies or trusted third parties would then be engaged to closely scrutinize the fairness of these tools.

While existing law provides some levers for addressing the discriminatory impacts of AI, U.S. law has been far less concerned with power imbalances and the social harms they might produce. The law intervenes to prevent extreme forms of exploitation, involving, for example, force or fraud, and occasionally steps in to prevent specific harms, such as threats to public health and safety. Otherwise, however, parties are free to structure their economic relations through private contracting, even when those arrangements are shaped by marked disparities in bargaining power or outright necessity of the part of the weaker party. Thus, the law as currently enacted is far more limited in addressing the ways in which predictive AI might contribute to economic inequality.

The power of AI comes from its ability to exploit large amounts of data, but legal limitations on the collection and use of personal information are quite anemic in the U.S. In the employment context, a handful of laws protect worker privacy, but they are ill-suited to an era of big data.⁵³ Constitutional limits apply only to state actors, not private companies. The common law privacy tort focuses on outrageous intrusions—a poor fit for the increasingly banal, routinized collection of information that characterizes today’s workplaces. Other sources of law, mostly statutory, protect certain specific types of information, such as medical information, thought to be particularly sensitive; however, access to large amounts of data and sophisticated statistical tools may allow an employer to infer that information without collecting it directly.

⁵¹ Kim, *supra* note 16, at 911-17.

⁵² See, e.g., Margot E. Kaminski, *Binary Governance: Lessons from the GDPR’s Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529 (2019); Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017).

⁵³ See Pauline T. Kim, *Data Mining and the Challenges of Protecting Employee Privacy under U.S. Law*, 40 COMP. LAB. L. & POL’Y J. 405 (2019).

For example, an algorithm may be able to predict whether or not an employee will get sick or become pregnant by analyzing other data available to the employer.⁵⁴

One possible response could be to strengthen workers' rights of control over their data. The difficulty is that traditional fair information principles have rested heavily on the notion of consent.⁵⁵ Given the unequal power that characterizes most work relationships, rights that are subject to waiver offer little protection. The EU's interpretation of the General Data Protection Regulation recognizes this reality by limiting employers from relying on consent as a basis for justifying data collection and processing.⁵⁶ Meaningful reform thus will likely require imposing non-waivable restrictions on how firms gather and utilize worker data.

Establishing mandatory rules regarding worker data faces another set of challenges, however. The types of data that can be collected and their potential uses are manifold and constantly evolving. In some situations, firms have good reasons for collecting data—perhaps to determine appropriate levels of inventory or to monitor the performance of their machinery—but that data may also provide information about workers' activities that could be used for decisions about pay, promotion or discipline. Employers have legitimate interests in information necessary to manage their workforces; however, some types of data gathering and use are unjustifiably intrusive or undermine workers' rights. The appropriate balance of those interests will vary depending upon the type of employer.

Precisely because of the wide variations in workplaces across industries and geographic regions, the law generally does not closely regulate the terms of employment other than setting some minimal standards. Instead, it attempts to address the disparity of bargaining power between employer and employees by permitting workers to organize and bargain collectively. The expectation underlying the labor statutes enacted in the early 20th century was that by leveling the playing field, workers would be able to bargain on an equal footing, protecting their own interests and negotiating fair resolutions appropriate to the needs of the particular industry or workplace.⁵⁷

The presence of a union generally improves wages and benefits, as well as provides a channel for voicing worker concerns,⁵⁸ including how new technologies are implemented. The reality, however, is that unionization rates, particularly in the private sector, have fallen to

⁵⁴ See, e.g., Rachel Emma Silverman, *Bosses Tap Outside Firms to Predict Which Workers Might Get Sick*, WALL STREET JOURNAL (Feb. 17, 2016), <https://www.wsj.com/articles/bosses-harness-big-data-to-predict-which-workers-might-get-sick-1455664940>.

⁵⁵ See, e.g., Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013).

⁵⁶ The GDPR states that consent to data processing is not considered valid “where there is a clear imbalance between the data subject and the controller,” GDPR, *supra* note 9, Recital 43(a). The European Union agency responsible for interpreting the GDPR has found that the employment relationship is one example of such a clear imbalance such that consent of employees should generally not be considered a legal basis for data processing. See Article 29 Data Protection Working Party, *Opinion 2/2017 on Data Processing at Work*, 17/EN WP 249 (June 8, 2017), http://ec.europa.eu/newsroom/document.cfm?doc_id=45631.

⁵⁷ See, e.g., Richard A. Bales, *The Discord between Collective Bargaining and Individual Employment Rights: Theoretical Origins and a Proposed Solution*, 77 B.U. L. REV. 687, 745–48 (1997).

⁵⁸ RICHARD B. FREEMAN & JAMES L. MEDOFF, *WHAT DO UNIONS DO?* (1984).

historically low levels, and very few workers have access to union representation.⁵⁹ The reasons for this decline are manifold, and include structural changes in labor markets such as the shift from manufacturing to service industries and increased global competition, as well as employer practices such as outsourcing and off-shoring jobs. Growing employer hostility to labor organizing, ineffectual legal remedies, and judicial retrenchment of workers' labor rights have also contributed.⁶⁰

Regardless of the reasons for the decline, the important point here is that in the absence of robust collective activity, workers have little voice in how algorithmic management tools are used and firms face few constraints on further increasing their control. Thus, addressing concerns that predictive AI tools are contributing to inequality will require broader legal reforms that strengthen employee voice in the workplace.

6. Conclusion

When AI is used in social domains to make predictions about people, it poses heightened risks not only for those individuals, but also of broader societal consequences. As the example of the workplace shows, AI tools can contribute to inequality—both by reproducing discriminatory patterns that disadvantage marginalized groups, and by further concentrating power in ways that increase economic inequality. These consequences of the use of AI are not inevitable; much depends upon the choices made in developing and deploying these tools. Legal and policy reforms will be required to ensure that AI applications do not simply reproduce or exacerbate existing inequalities, but instead promote broader social goods.

⁵⁹ JAKE ROSENFELD, WHAT UNIONS NO LONGER DO (2014).

⁶⁰ Cynthia L. Estlund, *The Ossification of American Labor Law*, 102 COLUM. L. REV. 1527 (2002).