Economics, Fairness and Algorithmic Bias*

Bo Cowgill
Columbia University

Catherine Tucker
MIT & NBER

May 11, 2019

Abstract

We develop an economic perspective on algorithmic fairness. Algorithmic bias and fairness issues are appearing in an increasing variety of economic research literatures. Our perspective draws from obvious connections to the economics of discrimination, crime, personnel and technological innovation; as well as less obvious connections to environmental economics, product safety regulation, behavioral economics and economics of information. We survey the small but growing literature in economics that directly examines this topic theoretically and empirically. Algorithms are not only growing in social impact, but also have attractive measurement properties that ease challenges for research economists. We conclude by discussing economic policy implications and future research directions.

*Bo Cowgill is an Assistant Professor at the Columbia Business School. Catherine Tucker is the Sloan Distinguished Professor of Management Science at MIT Sloan School of Management and Research Associate at the NBER. The authors thank David Blei, Charlie Brown, Sam Corbett-Davies, Fabrizio Dell'Acqua, Michael Impink, Ray Horton, Daniel Kahneman, Bruce Kogut, Zach Lipton, John Morgan, Alex Miller, Rob Seamans, Orie Shelef, Olivier Sibony, Megan Stevenson, Neil Thompson, Mike Yeomans and Angela Zhou for helpful discussion and feedback. Cowgill thanks the Kauffman Foundation and the W.E. Upjohn Institute for supporting this research.

1 Introduction

Algorithms – the application of mathematical formulae to observed data – are increasingly engaged in economically important decisions. They are used to to make decisions regarding sentencing in criminal courts, resume screening, pricing, ad-placement, lending decisions and the news media that citizens consume. This development has generated a public debate about bias and unfairness in machine-guided decisions, including several high-profile allegations in finance (Bartlett et al., 2018), criminal sentencing (Dressel and Farid, 2018), hiring,[1] and ad targeting (Datta et al., 2015).[2] Fairness concerns have resonated with policymakers in multiple countries, who have adopted or are considering fairness-related regulations for algorithms.[3]

There is a large and quickly-growing literature on the Computer Science (CS) issues around algorithmic fairness and bias.[4] This paper provides a notably missing perspective: One from theoretical and empirical economics.

In this essay, we develop an *economic perspective* on algorithmic bias and fairness. Our perspective draws not only from obvious parallels in the economics of discrimination, crime, personnel and technological innovation; but also on less obvious connections to environmental economics, product safety regulation, behavioral economics, and economics of information. Algorithmic fairness is a shifting and normative concept. We do not address how diverse beholders should conceive of fairness, but instead the economic implications of these diverse conceptions.

Our central claim is that a well-designed algorithm can be an enormous force for positive social change. Algorithms have potential to reduce demographic disparities, reduce the effects of behavioral biases and improve outcomes in substantive areas ranging from crime, to finance to labor

[1] "Amazon scraps secret AI recruiting tool that showed bias against women," https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Cappelli et al. (2018) discuss challenges in in the AI/hiring setting.

[2] There have also been lawsuits challenging the fairness of algorithms used to evaluate teachers: *Houston Federation of Teachers v. Houston Independent School District*. https://www.houstonpublicmedia.org/articles/news/2017/10/10/241724/federal-lawsuit-settled-between-houstons-teacher-union-and-hisd/.

[3] German chancellor Angela Merkel stated, "Algorithms, when they are not transparent, can lead to a distortion of our perception": https://www.theguardian.com/world/2016/oct/27/angela-merkel-internet-search-engines-are-distorting-our-perception. The Obama White House published a report entitled "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights" (Smith et al., 2016) which highlighted the potential for algorithms to lead to 'disparities in treatment and outcomes' in credit decisions, employment, higher education and criminal justice: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf. Other US government bodies such as the US Equal Employment Opportunity Commission and FTC have have reacted with public statements and policy guidance. In October 2016, the U.S. EEOC held a symposium on the implications of "Big Data" for Equal Employment Opportunity law. https://www.eeoc.gov/eeoc/newsroom/release/10-13-16.cfm. FTC hearings in December 2018 on the use of algorithms focused on questions of their fairness, transparency and ethical uses: https://www.jdsupra.com/legalnews/ftc-hearings-exploring-algorithms-52122/.

[4] Several new focused computer science conferences have started (for example, "ACM Conference on Fairness, Accountability, and Transparency," https://fatconference.org/ and "AAAI/ACM conference on Artificial Intelligence, Ethics, and Society" http://www.aies-conference.com/). In addition, the proceedings and best paper awards of top machine learning conferences such as ICML and NeurIPS feature multiple articles about fairness and machine learning. Computer science research funders have also targeted this topic. For example, in June 2017, computer scientists at the University of Wisconsin, Madison were awarded a $1M grant to study algorithmic bias. http://host.madison.com/news/state-regional/wisconsin-researchers-awarded-grant-to-fix-algorithmic-bias/article_d3561a49-1545-5e4a-b2cf-550b4703bcef.html.

markets.

This paper is also intended to encourage research by economists into these topics. Many issues raised in this article apply to a variety of economic sub-disciplines. The ability of algorithms' to reduce or remove bias may partly explain firms' decisions to replace human labor with capital (Brynjolfsson et al., 2018; Furman and Seamans, 2019). Algorithms are already used widely in many settings where economists have historically concerned themselves with bias.[5]

Towards that end, our paper highlights five key themes:

First, focusing on internal algorithmic characteristics, in isolation of the economic context in which they operate, is misleading. The properties of algorithms – what data was used in training them, how certain variables are weighed – may misguide observers about the practical effects of new algorithms on pre-existing social systems.

Much intuition about algorithmic characteristics is incomplete. Utilizing biased data to train algorithms may yield reductions in bias, particularly if these datasets contain noisy behavior that effectively act as experiments. Variables utilized in algorithms need *not* have a causal interpretation – or any interpretation – in order to have economic value to decision-makers. Algorithms can perpetuate biases (or not) while directly utilizing sensitive variables (or not). How algorithms affect job-seekers, criminal defendants and loan applicants depends not only these internal characteristics, but how these algorithms interact with the outside world.

Effective policy for algorithms should therefore regulate outcomes, not inputs. Regulating inputs – for example, by requiring certain engineering practices or technologies – exhibit the "command-and-control" approach economists have long opposed in arenas such as environmental policy. Regulations focusing on outcomes exhibit more flexibility, fewer loopholes, greater efficiency and stronger incentives for innovation. Causal inference techniques offer tools for assessing outcomes, but the economics approach differs from the way that computer scientists in algorithmic fairness currently conceive of "counterfactual fairness."

Our emphasis outputs on has implications for policies supporting transparency. These policies are not only potentially misleading, they also have economic drawbacks that must be weighed against their merits. Transparency facilitates collusion in many common settings. It permits theft of costly innovations and thereby reduces incentives for innovation. It weakens security. While the "black box" nature of machine learning is often criticized, it has merits. By accident or design, many existing processes are opaque. In many cases this is a preferable, perhaps essential part of a sustainable equilibrium.

The *costs* of AI development are an exception to our emphasis away from internal characteristics. The cost structure of AI engineering suggests that outperforming humans is *not* an economically natural stopping point. Fears that firms will reduce bias "slightly below human levels, and then stop" are not justified. Comparisons against status-quo human decisions are therefore a reasonable

[5] We estimate that approximately 25% of the US population lives in a jurisdiction where algorithmic guidance is provided to judges in state or local criminal courts. Pretrial risk assessment algorithms are already used in all federal criminal courts in the United States. A 2012 *Wall Street Journal* article estimates that the proportion of large companies using resume-filtering technology as "in the high 90% range," and claims "it would be very rare to find a Fortune 500 company without [this technology]." http://www.wsj.com/articles/SB10001424052970204624204577178941034941330, accessed January 19, 2019. Approximately two-thirds (67.5%) of the public believe that most large companies computer algorithms to sort job applications (Cowgill, 2017).

starting point for evaluating bias in AI.

Second, algorithmic fairness must be evaluated in a strategic context. Although predictive signals need not be interpretable to yield economic value, they must be must be costly to acquire; if they are not, their value will be diminished through strategic manipulation.

Algorithms utilize existing incentive and create new ones. There is a strategic component to "explainable algorithms," an increasing requirement for software developers. The economics literature about "cheap talk" suggests that lack of explanations has little to do with technological shortcomings of the message space (i.e., lack of data visualizations, statistics or examples). Explanations are made possible through better alignment of incentives between the sender of the explanation (the algorithm, its designers and owners) and the receiver (a human client), and making alignment common knowledge (or otherwise addressing the alignment, perhaps through vertical integration). Otherwise, the finest visualizations cannot be distinguished from strategic, self-serving "cheap talk."

Strategizing also relates to optimal policymaking. Efforts to police algorithmic bias will produce behavioral responses similar to those for other crimes: People could reduce crime, or increase evasion. Algorithms play divergent roles: They may reduce bias below current levels, but leave the remaining bias more visible and exposed to policing. Firms might then reduce any remaining algorithmic bias, or they may shift to more opaque forms of decision-making such as human discretion. Policies that do not anticipate opportunities for evasive behavior will fail to reduce unfairness, and may lead firms to hide bias behind less transparent processes.

Third, behavioral economics affects algorithmic fairness in many directions. A large literature documents systematic failures in predictions by well-intentioned humans, including both biases as well as noisy judgments affected by superfluous factors. Several of these shortcomings have theoretical microfoundations that machine learning can plausibly correct, even without extensive "fairness adjustments" to the technology. In some cases, behavioral-economics style prediction errors may actually *help* machine learning arrive at better conclusions by inadvertently exploring the space of potential decisions.

A related error in human judgment, supported by a large body of evidence (including several recent economics papers), is a *reluctance* to trust algorithms. This is particularly striking because public discourse frequently alleges that algorithms hypnotize the public into obedience through "scientific veneer."[6] There is little research documenting a deference to "scientific veneers."

If the public is indeed naive about quantitative arguments, this should not apply asymmetrically to *praise* of algorithms. *Critiques* of algorithms also benefit from "the veneer of scientific objectivity." Skeptics can systematically interrogate algorithms in ways that will almost certainly uncover some form of unfairness. Multiple theoretical papers demonstrate the mathematical impossibility of satisfying all fairness criteria, particularly simultaneously. Human- and other status-quo decision processes do not escape these impossibilities. They are simply more resistant to critical inspection.

Fourth, economic theory suggests that firms have profit-oriented motives for reducing bias.

[6]For example, a recent statement by the AI Now Institute and NYU Law's Center on Race, Inequality and the Law states that "[t]he use of risk assessments can give judges own biases and uncertainty about individual behavior a false veneer of scientific objectivity[...]" https://ainowinstitute.org/sentencing-risk-assessment.pdf.

This is true even without regulatory punishments, fines, lawsuits or bad PR. Firms face normal production-and-sales reasons to use the most accurate predictions whenever possible. Impediments to adoption may not arise from profit alignment, but from other frictions such as awareness, uncertainty about techniques, unavailability of expertise or raw inputs necessary to de-bias algorithms. This does not guarantee firms will give bias reduction their highest priority, but it does suggest that if regulators, vendors and activists can make de-biasing easy, then firms will do it.

The profit-maximization value of reducing bias suggests that *public* policies can be advanced through *private* processes. In other words, the benefits of de-biasing are partly privatizable.[7] This suggests that a sustainable, for-profit marketplace for bias-reducing technology is viable, even if the bias-reducing technology is not marketed as such.[8] In the profit-oriented world, however, there may be little incentive for sharing innovations without the development of a separate market for bias-reducing technology. Although bias-reduction is profitable, other conceptions of fairness are not.

Fifth, algorithms can improve social outcomes despite the concern about algorithmic bias. Several early, high-quality empirical studies demonstrate improvements in fairness, diversity and bias metrics compared with status-quo processes, often while simultaneously improving performance. These empirical studies validate specific theoretical arguments about mechanisms for bias reduction. Social scientists across disciplines, including economists, have found extensive evidence of bias in the pre-algorithmic world. Some of the problems associated with "algorithmic bias" are not algorithmic problems. Negative side-effects and self-fulfilling prophecies may result from perfectly accurate, unbiased predictions. These problems are the consequences of prediction, not bias, and they reflect larger social problems.

We elaborate on these five themes throughout the paper using theoretical and empirical evidence. Many popular and policy articles about algorithmic fairness and conclude with vague admonitions, for example, "be especially careful" (Byrnes, 2016).[9] Arvind Narayanan, a computer scientist at Princeton, wrote that "AI won't replace careful social science and statistics." We aim to provide specifics about what being careful means in this setting, and to provide examples of careful social science on the key economic questions about this phenomena.

The ideas in this paper apply to a variety of settings where humans, machine learning algorithms, or simple tests make decisions (such as in lending, criminal justice or advertising). For exposition, we mostly use language around *hiring* decisions and the possibility of gender bias throughout the essay. However, analogous ideas can be applied to many other settings, and towards other types of bias (including behavioral economics-style cognitive biases and/or biases against other characteristics). For example, employers may exhibit "pedigree bias" against non-

[7] By contrast, other problems at the intersection of business, engineering and public policy feature externalities that distort incentives for socially desirable outcomes.

[8] The viability of a for-profit marketplace for de-biasing technology is particularly strong in the presence of outcome regulation (mentioned above) which preserves these incentives for production and innovation in bias-reducing technology. By contrast with "command-and-control" regulation shuts down these marketplace and related incentives.

[9] The Obama' Administration's statement on algorithmic bias (Smith et al., 2016) similar stated, "Without deliberate care, these innovations can easily hardwire discrimination, reinforce bias, and mask opportunity." The statement conceded "The purpose of the report is not to offer remedies[.]" A 2017 Quartz article concluded, "But we must also be mindful of the specter of harms like algorithmic discrimination and implicit harmful bias in AI-enabled recruiting, and do our best to counter them." https://qz.com/work/1098954/ai-is-the-future-of-hiring-but-it-could-introduce-bias-if-were-not-careful/

elite colleges. This may not be illegal or a controversial political issue, but would nonetheless distort fairness or optimal hiring. Some of the ideas here apply not only to computer algorithms and complex machine learning, but also to simpler forms of evaluations such as psychometrics or other job tests.

The paper proceeds as follows. We begin in Section 2 by introducing and applying the classic economics conceptions of bias to algorithms. Section 3 discusses econometric and behavioral theories for why algorithms may exhibit human biases. In Section 4 we discuss the strategic and behavioral economics considerations for algorithmic bias. Section 5 reviews the nascent "algorithmic fairness" literature in computer science. In Section 6, we discuss theory and evidence on the topic of algorithms correcting, rather than amplifying, biases in human decision-making. In Section 7, we discuss implications for public policy and firm practices, and conclude with a discussion in Section 8.

2 Classic Economic Approach to Bias

2.1 Definitions of Bias in Economics

We begin by defining bias as typically used by economists. Economic decision-making is biased if it produces unequal productivity across groups *at the margin*. This standard is sometimes called the "Becker test" after its originator.[10] Below, we unpack this definition and discuss its applications to algorithms. We then discuss the behavioral theories of humans that give rise to unequal outcomes at the margin, and how these relate to algorithms trained from human data.

Productivity Becker's tests of equal productivity means "equal payoff for the firm" or equal benefit for whatever the decision-maker should be optimizing. This formulation presumes the existence of a well-defined objective function for the decision-maker. It also presumes that this objective can be measured error-free. In many settings it may be obvious what the decision-maker should optimizing. In bail decisions (Arnold et al., 2018), judges are often charged with minimizing failures to appear. In lending (Dobbie et al., 2018a), financiers should maximize the probability of repayment. Performance outcomes in many occupations (such as those in finance, sales and some forms of manual labor) can be measured objectively, and it is clear what decision-makers should maximize.

However in other settings, particularly white-collar or creative jobs, the objective function is less clear. Should hiring policy seek to maximize innovation? Or efficiency? Should University admissions maximize academic performance, post-graduate employment or alumni donations? Various answers to these questions have different implications for bias. There may be no "correct" answers to these questions, distinct from normative preferences and tastes. Insofar as correct answers truly exist, researchers and practitioners may lack the data and experiments necessary to determine optimal strategies. Obermeyer and Mullainathan (2019) find evidence of racial dispar-

[10] The foundational ideas about discrimination in economics were published in Gary Becker's 1955 PhD thesis, later published as a book (Becker, 1957). In this and later writing (1993), Becker suggested empirical "outcome tests" for bias that were motivated by economic theory. Applied to hiring, the Becker outcome test suggests that if hiring is unbiased, then the productivity of the marginal male and female employees should be equal.

ities in a commercially available health predictor algorithms driven by this issue: The algorithm was optimized around health care costs, rather than health.

Choosing the "objective" function may therefore be highly subjective. Even in cases where the objective function is conceptually clear, measurement of the objective may be problematic. Many important abstract objectives ("customer satisfaction," "cultural fit," "changing public sentiment") resist easy quantification.[11] The introduction of algorithms makes objective functions – and the implicit tradeoffs within – more explicit and transparent. Disagreements nominally about bias conceal differences about what objective function to use (and *vice versa*).

Marginality The Becker outcomes test examines productivity between *marginal* male and female hires. The "marginal female hire" is the least qualified woman hired. If an employer dislikes hiring women, the least-qualified female hire should exhibit higher performance than the least-qualified male hire.

Becker focused on the *marginal* hires as distinct from the *average* hires by gender. There may be large differences in the outcomes of *average* hires that are *not* the result of bias, for example, if one group's applicants performed better on average.[13] Though is not the norm, such differences may be more common in (say) criminal justice, where the base rates of criminal behavior may be truly different between men and women.[14] The idea of discrimination on the margin has natural application in non-hiring settings such as criminal justice and loan decisions, where machine learning applications have also been explored (Kleinberg et al., 2016; Hardt et al., 2016).[15]

Measuring marginal candidates is difficult in non-algorithmic settings. Hiring decisions by human recruiters typically label who is hired or who is not, but do not create a rank order that would allow a researcher to isolate marginal candidates. Some researchers use structural assumptions to infer rank orders or thresholds (Simoiu et al., 2017). An approach requiring fewer assumptions requires a source of random variation that shifts who is selected, but doesn't affect outcomes beyond the selections.

Candidates affected by such random variation are *marginal* because their hiring outcomes are so precarious that they could be affected by random environmental variation.[16] Several papers use the random assignment of a case to a judge to examine marginal candidates; Arnold et al. (2018) exploit the random assignment of judges to measure bias in bail setting; Dobbie et al. (2018a) exploit the random assignment of loan examiners to measure bias in lending. This strategy could

[11] In some cases, the deployment of an algorithms actually *changes* subjects' outputs rather than revealing their latent types. For example: Many papers in computer science (Dressel and Farid, 2018) examine criminal courts where judges were shown algorithmic recidivism scores. These papers use rearrest outcomes within two-years to measure whether the right people were jailed. A variety of prior literature in criminology, economics and political science suggests longer prison sentences may *cause* defendants' likelihood of re-arrest to increase.[12] This could be the result of greater criminality or greater police monitoring. Either way, comparisons of recidivism between marginally imprisoned black (vs white) defendants may not reflect different standards. It may instead reflect asymmetric negative effects of prison on racial minorities.

[13] Of course the measurement of performance may itself be skewed by gender.

[14] Ayres (2002); Simoiu et al. (2017) discuss other problems of infra-marginality in measuring bias.

[15] Criminal trials produce decisions about guilt. The "marginal defendant" is the defendant least likely to be guilty among those.

[16] By contrast: Non-marginal candidates' chances are unaffected by this random variation, either because they are well-above the threshold (and always admitted), or because they are far below (and thus never admitted).

be used in many other settings where evaluators are quasi-randomly assigned.[17] Researchers also could use other sources of randomness in decision-making introduced as studied in behavioral economics.[18]

The difficulty of studying marginal candidates complicates the evaluation of human biases in natural settings. By contrast, a hiring algorithm enable not only binary decisions, but rank ordering (and distances between) all candidates' scores. This enables marginal candidates to be examined using regression discontinuity-style procedures (Lee and Lemieux, 2010).[19]

This is one example of the many ways that bias in algorithms is more easily measured than bias in humans, a theme that has implications for who adopts algorithms and how they are regulated by policymakers (discussed in Sections 6 and 7). Firms facing high penalties for discrimination may prefer to employ human decision-makers, since their biases can be more easily concealed.

Challenges in Implementing Standard Economic Tests of Bias We conclude with four observations about Becker's tests. First, these empirical tests say nothing about whether a decision-maker directly utilizes sensitive variables as inputs to decision-making. It is well-known that a policy can fail these outcome tests, even without directly using demographic variables. For example, if an advertisers targets an ad at people who express an interest in Assasin's Creed on Facebook, the ad will be seen by over 89% men.

Less well-known is that it is possible to *pass* the tests while directly using using these variables extensively. The Becker test uses sensitive variables only in evaluations of outcomes.

Second, the Becker outcome test is an *ex-post* standard. It requires an employer to *implement* the selection procedures in practice so that outcomes can be measured. The test doesn't offer pre-certification selection procedures – only a way to evaluate them after the fact.

Third, the Becker test evaluates selection procedures by an absolute standard rather than by comparison to alternatives. Procedures that fail the Becker test may nonetheless reduce bias beneath a status quo. Much of economics is concerned with counterfactual comparisons (Angrist and Pischke, 2008), which the (classic) Becker test is not about. We discuss the role of counterfactual thinking in algorithmic fairness later in Section 7.3.

Finally, the Becker tests examine *bias*, a particular form of unfairness. However, other types of unfairness and inequalities exist. Procedures that pass the Becker test may fail these other criteria, as we discuss in more depth below.

[17] This includes criminal cases to judges (Kling, 2006), patent applications to patent examiners (Sampat and Williams, 2014; Farre-Mensa et al., 2017); foster-care cases to foster care workers (Doyle Jr et al., 2007; Doyle Jr, 2008); disability-insurance applications to examiners (Maestas et al., 2013); bankruptcy judges to individual debtors (Dobbie and Song, 2015) and corporations (Chang and Schoar, 2013); and job seekers to placement agencies (Autor and Houseman, 2010).

[18] For example, weather affects financial decisions (Rind, 1996; Hirshleifer and Shumway, 2003; Busse et al., 2015), sports victories affect financial decisions (Edmans et al., 2007) and relationships (Card and Dahl, 2011), stock prices affect decisions about effort, innovation, job candidates (Cowgill and Zitzewitz, 2008) and health (Engelberg and Parsons, 2016).

[19] Cowgill (2018a), Stevenson and Doleac (2018) and Berk (2017) use regression discontinuities in various ways to examine algorithmic fairness questions.

2.2 Economic Theories Explaining Biased Behavior

Having defined a set of empirical tests, we now examine the preferences, strategic interactions and learning technologies between people and firms that generate biased decisions (i.e., unequal outcomes on the margin). These economic structures are the data generating processes that produce training data used in machine learning. Typical discussions of bias in economics focuses on two types of discrimination: *Taste-based* discrimination and *statistical* discrimination. In addition, statisticians have noted the potential for discrimination based on prediction errors.

Taste-Based Discrimination *Taste-based discrimination* arises directly from preferences. A recruiter exhibiting taste-based discrimination enjoys direct utility for selecting his favorite type of worker. That is, they feel subjectively better (or worse) for selecting particular workers for reasons unrelated to performance.

Employers' tastes for discrimination is typically modeled as substitutable with other forms of utility. For a taste-based discriminator making employment decisions, a worker's poor performance can be offset by taste-based preferences in the worker's favor. In most models, this generates productivity differences on the margin of hiring across groups (as described in the empirical tests above).[20]

Psychologists have proposed *conscious* vs *unconscious* bias (Greenwald et al., 1998). Neither taste-based discrimination nor statistical discrimination requires subjects to recognize their bias. Although we devote less space in this essay towards taste-based discrimination, several scholars suggest that tastes could be a major contributor to inequality in a variety of settings (Guryan and Charles, 2013). Taste-based biases could bias algorithms both through the selection of either inputs or optimization criteria.

Statistical Discrimination In contrast to tastes, discrimination also can arise from signal extraction problems even if tastes are demographically neutral. "Statistical discrimination" (Phelps, 1972; Arrow, 1973) refers to educated guesses about a subject outcome based on limited information. If performance of workers are (on average) correlated with observable characteristics, employers may be tempted to use these variables as proxies for unobserved skills. Employers exhibiting only statistical discrimination would be completely indifferent between candidates of varying demographics if workers' quality were observable and equal across groups. Statistical discriminators utility functions are only about performance; they care about demographics insofar as they help predict performance.[21]

Statistical discrimination theory makes an optimistic prediction about machine learning and demographic bias: As better data becomes available about workers' characteristics and their performances, statistical discriminators should ignore crude demographic proxies. They will instead

[20] Some models predict productivity differences both on average and on the margin between genders (Knowles et al., 2001).

[21] Note: This applies only for jobs in which performance is truly uncorrelated with demographics. It may not apply, for example, to sales jobs where potential buyers have a taste for pitches from salespeople of particular backgrounds. For this type of job, demographics may truly does predict performance. Statistical discrimination theory predicts that firms concerned only about profits would prefer to continue using demographic signals.

place more weight on variables tightly linked to performance. The magnitude of the shift depends on how much demographic factors are truly correlated with performance.

In this sense, "big data" could be a positive force for demographic equality by providing new predictive variables that are uncorrelated with demographics. Some of these new variables may be available to humans already, but unincorporated into learning due to cognitive limitations around noticing (Hanna et al., 2014; Schwartzstein, 2014).

Many economic models feature decision-makers with perfectly accurate statistical predictors.[22] It is unclear how humans could arrive at perfectly accurate statistical discrimination, even if they were aided by computers and data. Humans may exhibit similar obstacles as machine learning engineers – unrepresentative training samples, flawed labels within their training sample, and other contributors to biased algorithms outlined in Section 3.

The challenges around algorithmic bias and fairness are similar to others at the intersection of computer science and economics, where behavior in economic models requires intractable computations.[23] Reducing bias similarly requires intractable computations (learning is generally computationally complex, i.e. NP-hard, Kearns, 1990, Guruswami and Raghavendra, 2009, Daniely, 2016; learnability itself may be undecidable Ben-David et al., 2019). However, computational hardness has not prevented progress in many applied machine learning problems in recent decades.[24] The general hardness of learning may not prevent progress in reducing bias from pre-existing levels, even if some bias remains.

Classic models that assume decision-makers overcome these challenges show that even in this idealized world – one featuring demographically-blind preferences and flawless human statisticians – inequalities may arise for signal-extraction reasons. A world of accurate statistical discrimination still faces many problems. For example, demographic profiling could pass the marginal outcomes test.

Another such problem is the "self-fulfilling prophecies" phenomena in Arrow (1973) and subsequent models featuring endogenous skill acquisition (Lundberg and Startz, 1983; Coate and Loury, 1993). In these models, employers' beliefs about workers' ability levels affect employers' hiring decisions, which affect the rate of return on human capital investments, which in turn determine workers' choices of training and skill investments and eventually their realized skills. Employers' negative prior beliefs are self-confirming in equilibrium thanks to their downstream effects on incentives for human capital.

[22] At a 2018 NBER AI conference, economist Joshua Gans said, "AI is terrible for economic theorists. We already have in our models people able to do *perfect* statistics – who could apply statistical analysis at a frontier that [actual humanity] hasn't reached." Alternatively perhaps, AI is good for economic theory insofar as it aligns real-world phenomena with optimizing behavior modeled in theory.

[23] Summarizing research at this intersection, Aaronson (2013) wrote, "[E]ven in the idealized situation [...], it will often be *computationally intractable* for those agents to act in accordance with classical economics." The computational complexity of deriving market-clearing prices in Arrow-Debreu markets and/or Nash equilibrium in bimatrix games is relatively computational intractable (Chen et al., 2009); Aaronson (2013) writes that the worst-case complexity of deriving Nash/Arrow-Debreu equilibria is "as close to NP-complete as it could possibly be." Even the game of billiards, featured in Milton Friedman's celebrated analogy ("The hypothesis that the billiard player made his shots as if he knew the complicated mathematical formulas," Friedman, 1953) has proven intractable for computer scientists (Archibald and Shoham, 2009; Archibald et al., 2010).

[24] The complexity of learning is based on worst-case scenarios for input data. Real-world datasets may exhibit characteristics that make learning more feasible than worst-case. The same may be true about learning to reduce bias.

Importantly, this type of self-fulfilling prophecy does not require that the statistical discrimination be *inaccurate*. These models suggests that some of the negative byproducts of algorithms may arise even if the algorithms are unbiased. For these reasons, economists have studied other models of fairness discussed in Section 2.3. We revisit statistical discrimination in later sections about behavioral economics and *inaccurate* statistical discrimination (Section 4.2) and the strategic content of variables used for statistical discrimination (Section 4.1).

Discrimination Stemming From Prediction Errors A separate approach to statistical discrimination emphasizes differences in the variance (or dispersion) of outcomes within members of the group. Computer scientists have noted that a group-blind classifier that minimizes overall error will fit the majority population better (Chen et al., 2018a; Chouldechova and Roth, 2018). As these papers discuss, is possible that variables should simply be weighed differently for minority populations, but once they are, comparable levels of predictive accuracy are possible for minority and majority groups.

However, prediction errors for minority groups can arise from both bias as well as variance. The latter possibility is raised in a literature emphasizing differences in within-group variance. These models suggest that minority groups outcomes may be more unpredictable because productivity outcomes are truly more uncertain (Phelps, 1972; Aigner and Cain, 1977). For example, as a result of discrimination, minorities may be forced to assume greater risk. Alternatively, unpredictability could arise because signals, rather than outcomes, are noisier and less informative (Morgan and Várdy, 2009), no matter how they are assembled into a model. These models provide theoretical microfoundations, empirics and policy prescriptions for the idea that prediction disproportionately fails for minority groups.

2.3 Other Fairness Considerations

Bias is a particular type of unfairness. Society may have other fairness-related goals. For example, organizations may have a preference for sacrificing efficiency to enhance diversity or to enhance social justice (for example, affirmative action). Traditional economic approaches to these tradeoffs involve a formalized social welfare function that combine payoffs from diversity, equity and efficiency.

In some cases, it may suffice to change the objective function used to measure individual-level productivity and proceed with the Becker test. For example, a welfare function could proceed by specifying a set of acceptable individual-level tradeoffs between a worker's additional contributed units of productivity and the utility from hiring from underrepresented background.

Ludwig et al. (2018) formalize the social-welfare approach for regulating fairness in algorithms. Other fairness considerations face two complicating issues. Some outcomes (such as diversity) are group-level outcomes that create interdependencies between applicants. A worker's contribution may depend on who else was simultaneously (or previously) admitted. Second, the presence of other fairness requirements affect incentives on the parties affected by screening, for example by changing the returns to effort or human capital investments.

2.4 Traditional Policy Options for Reducing Bias

Starting with Becker's 1957 article, economic research finds that competition disciplines bias. Bias from tastes or from inaccurate statistical discrimination may cause employers to hire unproductive workers, reducing profitability. Under some conditions, this will lead firms to go out of business or be penalized by capital markets; this threat may discipline firms to reduce or eliminate bias. While this form of self-correction is theoretically possible, many economists acknowledge government action is required to eliminate discrimination.

The discrimination literature turns to the economics of crime for policy guidance, which featuring three primary policy levers for shaping discriminatory behavior: Detecting discrimination, punishing known discriminators and reducing the benefits from discriminating. We mention two particularly salient issues for algorithms.

Accuracy of Statistical Discrimination The accuracy of firms' statistical predictions is critical for policy implications. Punishments for over-hiring men could offset an employer's utility payoffs for men. This discipline could induce employers into equalizing genders on the margin.

The significance of accurate statistical predictions does not end there. The accuracy of forecasting also affects how a guilty employer responds to the punishment threat. Suppose government policy fined all employers whose workforces exhibited gender productivity differences on the margin. If employers could accurately predict candidates' performance, they would know exactly how to respond to this punishment. They would know exactly who should be hired in order to achieve equality on the margin. Such firms would simply lower the "expected performance" threshold for hiring a woman until it equals the threshold for men (or raise the threshold for men until it equals women's). This would result in the efficient set of candidates being hired.

On the other hand, if firms are *not* accurate statistical discriminators, this raises more challenging questions about how guilty firms respond. How does an employer know who needs to be hired in order to achieve equality on the margin? The employer may have no idea; she may have felt she was already compliant. She may be correct; all forms of policing feature some level of false positives. Why would a government regulator have expertise in how to weigh candidates' strengths and weaknesses in a given industry?

Adjustments intended to equalize one margin (say, by hiring more women) could create new inequalities or problems on other margins, unless employers were perfectly accurate predictors. A firm could over-hire women and create inequality against other gender identities. Efforts to address gender inequality could affect inequality on other vulnerable groups. Firms could alternatively lower thresholds by too much and hire unprofitable or destructive workers.

Optimally adjusting hiring policy for compliance is not obvious. If it were, discrimination may not be so difficult. The section above catalogues reasons for inaccurate statistical discrimination by humans. As we mention, even computer scientists with PhDs, data centers full of processing power and extensive historical performance data cannot guarantee unbiased predictions. As highlighted by computer scientist Arvind Narayanan (2018a; 2018b), "Bias in machine learning is the rule, not the exception." His reason is that the underlying training data comes from historical systems by humans containing bias. Why would human cognition be different?

This uncertainty is why field experiments may be a useful practical and policy tool. Finding truly optimal hiring policy may not be practical, but finding improvements to the status quo may be. A well-designed field experiment may allow employers to estimate the effects of counterfactual screening policy, including whether the policy's affects are directionally positive. We discuss the design of field experiments for assessing algorithmic fairness in Section 7.

"Input Regulation" and Evading Detection Firms can circumvent regulation through evasion. Evading detection is particularly relevant for deterring bias, particularly in algorithms. As economists have considered policy solutions to bias, directly regulating preferences or statistical technologies used has been mostly off the table, because of the impracticality of legislating the permissible variables to appear in an employer's utility function or predictive reasoning. No regulator can tell a human recruiter, "Pretend you cannot see this variable and don't care about it." This would constitute unenforceable regulation of decision-making hidden inside brain cells. Employers may not even consciously know their own preferences or statistical abilities.

Some regulation attempts to do this anyway. "Disparate treatment" is a legal framework forbidding employers from directly using certain variables in decisions. However, this framework affects behavior only insofar as this behavior can be monitored. Firms can avoid detection by leaving no evidence. This may be a deliberate strategy to avoid intrusive searches. Alternatively, some firms regularly expunge documents, even if they believe they are innocent, in order to avoid policing false-positives.

Even without an evasion strategy, many decisions are made inside managers' minds or in verbal discussions. There may be no reliable record about which variables were considered.

The difficulty of detection is one reason why economists disfavor "regulating inputs" (i.e., the content of preferences and predictions) rather than regulating outcomes. Courts and regulators have embraced "disparate impact" regulations based on outcomes, including the federal government's "4/5ths rule" (discussed in Section 5). Becker's emphasis on *marginal* applicants has mostly *not* been incorporated into policy, perhaps because of the aforementioned trouble identifying marginal subjects in human decisions.

Unlike human decisions, directly regulating the variables inside algorithms is technologically feasible, and has political and legal support. Monitoring compliance is easier through electronic discovery, particularly given the popularity of version control software for engineering teams developing large amounts of code (Gentzkow and Shapiro, 2014). This does not mean that input regulation is a good idea. Statistical discrimination theory suggests when such regulation may be effective. Given the implications for monitoring, firms may prefer to forgo algorithms altogether and embrace human-based alternatives where bias is actually worse, but that resist invasive searches and monitoring. We discuss input regulation and evasion for algorithms in Section 7.3.

3 Sources of Bias in Algorithms

Why would algorithms exhibit bias?[25] We discuss four hypotheses about the origins of algorithmic bias, and their sources in economic phenomena. There is little research seeking to quantify the relative contribution of each of these types of bias, although one registered RCT proposes to do this (Cowgill and Dell'Acqua, 2018).

3.1 Unrepresentative training samples

Programmers' training data about a phenomena is often missing data for some applicants non-randomly. Economists are familiar with this problem through the sample-selection issues raised by Heckman (1979). For example, performance outcomes about job candidates who are not hired, or about loans applications that are rejected, could be missing for training data. Instead, outcomes may be available only for a non-representative selective group, which would result in biased predictions. As Brown (1978) notes, the more recruiters anticipate that educated workers are better, the harder it will be to find any evidence of this in the sample of hired workers. This is problematic when the goal is to develop an algorithm for a larger population (i.e., the set of all loan or job applicants) for use in screening.

Unrepresentative training data could come about either for taste-based or statistical discrimination by human decision-makers. Even accurate statistical discrimination would produce unrepresentative samples. For example, suppose that an employer predicted that test scores correlate strongly with employee performance and hired only good test-takers. The resulting sample of employees at the firm would be highly unrepresentative of the applicant pool. Should the firm give employee performance data to an engineer to train a hiring algorithm, this data would suffer from unrepresentativeness even if the employer's predictions were correct.

Cowgill (2018c) presents a theory model that endogenizes both the *selective labels* and *omitted payoffs* problems (discussed below). The paper then characterizes the space of human decision-making processes under which automated learning technologies can reduce or eliminate the underlying human biases.

In the model, human decision-makers generate a historical dataset by making a series of biased decisions (either from taste-based discrimination, or poorly-calibrated statistical discrimination) that generates both *selective labels* and *omitted payoffs* problems. The model suggests that machine learning algorithms can remove human biases exhibited in historical training data, but only if the human training decisions are sufficiently *noisy*.[26] Otherwise the algorithms will codify or exacerbate existing biases.

The key feature of this model is that better learning technology is complementary with greater experimentation. From the perspective of machine learning, noisiness in human judgment is ef-

[25] Mullainathan and Spiess (2017) and Varian (2014) review and introduce machine learning methods for economics audiences.

[26] At a recent NBER conference on AI and decision-making, Economics Nobel Laureate and psychologist Daniel Kahneman stated "We have too much emphasis on bias and not enough emphasis on random noise [...] most of the errors people make are better viewed as random noise [rather than bias]" (Kahneman, 2017). Kahneman has a longer article and book about the cost of noise in decision-making (Kahneman et al., 2016).

fectively a form of experimentation. This facilitates learning by exploring the space of alternative, less-biased decision-making strategies. Without sufficient noise, the learning technology will codify or exacerbate existing biases.

Given abundant evidence of noisiness in human decisions documented by psychologists and behavioral economists,[27] the model makes optimistic predictions about the effects of machine learning on bias – even in the presence of bias in the training data. The results suggests that learning technology needs only a small amount of noise to correct biases that cause large productivity distortions. As the level of human-related noise increases, machine learning can correct both large and increasingly small productivity distortions. However, the theoretical conditions necessary to completely eliminate bias are extreme, and are unlikely to appear in real datasets.

The Cowgill (2018c) model also suggests that the high levels of noise in decisions, a necessary condition for debiasing, actually harms traditional "goodness of fit" measures often used to measure model quality in practice among software engineers. If engineers and entrepreneurs use these metrics to guide their choice of applications, they will be lead towards applications most likely to codify, rather than relieve biases.

Taken together, these results have implications for the way that expertise interacts with machine learning. A variety of research suggests that the benefit of expertise is lower noise and/or variance, and that experts are actually *more* biased than non-experts (they are biased towards their area of expertise, Li, 2017).

If this is true, then the Cowgill (2018c) model of noise suggests that using expert-provided labels for training data in machine learning will codify bias because experts *experiment too little*. Even if experts are ultimately better than a non-expert at performing the job directly (as Li, 2017 finds), training algorithms using experts' historical data may not be as useful if the experts fail to explore.

3.2 Mislabeling outcomes in training samples or "omitted payoffs."

Conditional on appearing in a sample containing performance outcomes, some outcomes may be misleadingly labeled. For example, a worker who manages to be hired, but then faces discrimination by a supervisor, may be wrongly labeled as low-performing. An algorithm would associate their characteristics with low performance, even though it should be associated with discrimination. As with "biased training samples" noted above, mislabeling may also come about either for taste-based or statistical reasons.

"Omitted payoffs" also encompasses themes of the multitasking literature in contracting, which is often summarized as "you get what you pay for" (Kerr, 1975; Lazear, 1989; Holmstrom and Milgrom, 1991; Baker, 1992; Gibbons, 1998). In this same spirit, in supervised machine learnings "you get what you trained for." Excellent employee performance often requires a combination of easily measurable objectives and abstract goals and behaviors that are difficult to quantify.[28] Even if all important behaviors were measured, firms would have to specify trade-offs between them by

[27] For example, weather affects financial decisions (Rind, 1996; Hirshleifer and Shumway, 2003; Busse et al., 2015), sports victories affect financial decisions (Edmans et al., 2007) and relationships (Card and Dahl, 2011), stock prices affect decisions about effort, innovation, job candidates (Cowgill and Zitzewitz, 2008) and health (Engelberg and Parsons, 2016).

[28] For example, "organizational citizenship" behaviors (Meier, 2006; Bolino and Grant, 2016).

combining them into a single payoff function.

The difficulty of quantifying these goals creates problems both for incentive contracts and for algorithms. If employers offer contracts that pay only on easily measurable outcomes, they will substitute effort away from abstract goals. The same is true in supervised machine learning. If algorithms are trained only to predict outcomes that are easily measured, they will optimize these outcomes at the expense of other objectives. This exhibits one of several parallels between machine learning and mechanism design which we also discuss in Section 2.2.

3.3 Feedback Loops

Researchers about algorithmic bias have concerns about feedback loops and self-fulfilling prophecies (Lum and Isaac, 2016; Ensign et al., 2017; O'Neil, 2017). The critical point of these concerns is that many algorithmic prediction applications are not arm's-length, disinterested predictions; they are instead used to affect the outcomes they are supposed to "predict." The use of these outcomes either by a decision-maker herself, or by subjects responding to or anticipating to those decisions, affects whether the predictions are "correct."

Similar issues affect prediction in other domains. For example, corporate prediction markets (Gillen et al., 2017; Cowgill and Zitzewitz, 2015a) try to help executives forecast business outcomes. If managers use these forecasts, they interact with the reality the markets are designed to predict, which changes the informational content of the prices (Siemroth, 2019). A prediction market could therefore appear "wrong" to a naive *ex-post* observer, even if it has given managers highly actionable information.[29]

Algorithmic feedback loops go beyond this type of feedback. Not only are outcomes affected by predictions, but these tainted outcomes are then codified as "ground truth" for use in future algorithms. This may reinforce or amplify biases in the original predictions. One setting where this appears to be happening is credit scores and employment. A growing number of employers are using credit scores to evaluate job applicants.[30] The theory appears reasonable – if a worker is responsible enough to repay bills, they may be responsible enough to perform a job.[31] Even if true, believing credit scores predict job performance leaves low-credit workers without income, which further damages credit scores, which further damages job prospects, and so forth in a self-reinforcing loop. As discussed in Section 2, Arrow's 1973 classic work about "self-fulfilling prophecies" shows these feedback loops may happen even if statistical discrimination is accurate and unbiased.

Fudenberg and Levine (1993) note that it is possible for individuals to maintain *incorrect* beliefs about the payoff consequences of actions that have rarely been tried – for example, by hiring non-traditional candidates – and for these beliefs, in turn, to support suboptimal actions in equilibrium.

Causal inferences about "algorithmic feedback loops" are inherently difficult for empiricists. In

[29] For example: Suppose a market forecasts disaster with 90% probability. Managers react to this forecast by changing their plans, thus averting the disaster. The ex-post 90% forecast may appear wrong to a naive observer because the disaster was avoided, but it was premised on the state of the world before the intervention.

[30] Bartik and Nelson (2016); Clifford and Shoag (2016); Friedberg et al. (2016); Cortes et al. (2018) discuss the effects of this information on hiring.

[31] Koppes Bryan and Palmer (2012) evaluate how well credit scores correlate with on-the-job performance.

many cases, job applicants labeled "high expected performance" by an algorithm may be more likely to be hired anyway, even without digital labels. Attributing a hire to an algorithm requires a quasi-experimental intervention. Identifying a feedback loop requires a researcher not only locate such an intervention, but also that they must trace intervention as it propagates into codified outcomes as well as future actions and conclusions drawn from contaminated data.

To our knowledge, Cowgill (2018a) presents the only well-identified causal evidence of the "algorithmic feedback loop" phenomena. The setting is Broward County, Florida, where bail judges are provided predictions about defendants' recidivism using an algorithm derived from historical data.[32] The output of the prediction algorithm, called "COMPAS," was continuous. But the scores were shared with judges in rounded buckets (low, medium and high). Using the underlying continuous score, the paper examines judicial decisions close to the thresholds using a regression discontinuity design.

Defendants slightly above the thresholds spend an average extra one to four weeks before trial, which suggests the judges incorporated the signal into decisions. When bail is linked to outcomes, the extra pre-trial detention given to defendants above the thresholds corresponds to a small increase in recidivism within two years. This was the outcome the algorithm was originally designed to predict. Black defendants' outcomes were shown to be more sensitive to the thresholds than white defendants'.

These results suggest that algorithmic suggestions have a causal impact on criminal proceedings and recidivism. Showing this label to judges affects whether or not the original assessment was "correct" by traditional predictive-accuracy measures.

However, algorithmic labels not only affected defendant outcomes. They also affect future training datasets, future research conclusions and future algorithms. As mentioned earlier, the COMPAS dataset has been extensively used in computer science. Many papers featuring this data use re-arrest outcomes as "ground-truth" for new methods. They do *not* use the rearrest data as if it could be contaminated by upstream interventions by biased judges and their algorithmic guides.[33] These papers instead use re-arrest outcomes as if they were fair measures of criminality, untainted by judges' utilization of the very algorithm (COMPAS) these papers criticize. This completes the algorithmic feedback loop: The original COMPAS intervention affected judges' decisions, which affects re-arrest probabilities, which affects the training data used by researchers for future algorithmic development.[34]

[32] This is a setting of many papers about bias in algorithms, thanks to a high-profile investigation and report by *ProPublica*, a media outlet that shared the results of its FOIA requests for Broward's criminal data freely online (Larson et al., 2016).

[33] For example, several recent computer science papers evaluate new proposed algorithms or approaches, both purely algorithmic (Zafar et al., 2017; Corbett-Davies et al., 2017) and incorporating human discretion (Tan et al., 2018; Dressel and Farid, 2018).

[34] Because of the secrecy of the COMPAS algorithm, we cannot know whether Northpointe takes feedback loop into account in training the next generation of their algorithms. However, a 2014 government report (Austin, 2014) to Broward County policymakers recommended that "the COMPAS system could be easily replaced with a customized risk assessment scale [...] tailored to Broward County." If this happened, the recidivism outcomes caused by COMPAS could find their way into training datasets used for future algorithms. Similarly, we also do not know whether doing so would have a meaningful impact on their algorithm's suggestions. It is possible that even without corrections for feedback loops Northpointe's suggestions are more fair than a counterfactual judge. Arnold et al. (2018) uses random assignment to judges to suggest that human bail decisions, the same decision studied by *ProPublica*, are already biased. Like *ProPublica*, the Arnold et al. (2018) paper specifically examines county bail decisions from the Miami metropolitan area.

3.4 Biased Programmers

Software engineers may unconsciously (or overtly) exhibit bias during development. According to the Bureau of Labor Statistics in 2017, software engineers are more often white, male, well-educated and better-paid than America as a whole. These engineers may not be consciously biased, but their life experiences may influence their approach to developing an algorithm.[35]

"Biased programmers" could create both unrepresentative training samples as well as omitted payoff problems. In many practical settings, software engineers are responsible for assembling training data and formulating the computational problem as well as developing and implementing software. Mitchell et al. (2018) catalogues the choices made by software engineers as they build models. Biases of unrepresentative programmers could enter at any point. They could select training data from a familiar but unrepresentative setting. Or they may pay disproportionate attention to particular training examples, measures of accuracy, or empirical applications during development.

One alleged example of this may have taken place in 2015, when Google released a photos product that offensively mislabeled African Americans.[36] Had Google's product developers been more diverse, the problem may have been avoided. Separately, Blodgett and O'Connor (2017) find that machine translation tools by Google, IBM, Microsoft, Twitter and others translate text and speech by African-Americans and women worse than white males'. The authors highlight that blacks and women are a much larger portion of the American population (and of Internet users) than they are of engineering teams of these firms and computer scientists' workforce more broadly. Had these programmers been more diverse, the suggestion is that these problems would have been identified. They may also pay less attention to covariates in the training data that are correlated with vulnerable groups.

Anecdotal evidence for the biased programmers hypothesis suggests that it is not driven by deliberate animus but by programmers failing to consider their diverse audience.

Statistical evidence that causally links a programmer's identity to algorithmic behavior is rare. In a recent paper, Silberzahn et al. (2018) give twenty-nine research teams (61 individuals) an identical dataset about soccer. Each researcher was asked the question: "Are referees are more likely to give red cards to dark-skin players than light-skin players?" Answers varied widely across the researchers. Estimated effects ranged from 0.89 to 2.93 in odds-ratio units. The twenty-nine teams used twenty-one unique combinations of covariates. The authors did not report systematic differences grouped by the researcher's demographics, but the results suggest that programmers may reach widely different conclusions using the same data.

Several papers give reason for hope regarding biased programmers. Some economists explicitly model limited attention (Schwartzstein, 2014; Hanna et al., 2014) in general settings, finding that human prediction may be mis-calibrated because of cognitive constraints that focus attention on a limited subset of variables. Given the role of attention in the biased programmers hypothesis, this suggests that predictions can be better calibrated if delegated to agents (computers) that can attend to a greater number of variables. The Bordalo et al. (2016) model suggests that stereotypes come about partly because of human misunderstanding of the prediction problem – humans provide

[35] This could come about on account of either taste-based or statistical discrimination in the developer.

[36] (Forbes.com, 2015)

diagnostic variables when they are asked to make predictions.

Insofar as machine learning know what to program for, they may be able to avoid replicating this mistake. However, there are several other reasons to be less optimistic. Biased programmers may often face incentives to create algorithms that confirm prior beliefs. A famous example of this algorithmic ranking was in US News and World Report, which was repeatedly adjusted in order to conform to the public's images of great universities (Thompson, 2000). Many computer science papers validate models based on users labeling the classifications as agreeing with prior intuition. These problems suggest incentive contracting issues underlying the "biased programmers" problem.

3.5 Spillovers and Composition

Many economic settings feature interdependence between actors in a manner that leads to biased outcomes, even if each actor is individually unbiased. Computer scientists refer to this as "fairness under composition" (Dwork and Ilvento, 2018). Issues around spillovers and composition are particularly salient in digital economics.

A key example is found in Lambrecht and Tucker (2016), which studies how online advertising for STEM jobs are displayed differently to men and women. The authors examine two sets of online ad campaigns – one targeted at men and the other at women, but otherwise identical in their budget and targeting. The authors find that the STEM ads are shown to men more often. At first glance, the culprit might be thought to have been biased training data or algorithms that associate engineering jobs with masculinity for historical reasons. However, the authors' closer analysis reveals a different explanation. Women are less likely to see ads for job opportunities because of competition from other advertisers, particularly those selling consumer-packaged goods ("CPG"). Ad auctions for female eyeballs contain more bidders and thus have higher clearing prices.[37] This means that the STEM ads targeted to women are crowded out because of spillovers between industrial sectors. This does not happen in the markets for male eyeballs as often. As a result, otherwise identical campaigns reach fewer women because there is more competition for female eyeballs. Demand for female attention coming from cosmetics companies (and other sources) spills over to affect prices and quantities for STEM ads, effectively crowding them out.

This mechanism is arguably a form of "omitted payoffs"; many observers could view equality in job advertising as a payoff worth preserving. However, the auction-based allocation system attempts to maximize the profitability and efficiency of ad targeting; these are worthwhile goals in their own right, but they omit the goal of gender equality. Other allocation methods could produce the same outcome so long as they also respond to advertisers' preferences. Even if the underlying bidders are not biased, aggregation through the price system creates an unequal outcome.

The Lambrecht and Tucker (2016) research raises other policy issues. One frequent proposal for addressing bias is to make platforms gender-blind. However, gender targeting may sometimes be necessary to achieve gender balance. An employer could address the STEM jobs issue by bidding higher for women, guaranteeing that the STEM ads were shown ahead of CPG ads. Without gender targeting, it would not be possible to increase bids for women in order to out-bid CPG

[37] The phenomenon of "Female eyeballs are more expensive for advertisers" is a broader phenomenon that appears in advertising settings beyond the platform they study.

advertisers. But this too, creates other fairness issues: Recruitment strategy would feature unequal resources to recruit men and women. For this reason, Facebook actually forbids gender-targeting employment ads on its platform.[38]

Similar issues could explain the patterns in Larson et al.'s 2015 study of price quotes on the Princeton Review's website. When users enter a ZIP code from an Asian-American neighborhood, they are quoted higher prices. Spillovers may explain how this came about. Like female advertising inventory, Princeton Review classes are priced by the market. The supply of neither is perfectly elastic. If there is strong demand in Asian neighborhoods for classes, Princeton Review will have to allocate scarce slots among competing families. If they use the price system for allocation, then one family's demand for SAT classes may spill over onto others in the form of higher prices.

Using AI for matching in labor or other markets is similar to advertising. AI is often used to improve targeting for buyers and sellers. Several theory papers examine the effects of better targeting on markets. As targeting improves, markets may fragment into smaller, segmented pools of demand. These segments may contain higher match quality, but suffer from "thinness" because the quantity of participants decreases per segment. Thinness may complicate bargaining and price-setting by exaggerating negotiation power. Hummel and McAfee (2015) and Fu et al. (2012) examine the targeting/thinness tradeoff theoretically.

Cowgill and Dorobantu (2018) study the effects of greater targeting in ads using a differences-in-discontinuity design across geographic markets. They found that greater targeting options made ad markets thinner. Measures of the number of bidders per ad inventory decrease. As competition subsided, average prices decreased as well. However, the composition of ads also changed. The quantity of ads sold went up because new targeting enabled unsold inventory to find a buyer. The net effect on ad revenue was positive.

These results were not inevitable or a mechanical byproduct. They required advertisers to exhibit heterogeneity in preferences. If advertisers' target consumers were undifferentiated, targeting may not have affected market thickness and prices may have risen.

Similar phenomena could occur in other settings where algorithms improve targeting. The consequences may impact diversity and inequality directly. Cowgill (2017) found that the introduction of machine learning into hiring decisions led to more offers to non-traditional candidates. In a labor setting, these candidates are analogous to "previously unsold inventory" in the Cowgill and Dorobantu (2018) advertising market. Despite being overlooked by the firm's human recruiters, these candidates performed slightly better in interviews and on-the-job performance.

However, like the unsold ad inventory, these non-traditional job candidates faced a thinner market for their services. Non-traditional candidates identified by machine learning and rejected by human screeners were 15 percentage points more likely to accept job offers (when extended), and were 12 percentage points *less* likely to show evidence of a competing offer in negotiations surrounding offers.

The employers in question did not adjust wages substantially in response to these outside offers, leaving the traditional and non-traditional candidates equally paid in their jobs. However, many employers would engage in such bargaining (including universities and faculty). For employers

[38]This would arguably constitute a failure of "disparate treatment," the legal framework discussed in Section 2.4.

who do, wage inequality between traditional and non-traditional workers at the same company would increase. Note that inequality among overall job seekers may have decreased, since some applicants would be paid a positive wage in this industry rather than being paid nothing (the "unsold inventory" candidates who found a buyer). However, inequality among co-workers at the same company or job would increase because the algorithm selects qualified workers with fewer outside options.

These results suggest that machine learning identified valuable candidates who were not only overlooked by the human recruiters at one firm, but also by the entire remainder of the labor market. This took place in the market for programmers, a labor market so tight that employers lobby Congress for expanded H1B visas. Like the ad-targeting market, these results were also not mechanical or inevitable. It is possible that overlooked candidates were, in fact, adversely selected. If employers' demands were undifferentiated (i.e. a strictly "common-value" labor market), better targeting of candidates may have led firms to concentrate recruitment on a narrow group that is equally desired by all employers, and to *avoid* non-traditional candidates.[39]

Just as in the Cowgill and Dorobantu (2018) ad market, better targeting could have increased price pressure (wage pressure) in situations where employer preferences were undifferentiated. However, both papers find decreased price pressure, increased quantity and a shift in the composition composition of transactions. Given the role of AI in these outcomes, some of these results could be misunderstood as algorithmic bias. Like the high-priced female eyeballs in Lambrecht and Tucker (2016), the results in the hiring example are driven not only by algorithms but by inequality of competing offers. In both these cases, the effects of algorithmic selections depend on how they interact with with an existing set of outside circumstances which are *not* determined by an algorithm.

4 Strategic and Behavioral Economics Considerations

As mentioned in Section 2.1, decisions can fail the Becker test without directly using demographic variables. Similarly, they can pass the test, despite using such variables. Can we know anything about what variables will be used? In this section, we review what economic models say about the *content* of algorithms and their relationship with strategizing agents. We also discuss the application of behavioral economics to algorithmic fairness problems.

4.1 Economics Signalling Applications to Algorithmic Fairness

Statistical discrimination theory makes an affirmative prediction about which variables will be used to discriminate: Viable signals must be *costly* for low-quality job candidates to acquire (Spence, 1973). If a signal elicits favorable treatment but is cheap, all candidates will acquire the signal and it will lose its screening value.

The Spence model suggests that certain characteristics may be useful *only* for screening, and

[39] To our knowledge, Agan et al. (2018) contains the only empirical estimate of how correlated firms' demands for workers are and finds moderate sized correlations showing a large role for heterogeneous preferences between employers in the same industry.

otherwise offer no utility to employers. These characteristics are valuable insofar as they signal hard-to-observe traits. Spence used this concept to explain the popularity of educational credentials whose requirements are unrelated to job function. Diplomas are useful for signaling characteristics such as intelligence and commitment, even if they do not teach job skills. Several nuances in education and labor statistics suggest a role for signaling.[40]

Signaling theory has implications for algorithms. First, if algorithms do not use costly signals, agents can "spam" the system. Without costly signals, the usefulness of certain machine learning applications may be limited. In résumé screening, certain job signals (i.e., qualifications such as graduate degrees) are costly proxies for abilities and are verifiable. For other signals, such as the use of complex vocabulary words, may be easier for low quality candidates to adopt.[41] Already, researchers and journalists have reported the phenomena of "résumé stuffing" or adding strategic content to résumés in order to game humans or algorithms.[42] Similar issues have not stopped *actual* spam filters (for email) from success.[43]

For settings where manipulation is possible, strategizing may be cheaper for some than others. Pathak and Sönmez's 2008 analysis of the Boston mechanism for school choice showed that gaming the system by misrepresenting preferences could be profitable. However, this was probably exploited only by more sophisticated families (possibly wealthier or better-educated parents). Hu et al. (2018) formally modeled screening games in which some agents have a cost advantage for manipulating. The authors show how subsidies intended to equalize equilibrium outcomes may could make all groups worse off – including the group receiving the subsidy – while only improving outcomes for the learner.

In adversarial settings, the uninterpretable "black box" nature of machine learning is an advantage. Spamming a decision system if harder if its internal process is opaque. Cybersecurity researchers refer to this protection as "security through obscurity." Ederer et al. (2018) analyze this tradeoff formally, showing that deliberate opacity reduces gaming, and document the long intellectual history of this concept. Human decisions, which are also opaque in some settings, may also enjoy spam resistance for this reason.

Machine learning features an additional anti-gaming feature. A common practice in machine learning is to limit to how influential any single variable can become through regularization. This makes machine learning models more robust outside of training samples, but would also limit the effectiveness of gaming. For someone to change his or her classification, regularization requires a large number of variables to be modified. Insofar as human judgment can't "regularize," possibly because of limits on the number of variables they can process (Hanna et al., 2014; Schwartzstein, 2014), they can't enjoy this advantage.

Costly signaling and interpretability Signaling theory and related empirics suggest that many good screening variables may have no clear relationship to the employer's task. Many economists

[40] Some research suggests that workers learn little in school that is useful for employment. Caplan (2018) summarizes this perspective.

[41] Some claim these words are useful in screening (Weaver, 2017).

[42] *How To Trick The Robots And Get Your Résumé In Front Of Recruiters*, *Fast Company*, `https://www.fastcompany.com/40422836/how-to-trick-the-robots-and-get-your-resume-in-front-of-recruiters`, accessed November 14, 2018. For a scholarly discussion, see (Amare and Manning, 2009).

[43] Rao and Reiley (2012) review spam problems from an economics perspective.

suggest that high school and college curricula, which sometimes include trigonometry, cellular biology and world history, have little direct application in the vast majority of employment. For example, venue contracts for the 1980s rock band Van Halen included a rider requesting a bowl of M&Ms with no brown ones. The band later revealed it was a test – a costly signal for checking who read the band's elaborate concert instructions. Failing the M&M test triggered an inspection of the concert setup. Brown M&Ms were otherwise irrelevant.

Policymakers and businesspeople often request "explanations" for the output of machine learning algorithms (Guidotti et al., 2018) . The European Union General Data Protection Regulation (GDPR) and prior EU regulations asserts a "right to explanation" when influenced by algorithmic decision-making (Goodman and Flaxman, 2017).

However, many useful screening variables, such as diplomas or bowls of non-brown M&Ms, have little obvious relationship to performance objectives. If Van Halen's concert instructions were an algorithm, inspectors might have suspected a bug or spurious correlation when they discovered the M&M variable, but they would be wrong. If hiring baristas were algorithmic, inspectors may have questioned the trigonometry prerequisite (implicit in the high school degree requirement).

Economic theory and empirical studies indicate why such signals are already incorporated into screening. These characteristics are costly signals of hard-to-observe qualities. Machine learning can scale the search for non-obvious signals. However, this search can yield variables requiring further inspection for a human to understand. We are aware of no algorithmic approach that automates explanations for why variables such as brown M&Ms (or trigonometry requirements) are useful in screening staff. In the Van Halen case, disclosing the explanation would undermine the test's purpose. If the test were explained per GDPR regulations, then concert staff would realize they could pass the test by providing M&Ms without reading the instructions.

We discuss fairness in adversarial situations more detail in Section 7.2. One benefit of the Becker test is that it examines outcomes only. As this example shows, transparency about internal characteristics introduces gaming.

Costly signaling is also related to "recourse," a common perspective on regulating algorithms. Spangher and Ustun (2018) and others suggest that algorithms should have "actionable recourse" and rejected applications "should be able to do something." This appears to be motivated by the reasonable goal of eliminating demographic considerations from selection since these variables are costly to acquire. The concept of recourse creates a principled, abstract way to differentiate "sensitive" variables from others (without referring to contemporaneous law).

While this seems reasonable when applied to demographics, the Spence model and its successors suggest that it is possible for workers to have so much "recourse" that it will obliterate the usefulness of screening. Lack of recourse is not a bug in these models, but the entire purpose, allowing good and bad candidates to be separated.

Demand for recourse should must also be weighted against its costs.[44] Ultimately, "recourse" in many settings may be illusory.[45]

[44]Suppose NBA scouts are prohibited from discriminating on player height ("no recourse for being short"). Setting up tryout systems would be costly. While it may be worthwhile, the benefits must be measured against these costs.

[45]Using test performance to screen candidates may appear to offer recourse (pass the test). However, performance on certain tests may be heavily influenced by fixed characteristics. Tall players would probably win the NBA tryouts

Costly signaling theory provides a theoretical foundation for using seemingly-irrelevant variables in decisions. However, the theory is too broad to be used as a regulatory strategy to discipline bias. Can an employer claim that anything is an M&Ms-style costly signal? Can this loophole be used to mask harmful discrimination? How can a judge determine which signals are truly legitimate?

The Becker test (and other outcome-based tests) do not require regulators to examine inputs, and instead only look at the outputs. If the marginal male and female are not equally productive, the Becker test suggests that hiring is biased, irrespective of the internal characteristics of the algorithm.

Nonetheless, knowing whether seemingly-irrelevant variables are truly predictive is often valuable. A firm caught using biased practices may want to know how to screen. Courts have asked firms to demonstrate "business relevance," which can be established by showing a statistical correlation between a characteristic and performance outcomes. However, firms may lack clean variation in screening criteria in their historical decisions.[46] Field experiments in screening offer firms and regulators a way to address these questions.[47]

Contract theory and mechanism design Contract theory and mechanism design are subfields of economics concerned with strategic design of the message space for costly signals. A subfield of computer science examines the computational properties of mechanism design (Nisan and Ronen, 2001). In a typical mechanism design problem, an employer wants to screen certain workers to join a firm. She cannot directly observe worker quality or effort and therefore cannot contract directly on these variables. However, she may observe a noisy measure of output for workers who join the firm. She then can strategically design a contract to convert these signals into payments. Written optimally, this formula will induce only the right workers to join the firm.

As imagined by economists, contracts such as formulae converting performance metrics into payments are essentially algorithms. They convert costly signals into payments, and thus generate credible information through incentive design. There are several parallels between contracts and algorithms that we explore in more detail in Section 3. Several economists studying AI have backgrounds in contract theory.

4.2 Implications of Behavioral Economics for Algorithmic Fairness

Many attempts at statistical discrimination are *not accurate*. Behavioral economics is partly about systematic errors in beliefs and statistical reasoning.[48] These errors include confirmation bias (Ra-

anyway.

[46] It would be hard for Van Halen to show the value of the M&Ms test using observational data from their tours if there was no variation (or no unconfounded variation) in their screening methods.

[47] If Van Halen desired to test if brown M&Ms were a truly valuable costly signal, the band could randomly divide their touring venues into treatment and control. For treatment, they could utilize the brown M&Ms method for testing compliance. For control, they could remove the rider. They could then examine measures of success and compliance for both groups. If someone questioned whether brown M&Ms was a reasonable screening criterion, the band could share the results of this experiment.

[48] The other major part of this literature is about non-standard preferences and non-standard decision-making (Rabin, 2002).

bin and Schrag, 1999), framing (Tversky and Kahneman, 1981), overconfidence (Malmendier and Tate, 2008; Cowgill and Zitzewitz, 2015a), herding (Devenow and Welch, 1996) and recency bias (Kahneman et al., 1993; Cowgill and Zitzewitz, 2015b).

Two recent papers are especially relevant to algorithmic bias. Theories about *selective attention* (Hanna et al., 2014; Schwartzstein, 2014) suggest that humans statistical discrimination may be mis-calibrated because of cognitive constraints. In this setup, even sophisticated humans who correctly calculate their prediction weights may be biased because they do not use all relevant variables. This effectively creates omitted variable bias in the humans. The model suggests that AI may be able to reduce bias insofar as it can use more variables.

Bordalo et al. (2016) contains a statistical formulation of stereotypes. The authors explain their model using the example of elderly Floridians. If stereotypes were based on the most common characteristics, Florida would be associated with 20-44 year-olds. Age distributions in Florida are actually similar to the rest of the United States. The authors claim that Florida is associated with elderly *not* because $P(Old|Floridian)$ is high, but instead because $\frac{P(Old|Floridian)}{P(Old|Not\ \ Floridian)}$ is high (i.e., advanced age is a "diagnostic" variable for Floridians, distinguishing Floridians from non-Floridians). The authors show this pattern of stereotype formation explains many common group stereotypes as well as other nuances of stereotyping. In this literature, diagnostic variables are more likely to "come to mind" (Gennaioli and Shleifer, 2010), and are therefore weighted more heavily in predictions about the characteristics of Floridians.

Bordalo et al. (2016) and other data-generating models for stereotyping relate to whether algorithms will embody these stereotypes. They suggest that humans misunderstand their prediction task. Decreasing bias may be possible if programmers are given incentives to predict the right outcome (i.e., $P(Old|Floridian)$) without overweighting diagnostic variables) and delegate math to processors. While this seems like an easy requirement, our later section about "biased programmers" (Section 3.4) shows where software engineers face incentives to create algorithms that conform to prior beliefs (which may be distorted because of the issues above). Fudenberg and Levine (1993) note that it is possible for individuals to maintain incorrect beliefs about the payoff consequences of actions that have rarely been tried – for example, by hiring non-traditional candidates – and for these beliefs, in turn, to support suboptimal actions in equilibrium.

Among academic psychologists, the superiority of evidence-based algorithms for prediction is supported by decades of research in many empirical settings. This includes employee performance (Highhouse, 2008), student performance (Dawes, 1971, 1979), criminal defendants' recidivism (Thompson, 1952; Wormith and Goldstone, 1984), demand for products (Schweitzer and Cachon, 2000) and medical diagnoses (Adams et al., 1986; Beck et al., 2011; Dawes et al., 1989; Grove et al., 2000). Review articles include Dawes et al. (1989); Grove et al. (2000); Meehl (1954).

Much of this research uses decades-old statistical methods (for a methodological critique, see Cowgill, 2018b). However, a recent review summarized academic psychology's views: "When choosing between the judgments of an evidence-based algorithm and a human, it is wise to opt for the algorithm." Psychology has in fact moved towards documenting and understanding biases *against using algorithms* (or "algorithm aversion," Dietvorst et al., 2015, Dietvorst et al., 2016) despite their superior performance. A nascent literature in psychology and economics (Yeomans et al., 2017; Bigman and Gray, 2018) seeks to explain when and why humans defer to algorithms, with some researchers reporting boundary conditions to "algorithm aversion," particularly for

non-experts (Logg et al., 2018).

There are several related topics we do not discuss in this essay. Some uses of machine learning and AI have been used to uncover hard-to-observe human biases such as ideological bias (Jelveh et al., 2015). A small literature studies the design of marketplaces that neutralize discriminatory behavior (Edelman et al., 2017). Algorithms also influence inequality through the surplus captured by their designers. One source of rising inequality in the last several decades has been the wealth created by technology entrepreneurs. This essay mostly focuses on the direct influence of algorithms on fairness outcomes in markets, rather than through economic gains from owning them.

5 Computer Science Literature about Algorithmic Fairness and Bias

The topic of social biases in computer systems has an unexpectedly long history. To our knowledge, the first paper on this topic was Friedman and Nissenbaum (1996). Although the authors do not identify many of the specific issues arising today, the paper was prescient in many ways. Early empirical papers about algorithmic bias were straightforward: They ran a script to collect algorithmic outcomes and then measure differences between demographic groups using straightforward statistics. This evaluation of bias did not typically present comparisons to counterfactual screening methods.[49] If these papers went further it was generally towards building a practical software tool allowing further data collection about bias.[50]

In this section, we review several of the major groups of papers in computer science and related literatures about algorithmic bias.

5.1 Law and Computer Science

When the computer science literature has looked outside its borders, it has focused mostly on the law. Legally-binding tests for discrimination, established by judges, regulators and politicians, strike these researchers as a reasonable place to begin. This is consistent with the field's practical and engineering-oriented preference for algorithms that can be used in practice (without incurring lawsuits). These computer scientists are not suggesting optimal public policy; they are developing technology to comply with existing policy. This literature often takes legal requirements as exogenous, and attempts to express the requirements as constraints within mathematical optimization.

Regulators sometimes offer precise legal requirements that can be incorporated into algorithms. America's federal employment discrimination regulator, the Equal Employment Opportunity Commission (EEOC), guides employee selection to abide by a "four-fifths rule"[51] which means that the

[49] In the computer science theory literature, one exception is Chouldechova and G'Sell (2017), which develops a methodological framework for comparing two algorithms. The paper thus contains an implicitly counterfactual setup although it does not use counterfactual language.

[50] For example, the AdFisher project (Datta et al., 2015) documented that Google in India was more likely to show ads for executive coaching to men than women. The project was focused on building a tool which would allow others to automate the process of creating a gendered Google account and measuring whether it saw different ads from an account of a different gender. See Section 5 of Datta et al. (2015) for an example.

[51] This policy was implemented in the Uniform Guidelines on Employee Selection Procedures ("UGESP"), a set of

pass rate for all demographic groups must fall within 80% to 100% of the group with the highest pass rate.[52] These policies sometimes themselves lack thinking about equilibrium.[53] Passing or failing the 4/5ths tests has no relationship with passing or failing the aforementioned Becker -style tests of equal productivity of marginal candidates. The 4/5ths test does not require comparisons to other methods.

A foundational paper in the CS literature on algorithmic bias (Feldman et al., 2015) formalizes the "four-fifths rule," links it to pre-existing loss-functions studied within statistics and computer science, and provides practical and theoretical guidance for building complaint algorithms.

Consistent with their engineering orientation, these papers focus on protected classes of subjects as defined by current U.S. law and with the empirical properties of recent data. The "sensitive variables" are usually demographic; some theoretical papers (Corbett-Davies et al., 2017) discuss more vague but exogenously-defined "sensitive variables" and "legitimate variables" without endogenizing the labels.

5.2 Definitions of Fairness

Some papers go beyond legal requirements and examine other plausible definitions of fairness (Romei and Ruggieri, 2014; Žliobaitė, 2017). For example, suppose programmers develop an algorithm forecasting loan repayment for screening applicants. Should there be separate algorithms for black and white borrowers? If only one scoring algorithm is used, could there be different thresholds for each group? Several answers to these questions could be characterized as fair. Fairness may also require that model's predictive accuracy be equal across groups (Corbett-Davies et al., 2017).

The definitions in this literature frequently contain one or both of the following two features. First: They may involve direct preferences for diversity and/or representation. "Demographic parity," the requirement an equal (or bounded) representation all demographic groups, exemplifies this direct taste for representation and is the focus of several CS papers.[54]

Such preferences for diversity could plausibly be divorced from the computer science of machine learning altogether. Machine learning is broadly useful for predicting empirical outcomes such as worker productivities. However, the usefulness of these predictions – and the tradeoffs of predicted performance with diversity with other goals – is a question of utility functions. As Kleinberg et al. (2018) wrote, "A preference for fairness should not change the choice of estimator."

1978 federal policies adopted by the Civil Service Commission, the Department of Labor, the Department of Justice, and the Equal Opportunity Commission in part to enforce the anti-employment discrimination sections of the 1964 Civil Rights Act. They state, "A selection rate for any racial, ethnic, or sex group which is less than four-fifths (4/5) (or 80 percent) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded as evidence of adverse impact."

[52]For example, suppose a technology company adopted a job test. The most successful demographic group in this job test was Asian females, who pass at 45%. The "four-fifths rule" says that the job test is discriminatory if any group's pass rate was below 36% (45%×4/5).

[53]Companies could game the EEOC's rule by manipulating who took the test. Cowgill (2018b) contains additional economic and econometric analysis of the 4/5ths rule. Such critiques may be irrelevant to algorithmic developers or practitioners who simply aim to comply.

[54]The aforementioned 4/5ths rule and related papers (Feldman et al., 2015) are an example of demographic parity.

Second: Fairness definitions impose constraints *on the process of learning from data* to identify and correct for distortions. These learning adjustments are more connected to the traditional expertise of statistical learning. They could improve both the quality of estimators and performance, as well as diversity and representation, without an explicit taste for diversity or representation.

Many real-world actors, of course, have tastes both for performance and for diversity. However, the literature in this area is sometimes unclear about which combinations of goals are driving applications and definitions of fairness. This can be tricky; for example, Kleinberg and Raghavan (2018) formally analyze the "Rooney Rule," a practice of guaranteeing at least one interview slot to a minority candidate. This may appear to be driven by tastes for diversity and representation, but the authors identify circumstances where the Rooney Rule is profit-maximizing.

On the whole, this CS literature avoids advocating a single definition of fairness and embraces the subjectivity of the choice. The literature aims to guide policymakers' selection of fairness criteria contextually, by highlighting tradeoffs between various definitions (Kleinberg et al., 2016). For whatever the chosen criteria, the literature then aims to provide practical tools and tests for implementation and to document their computational properties.

The literature also aims to formalize each notion of fairness and to examine the (in)compatibility of each notion with each other and other goals for algorithms (such as maximizing predictive accuracy).[55] Importantly, many papers contain impossibility results showing that many attractive fairness criteria cannot be simultaneously achieved (Kleinberg et al., 2016; Barocas et al., 2017; Chouldechova, 2017; Berk et al., 2017; Narayanan, 2018c).[56]

Many of the fairness definitions used in this literature – including those used in the impossibility results – suffer from the problem of infra-marginality introduced by (Ayres, 2002).[57] For example, Hardt et al. (2016) introduces an "equality of opportunity" fairness criterion, which the authors operationalize as a selection algorithm that equalizes the rate of true positives across all demographic groups. Hardt et al.'s 2016 false-positives – and other fairness metrics – typically refer to the *entire* admitted population in each group rather than just the marginal candidates on the cusp of rejection (Corbett-Davies and Goel, 2018).

To some observers, the Hardt et al. (2016) discussion of separate thresholds for white and black borrowers in FICO scores may relate to the aforementioned Becker (1957, 1993) "outcome tests" for bias (Section 2). However, Becker's test requires equal *repayment rates* for black and white borrowers at the margin (i.e., at the threshold of the decision to borrow or not), which is not necessarily the same as equal FICO scores. If FICO scores do not accurately predict creditworthiness, then demographic-specific thresholds in machine learning may be useful for reducing bias.[58] Similarly, if FICO scores are biased, then adopting a single FICO threshold for all demographics may nonetheless fail the Becker test.

[55] Some researchers dispute tradeoffs with predictive accuracy, claiming any tradeoffs are an artifact of badly-defined performance criteria. See Section 4 and Section 3 about the choice of performance variables for training.

[56] For example, Berk et al. (2017) defines six notions of fairness, and concludes "some of [these] are incompatible with one another and with [maximizing predictive] accuracy." The authors claim that satisfying all six notions simultaneously is impossible outside of trivial instances. Narayanan (2018c) similarly examines 21 plausible definitions of fairness.

[57] This is also discussed in Simoiu et al. (2017).

[58] Demographic-specific thresholds may have other uses in affirmative action.

5.3 Engineering *ex-ante* Fair Algorithms

Some computer science literature measures bias or fairness of outcomes *ex-post*. A related, smaller literature prescribes new computational techniques for developing algorithms that are *ex ante* more likely to be unbiased (Lum and Johndrow, 2016; Johndrow and Lum, 2017; Aliverti et al., 2018; Kallus and Zhou, 2018). A few of such approaches are familiar to economists. Some make precise adjustments to standard algorithms based on a model of the bias (Calders et al., 2009; Kamiran and Calders, 2009; Feldman et al., 2015; Agarwal et al., 2018; Chen et al., 2018b). These are similar in spirit to economists' structural modeling, although in some cases without links to equilibrium or economic theory. Other CS researchers attempt to reduce bias with fewer assumptions by gathering more data in a quasi-experimental fashion, for example, by using multi-armed bandits (Jabbari et al., 2017; Joseph et al., 2016).

5.4 Interpretability and Explanations

Lastly, a nascent CS literature develops "interpretable" machine learning (Doshi-Velez and Kim, 2017). Guidotti et al. (2018) surveys the CS literature on explanations. These papers are motivated by the hope that interpretable algorithms and/or explanations can help assess whether an algorithm is unfair. This literature criticizes the "black box" character of machine learning, and aims to build algorithms that ordinary humans understand. Some of these are focused on providing explanations for algorithms as a whole (Wattenberg et al., 2016; Krause et al., 2017), while others are interested in providing specific explanations for each decision (Martens and Provost, 2014; Ribeiro et al., 2016; Koh and Liang, 2017; Lundberg and Lee, 2017).

Several papers show that machine learning algorithms can be vastly simplified without harming predictive accuracy. Examples include point systems (Jung et al., 2017; Ustun and Rudin, 2016), additive models (Caruana et al., 2015; Lou et al., 2012, 2013) or checklists (Malioutov et al., 2017). Kleinberg and Mullainathan (2018) proposal a formal framework for simplified interpretable predictions, and show that every simple prediction function is strictly improvable in efficiency, equity and welfare; i.e., identifying tension between the goals of simplicity and equity and other social outcomes. Although simplifications may be easier to explain to users, they may not actually increase performance, comprehension or trust in the algorithm. Simplification may allow an algorithmic designer to bury problems less transparently.

Psychologists have found that humans in lab experiments are less biased when they suspect they will have to explain their decisions. This could influence either algorithmic developers (see "biased programmers" in Section 3.4) or human judges receiving guidance from algorithms. A separate line of research suggests that humans are excellent at rationalizing decisions, i.e. providing false explanations – this could limit the effectiveness of interpretable machine learning on "biased programmers" or the audience for their products.

Do interpretability requirements mitigate the biased effects of algorithms? What are the costs or benefits? Given the interest in interpretable machine learning, surprisingly few papers attempt to answer this question scientifically. Poursabzi-Sangdeh et al. (2018), present large-scale field experiments in which functionally identical AI models are shown to users with varying levels of interpretability. They find that interpretability indeed improved subjects' ability to simulate the

models predictions. However, they found no effects on measures of trust in the algorithms, or in subjects' performance in detecting sizable mistakes in the algorithm. The benefit from simple explanations may be what psychologists call "the illusion of explanatory depth" (Rozenblit and Keil, 2002).

Economics may have useful insights into the role of explanations in algorithms. A long literature examines strategic communication, featuring a "sender" who has better information than the receiver (Crawford and Sobel, 1982; Farrell and Rabin, 1996). In the setting of machine learning, the "sender" is the algorithm.

The receiver of the information is a humans relying on algorithmic advice. In the canonical setup, communication is possible only insofar as principle and agents' interests overlap, i.e., insofar as they are playing a coordination game. As the interests of sender and receiver diverge, only very coarse or no communication is possible.

Critically, these models say that the inability to communicate is *not* determined by shortcomings of the message space. It is not because of the lack of well-crafted sentences, data visualizations, check lists, historical examples or other innovations to "explain" decisions. It is because of two parties' inability to trust each other. As interests diverge, all communication becomes self-serving, manipulative, unverifiable "cheap talk." If interests were completely aligned (and this was common knowledge to both parties), no explanations would be necessary, only recommendations.

The challenge for explainable algorithms is therefore to align the interests of human and machines, and to make this alignment common knowledge. Do algorithms and their human interpreters strategic interests' diverge? An algorithm is an inanimate object having no strategic interests. However, algorithms are designed and owned by people who do. The interests of agents who design and own algorithms may not completely align with the principals receiving recommendations. Most obviously, if the software team aims to replace the principal's job with software or take credit for success, this would limit the AI's trustworthiness.

Preferences could misalign for other reasons. For example, humans and machines may face different penalties for Type I and II errors. A judge who wrongly acquits a murderer may face different punishment than the software executives whose algorithm guided the judge's decision. The same is true for wrongful convictions. Humans and and software owners may additionally have asymmetric ability to evade responsibility for these mistakes, again leading to differing interests and coarsened communication.

AI and human clients' interests may additionally diverge is if the AI vendor serves multiple clients simultaneously, including competitors. Alignment raises obvious industrial organization and organizational economics questions about vertical integration and incomplete contracting (Varian, 2018; Hadfield-Menell and Hadfield, 2018).

Several papers show that communication can improve if the sender can obtain a costly signal of verification (Kartik et al., 2007; Kartik, 2009; Halac and Yared, 2016). However, most attempts at "explainable AI" fail the "costly signal" criterion. The entire premise of explainable AI is to make explanations cheap by automation at low marginal cost.

The models above suggests that trust in AI may increase if software businesses can credibly align their interests with their clients', for example by enabling formal or repetitional risk-sharing. An

AI vendor's instincts to avoid responsibility for clients' failures is short-sighted; it shows clients that AI firms can avoid the downside risk of mistakes and therefore cannot be trusted.

Just as importantly, this requires alignment to become common knowledge. Even of software/client interests are aligned, this may not be obvious to rank-and-file human operators. These rank-and-file operators are often the people whose trust in AI is necessary for everyday business decisions such as processing job or loan applications.

For these reasons, explanations and interpretability, at least as as currently imagined by computer scientists, may not be effective. We further discuss the merits of transparency, interpretability and explanations as policy solutions to bias in Section 7.

5.5 Counterfactual Fairness

A recent strand of computer science research focuses on causal or counterfactual models for fairness (Kusner et al., 2017; Chiappa and Gillam, 2018; Spangher and Ustun, 2018). Kusner et al. (2017) assess fairness in algorithms by asking the following question: If a candidate's characteristics were counterfactually different, would an algorithm's suggestions change? If the suggestions would change in response to sensitive variables changing, this is interpreted as unfair.[59]

An attractive property of this concept is that a regulator can look at the code and weights of a scoring algorithm to assess whether changing candidate characteristics (gender, race, age, etc) would result in counterfactually different decisions. However, this counterfactual approach may be easy to evade, for example by leaving sensitive variables out of the model altogether. This would create the appearance of no effects. However, because many variables in large, modern datasets are correlated with the sensitive ones the algorithm may nonetheless utilize demographic categories (see Section 7.1). Counterfactual fairness as proposed above is a form of *input regulation*, a type of regulation with strengths and weaknesses we discuss more generally in Section 7.

The above notion of "counterfactual fairness" critiques algorithms by characterizing changes that would counterfactually affect an algorithm's decision. Some researchers go beyond this, claiming that algorithms should be designed from first principles to model the causal effects of worker characteristics on real-world productivity.[60] In this literature, firms should set a target productivity variable to maximize, and select applicants who have been treated with worker-level interventions that *causally* impact their productivity. For example, programming classes could causally impact a worker's ability to write software, but altering a worker's skin tones should not.

This approach would seem compatible with empirical economics. Counterfactual and causal inference approaches are central parts of econometrics. However, there are several differences with econometrics' perspective on counterfactuals. Bias against women and minorities is a clear motivating factor for algorithmic fairness. The role of fixed, unalterable personal characteristics (such as demographic variables) in the Rubin causal model underlying econometrics is unclear (Greiner and Rubin, 2011). "No causation without manipulation" is a mantra in this approach (Holland, 1986). The Rubin causal model that underlies causal inference focuses on measuring effects of ma-

[59] A related paper by Chiappa and Gillam (2018) pursues a similar approach. Spangher and Ustun (2018) expands the scope of excluded variables to "non-actionable" variables – any that a person cannot choose to change.

[60] Kusner et al. (2017) write, "Fairness should be regulated by explicitly modeling the causal structure of the world."

nipulating the world. Variables such as a worker's education could in theory be manipulated (for example, by school admission lotteries that randomize access to education). However, it is unclear whether and how somewhat fixed demographic features can be experimentally manipulated, or what this would mean.

For example, altering appearances early in life may *cause* workers to become less productive if the changes causes the subject to face discrimination in education (for example, in Arrow's 1973 "self-fulfilling prophecies" in Section 2.2).[61]

Even for variables that can be manipulated, using these variables raise separate fairness issues. Suppose that high quality causal inference studies demonstrate that education causally increases worker performance. The counterfactual fairness approach suggest that education is therefore a legitimate variable to use in screening. However, this does not entirely eliminate fairness concerns. Education causes productivity increases, education itself is not equally accessible across demographic groups.

Modeling the causal the structure of the world, even for variables that can be manipulated, is impractical for most applied settings. Causal models require clean sources of exogenous variation in the form of experiments, random shocks or interventions (deliberate or natural). Such randomization is rare, and lack of clean identifying variation frustrates many important research subjects, for example in macroeconomics (Nakamura and Steinsson, 2018).

Finally, the Spence (1973) model of costly signaling suggests a productive role for costly variables for *signaling reasons only*. Van Halen's fussy M&Ms do not *cause* the staff at concert venues to become effective. Instead, they signaled which workers read Van Halen's elaborate instructions. Similarly, many forms of education may *not* cause workers to become more productive. They may instead signal or credentialize abilities workers possess. Such variables are useful for decision-making, but would be excluded from the counterfactual approaches noted above. The implications of costly signaling for algorithms is discussed in more detail Section 4.1.

Counterfactual thinking can be applied to evaluating fairness in other ways, which we discuss in Section 7.3. Defined above, "counterfactual fairness" is about worker-level characteristics.[62] Rather than focusing on the effects of changing *counterfactual individual characteristics* of a person, Section 7.3 focuses on the effects of counterfactually changing *selection processes*, leaving personal characteristics fixed.

6 Correction of Human Bias by Algorithms

Despite ample concern, there are many reasons to believe that algorithms – including simple algorithms, naively trained from observational data – may correct or reduce human biases. We begin with an overview of the empirical evidence. We then review some theoretical reasons why these may be the case, particularly for applications experiencing growth in interest for this technology.

[61] Kohler-Hausmann (2018) critiques approaches to detecting racial discrimination based on counterfactually manipulating individuals' fixed characteristics.

[62] Causal coefficients about these characteristics are useful for comparing someone to his/her counterfactual self (without a treatment). They are not necessarily well-suited for comparisons between people with different fixed characteristics.

Then we discuss why there is so much alarm about algorithmic bias despite these results, and why some researchers feel that human comparisons are inappropriate.

6.1 Empirical Evidence Comparing Algorithmic and Human Bias

Several empirical economics papers suggest that algorithms decrease bias compared to human decision-makers. There are few such evaluations, but those we have conclude that algorithms reduce relative bias (Kleinberg et al., 2017; Cowgill, 2017; Dobbie et al., 2018a). As we describe in Section 7.3, these comparisons to human judgment (or pre-existing algorithms) are useful for policymakers and practitioners to evaluate bias.

Cowgill (2017) contains a field experiment in overriding human discretion with algorithmic judgment about employers' decisions about whom to interview. The experiment follows candidates into productivity realizations in their next jobs. Although the algorithm was trained on historical data, it *increased* hiring for several historically underrepresented groups at the firm.

Similarly, Kleinberg et al. (2017) develop an algorithm for predicting recidivism. To compare their algorithms' performance against human judges, they construct a simulation by exploiting the random assignment of judges to cases. The simulation models judges perfectly complying with the algorithm's guidance.

Although the counterfactual is simulated, their results suggest positive real effects on final outcomes, including prison sentences, crime and recidivism. Their simulations suggest large welfare gains from reduced crime (up to 24.8%) with no change in incarceration rates. The authors suggest prison populations can be reduced by 42% with no increase in crime rates. The authors also show that "the algorithm is a force for racial equity," allowing judges to incarcerate 40.8% fewer minorities without increasing the crime rate.

Dobbie et al. (2018a) use a similar strategy to measure bias in lending by using random variation in the assignment of loan examiners to applicants. The authors find significant bias against both immigrant and older loan applicants. Using simulations similar to Kleinberg et al. (2017), they find that a decision rule using machine learning can simultaneously eliminate bias and increase profits.

Fuster et al. (2017) examine the effects of better statistical technology on lending using a structural model of both loan rates as well as decisions on loan applications in equilibrium. The authors find that the machine learning models are more effective at predicting default, and extend credit to a slightly larger fraction of mortgage borrowers. This slightly reduces cross-group dispersion in positive decisions on loan applications. However, they also find increases in the differences in interest rates across groups, particularly by widening the interest rate gap between white and non-white borrowers. These may be driven by actual differences in the underlying probabilities of default and thus not the result of *bias* by the models. However, these differences may be troubling for other reasons.

The above papers examine algorithmic biases vs human bias, or against simpler algorithms. A related set of papers studies the effect of human judgement-guided, rather than replaced, by algorithms.[63] Stevenson (2017) examines how judges are influenced by algorithmic guidance in

[63] Cowgill (2017) contains some additional results about human screeners' willingness to defer to algorithms, and

pretrial release decisions. Using sharp pre/post variation around 2011, when algorithmic scoring in Kentucky became mandatory, she finds that the Kentucky results "should ease (but not eliminate) concerns that risk assessment tools will exacerbate racial disparities." However, the findings also suggest the guidance barely changed anything else. Most judges overruled the algorithmic guidance, and only small changes in average pretrial release outcomes were detected towards the beginning of adoption. Failures-to-appear and pretrial crime increased as well, but only by small amounts. These changes eroded over time as judges returned to their earlier habits. By contrast with the results from Kleinberg et al. (2017), based on simulations of fully replacing discretion with algorithms, these results suggest a role for limiting judicial discretion.

These papers complement a pre-existing literature on simpler forms of algorithmic selection such as job testing. Autor and Scarborough (2008) finds productivity increases from the introduction of job testing, but no effects on minority hiring. Hoffman et al. (2016) finds similar results from limiting discretion.

6.2 Theoretical Reasons Algorithms could Reduce Bias

This paper has covered several theoretical reasons why the use of algorithms may reduce human biases: Stereotypes, noise, settings where various issues don't appear, and inputting more refined variables that reduce the coarseness of the analysis.

Interest in machine learning is growing not only because of technological progress in IT, but also because of novel demands on decision-making. Agrawal et al. (2017) discuss at length the microeconomic effects of improved prediction. The arrival of the Internet dramatically changed job search by decreasing barriers to applying, resulting in employers overwhelmed by the quantity of job applications (Oyer and Schaefer, 2011).

Such trends create demand for lower-cost, more scalable decision-making. Importantly for bias researchers, if machine learning is *not* adopted, decisions will made by human decision-makers who simply lack the time to invest in researching each case.

Scholars studying hiring find recruiters spend 16 seconds or less on each résumé (Bartoš et al., 2016; Lahey and Oxley, 2018).[64] Scholars have documented bail judges spending on between two and three minutes average per case.[65]

Psychology and behavioral economics research has highlighted that settings where there is limited attention are where humans are most likely to resort to shortcuts and heuristics, some of which (stereotypes) may have adverse consequences along fairness lines (Bartoš et al., 2016). Gender dif-

also finds that human screeners are highly deferential to the algorithms' recommendations. When the algorithms' conclusions are hidden from human screeners, the screeners reject approximately 30% of machine-approved candidates. However, when they are told which candidates were approved by the machine learning algorithm, nearly all of the machine-backed candidates are approved.

[64] Lahey and Oxley (2018) note, "This is slightly higher than, but comparable with, the estimate of 15 seconds often given by human resource professionals when asked (Lahey, 2008)."

[65] Stevenson (2018) writes, "While there is no systematic survey of the length of bail hearing, they are reported to be very short in many jurisdictions: three minutes long in North Dakota (VandeWalle, 2013), less than two minutes in Cook County (Staff, 2016) and only a couple minutes long in Harris County (Heaton et al., 2017)." Austin (2014) and Cowgill (2018a) also measure between two and three minutes in Broward County, Florida, the setting of the COMPAS dataset used in many algorithmic bias computer science papers. Dobbie et al. (2018b) also report two to three minutes.

ferences in hiring appear strong in the earlier stages of screening (featuring high-volume and low attention) rather than in later stages after candidates are narrowed to a short list (Botelho and Abraham, 2017).

Famous social psychology papers about "the illusion of transparency" suggest most people overestimate how well their own behavior is understood by other people (Gilovich et al., 1998).

6.3 Politics and Law of Algorithmic Bias

The optimistic results above are not inevitable. They depend on parameter values in the environment, institutional settings and other details. However, theoretical and empirical evidence from several perspectives and settings suggests optimism about this technology for reducing bias.

Nonetheless, a new movement critiques machine learning applications in business and government decision-making on fairness grounds. The movement is fueled by the belief that decision-makers are rushing to adopt algorithmic systems before considering the social consequences. These critics claim that automating decisions using analytics, code and training data creates a false veneer of science and objectivity. O'Neil (2017) reminds readers, "Algorithms are opinions embedded in code."

The optimistic evidence described above – a product of formal theory and high-quality empirics using causal inference methods – is more than a veneer. Nonetheless, this quote contains an element of truth. In Section 2, we discuss the subjectivity of the "objective" functions used to inform or to evaluate decision-making. In addition, algorithmic tools are marketed by businesses facing less intellectual accountability than peer-reviewed academics and their controlled experiments. "You can't lie with math," wrote physicist Sabine Hossenfelder (2018). "But it greatly aids obfuscation."

Over-deference to quantification may also apply to critiques of algorithms. The aforementioned ProPublica article contained quantitative analysis containing barplots, logistic regressions, survival plots and statistical significance stars. ProPublica's report was initially widely accepted by both the public and academics; the results are still used to motivate computer science papers about fairness. However, later analysis of ProPublica's data by independent academic researchers has called the original conclusions into question (Doleac and Stevenson, 2016; Kleinberg et al., 2016; Flores et al., 2016; Chouldechova, 2017). As we discuss in Section 7, a 2018 Reuters story alleging bias in Amazon's hiring algorithm contained similar appeals based on numerical argument, and was similarly broadly embraced in policy discussion despite weaker claims than ProPublica. Hossenfelder's argument about obfuscation also applies to critiques of algorithms.

Over-deference to numbers may be particularly bad for algorithms. The code and weights of an algorithm can be inspected in ways that human decisions cannot. Recall from Section 2 how difficult Becker tests on marginal candidates are in non-algorithmic settings. In addition, machine learning applications are held to higher standards for bias under U.S. law than humans. In *Wal-Mart vs Dukes*, a 2011 US Supreme Court case, 1.6 million female workers sued Wal-Mart for gender discrimination in pay and promotion. Although the case had nothing directly to do with algorithms, the Court's decision created higher liability standards for machine vs. human biases, even though each has the same practical effects.

The Court ruled in Wal-Mart's favor, arguing that Wal-Mart's personnel policies were delegated to branch managers who enjoyed substantial discretion (Dessein, 2002).[66] As a result, plaintiffs were not subject to a common injustice necessary to certify a class action lawsuit.

Through the lens of *Dukes* (and related laws and precedents), machine learning may be conceived of as a form of organizational centralization (Alonso et al., 2008) that increases an organization's vulnerability to class action lawsuits. In settings ranging from employment to insurance claims, legal scholars and practitioners have cited *Dukes* as a reason *not* to deploy machine learning, irrespective of the quality of machine decisions (or their impact on vulnerable groups).[67]

As Justice Scalia wrote in his opinion for the Court in *Dukes*, "The whole point of permitting discretionary decision-making is to avoid evaluating employees under a common standard." To fairness advocates, "common standards" are a positive feature, arguably the entire goal of their movement. However, the Court effectively punished firms for creating common standards and implicitly allowed greater legal protection to bias – so long as it arises from human discretion.

The Court's decision weaponized the decentralized biases of human managers, enabling them to continue influencing business decisions throughout the economy, while creating additional liability for using machine learning systems.

The *Dukes* decision has been criticized for making class certification difficult without prohibitive amounts of discovery (Weinstein, 2011). Discovery for the 1.6 million plaintiffs against Wal-Mart's decentralized human system would be costly even if forthcoming and reliable records were kept. By contrast, an algorithm would be much easier to investigate through electronic discovery.

Similar issues have arisen in the student adoption of electronic medical records ("EMRs"). If EMRs improve quality of care, then they lower the likelihood of malpractice claims (Studdert et al., 2006). However, EMRs may increase the legal discoverability of details of patient care. These details offer doctors legal defenses if care was correct (Miller and Glusko, 2003). EMRs can serve as a "smoking gun" (or presented as a smoking gun) in situations of uncertainty (Korin and Quattrone, 2007) and subject hospitals to invasive searches and/or lawsuits, even if care were properly administered. According to the industry press, "lawyers smell blood in in electronic medical records" (Mearian, 2015).

The data about about EMR adoption suggests that EMR adoption actually *decreases* or has no effect on malpractice claims (Quinn et al., 2012; Virapongse et al., 2008; Victoroff et al., 2013). Healthcare IT adoption also reduces malpractice resolution times (Ransbotham et al., 2019) which are costly for patients and providers and reduces the number, severity, and disposition time of claims, while having no effect on the amounts paid (Ransbotham and Overby, 2010).

[66] Writing for the majority, Justice Antonin Scalia claimed that "On its face, [Wal-Mart's decentralization policy] is just the opposite of a uniform employment practice that would provide the commonality needed for a class action. It is a policy against having uniform employment practices." For the women suing Wal-Mart, Scalia wrote there is no "common answer to the crucial question *why was I disfavored?*"

[67] As Crews (2018) wrote, "The implication of Dukes is that policies that are more centralized, and applied uniformly, enhance the likelihood a class will be certified." "Algorithmic hiring" resembles similar legal cases in which employers administer a standardized test – itself a crude form of an algorithm – to applicants. These classes are typically certified. Empirical studies by Autor and Scarborough (2008) and Hoffman et al. (2016) suggest that these tests reduce bias against minorities and lead to more productive workers. Crews (2018) continued, "This suggests that a Big Data algorithm, applied uniformly and consistently throughout an employer's workforce, potentially provides the 'glue' [for class action lawsuits] missing from *Dukes*."

Nonetheless, "discoverability" and "smoking gun" fears have already deterred EMR adoption. Miller and Tucker (2012) examined the impact of state rules facilitating the use of electronic records in court, concluding that hospitals are one-third less likely to adopt EMRs after these rules are enacted. Similar issues affect the regulation of algorithmic bias: Greater measurability and litagability of alleged bias may deter adoption of technology despite its overwhelming potential to reduce overall bias.

6.4 Benchmarking vs Human Judgment

Given the counterfactual orientation of most economists, comparisons to human decision-making come naturally. If decisions are not made by algorithms, they will be made by someone else. With this in mind, it is obvious to many economists to compare which method minimizes bias.

Whatever the shortcomings of a particular algorithm, a rushed and "predictably irrational" human judge in the *status quo* could perform in a more biased manner and, in addition, fail to explain their decisions. Nonetheless, the very premise of comparisons to human bias is controversial among some researchers and policy advocates. These critics interpret human comparisons to encourage AI engineers to narrowly improve the *status quo*, stop prematurely, be congratulated for that achievement and fail to undertake additional fairness improvements. A typical response from this perspective is, "instead of just focusing on the least-terrible existing option, it is more valuable to ask how we can create better, less biased decision-making tools[.]"[68]

Our reading of the economics of AI suggests this is unlikely. In some settings, firms' economic incentives are well-aligned with reducing bias. Reducing bias would improve selection of productive workers, protect the firm from lawsuits and possibly net some public relations benefits from employing a more diverse workforce. As we outline in Section 4, many of the impediments for reducing bias are human cognitive shortcomings in statistical reasoning, which is something which algorithms are a complementary technology for.

Implementing machine learning involves large fixed costs of capital, labor and complementary organizational practices. Once these fixed costs have been paid on Version 1.0 of an algorithm, costs of improvements lower. The size of these fixed costs alone demand non-trivial improvements from Version 1.0. In addition, fixed costs paid to reach Version 1.0 can be leveraged repeatedly in upgrades to address fairness and bias. This makes it more likely that firms would take advantage of lower upgrading costs that may reduce bias and productivity even further.

Premature stoppage is most tempting *before* the large fixed cost of digitization are paid, not afterwards. Firms already face many non-digital ways to "narrowly improve the status quo, then stop prematurely and be congratulated," for example, diversity training programs that may have little impact. Remaining pre-digital is particularly tempting for employers because firms can more easily evade detection when hiring policies are not codified. As discussed in Section 6.3, public policy such as *Dukes* scrutinizes machine bias more severely.

There is little research about reactions to algorithmic fairness rhetoric, or about whether critics of advocates of algorithms benefit more from the halo of numeracy. The little research we have suggests the public is predisposed towards algorithmic fairness rhetoric. Pew Research Center's

[68] https://www.fast.ai/2018/08/07/hbr-bias-algorithms/, discussing Miller (2018).

2017 *Automation in Every Day Life* survey of four thousand Americans included several questions about hiring algorithms. Approximately 70% of respondents reported being worried about these applications, with only 3% expressing enthusiasm.[69] 73% reported that hiring algorithms would be bad or neutral for diversity. A survey of American public opinion by Zhang and Dafoe (2019) found similar skepticism about AI, particularly about bias in hiring and criminal justice. Whatever the merits of the public's views, the Pew findings and other evidence betrays the animating narrative of the algorithmic fairness movement ("rush to adoption without considering the social consequences.")

The 2018 outrage around Amazon's purportedly biased hiring algorithm also suggests the public is willing to believe stories about biased algorithms despite dubious numeracy. As we argue in Section 7 below, Amazon executives became concerned for misleading reasons. Amazon canceled the program altogether rather than fix the algorithm to "create better, less biased decision-making tools."

The fairness rhetoric intended to encourage thoughtfulness about the social content of machine learning caused Amazon to abandon its application of machine learning altogether, rather than leverage the technology's documented potential to reduce bias. Amazon reportedly resumed using human screeners, which many empirical studies across disciplines have show are biased.

7 Policy Implications

Regulation could play a useful role in thwarting algorithmic bias. The encouraging empirical analysis we have discussed so far are not inevitable. They depend on practical implementation decisions and parameter values of the environment that could differ in other settings. It would be foolish to claim that all algorithms improve fairness; surely someone could write a terrible one.

In this section, we examine policy proposals through the lens of economics and make suggestions of our own. Two particular themes reappear multiple times in our discussion.

The first theme comes from the economics of crime. Citizens can respond to policing either by increasing compliance, or by increasing efforts to evade detection. Even completely innocent parties may prefer not to facilitate inspection by police or regulators, who are capable of false-positive errors against the innocent.

Evading detection is particularly relevant to algorithmic bias. Human decision-making resists audits and electronic discovery. For reasons discussed in Section 6, using algorithms for decision-making increases the measurability of bias – both actual measurability as well as perceived.[70] Firms who want to evade inspections of bias possess a powerful tool: Let the humans decide.

Algorithms may frequently reduce bias, but make the remaining bias more visible, more inspectable and easier to target. Policies that do not grapple with this will fail to reduce unfairness, and will instead compel firms to hide bias behind less transparent processes. There, it will be harder for everyone to inspect and correct (both regulators and the firm itself). These policies may

[69] *Americans' attitudes toward hiring algorithms*, http://www.pewinternet.org/2017/10/04/americans-attitudes-toward-hiring-algorithms/, accessed November 11, 2018.

[70] Kleinberg et al. (2019) arrive at a similar perspective independently.

actually increase bias, even if the intent was to reduce it. Punishing algorithmic bias disproportionately to human bias reminds us of Kaplan's 1964's "streetlight effect," in which a man loses his keys in a park, but searches only near a streetlight because that's where the light is. Digital systems are where fairness is most easily measurable, not necessarily where it is most lacking.

The second theme comes from environmental economics. In environmental regulation, product and labor (and related fields), economists prefer to regulate firms' outputs rather than directly regulate technology. Environmental solutions such as carbon taxes, cap-and-trade and vehicle emission standards are forms of "output regulation." "Output regulation" is the policy equivalent of bosses telling workers, "Work however you'd like, as long as you finish your work on time."

By contrast, environmental "input regulations" (sometimes called "command and control" regulations) require firms to adopt specific technologies. As one textbook summarizes, "The simplest kind of regulation is to just tell people what to do."

By contrast, outcome regulation features less direct instruction, and allows firms decide how to comply with output requirements most efficiently given their circumstances. Without this flexibility, many firms claim they have special circumstances and therefore lobby for exceptions and loopholes that complicate the rules.

Command-and-control policies raise the possibility of firms complying with the letter of regulatory requirements while failing to make a difference. Output regulation allows direct contracting on the variables regulations are meant to change, and outsources the details of compliance to the private sector. This creates incentives for entrepreneurs to invest in technology for meeting standards. These investments may create technology that enables firms to go beyond the limits set by the government. With command and control, there are no incentives for an individual firm to ever go beyond what the government has asked.

The most popular ideas for regulating algorithms are input regulations. Asking firms to remove sensitive variables from algorithms altogether is a form of input regulation. Calls for "transparency" or interpretation of coefficients are also forms of input regulation.

Discrimination policy for algorithms could instead focus on outputs. There is already regulation of bias in many countries. While these regulations do not always follow economists' emphasis on marginal candidates, they are frequently outcome-based in other ways. As discussed in Section 2.4, "input regulation" of human decision processes is inherently difficult (although some policymakers try). Ludwig et al. (2018) formalizes output regulation, and traces how the incentives created by output regulation affect the downstream choices of machine learning engineers.

Although most economists do not advocate for government regulation of inputs, analyzing inputs is practically useful. Inspecting inputs is a useful diagnostic; they may help a single company – or an industry – develop best practices and innovations. Recall from Section 2 that the Becker test (and other pure outcome tests) are not constructive – they don't provide instructions to how to create good outcomes, only how to test for them. Innovations about the inputs and internal architecture of algorithms provide useful guidance about passing these tests. Researchers should continue developing new algorithms based on inputs and architecture improvements. The argument here is simply that *government regulations* should focus on outputs.

Lastly, as we mention in the introduction, the profit motive is aligned with reducing bias. Profit-

oriented firms should be naturally cooperative with reducing bias. The benefits of de-biasing can be at least partially privatized, which makes a sustainable, for-profit marketplace possible, particularly in the presence of outcome regulations that preserve incentives for production and innovation in bias-reducing technology.

As we discuss policy, a useful analogy may be the molecular structure of pharmaceuticals. Opening an "FDA for Algorithms" is a commonly-suggested policy solution for algorithmic fairness (Tutt, 2016). Even if pharmaceutical companies were forced to disclose molecular structures to patients, it is not clear that it would be helpful. What is important for understanding the efficacy of a drug is concrete data on how the drug interacts with its human environment. The FDA evaluates this using randomized controlled trials, which is the approach we explore for algorithms. For many popular drugs, the exact molecular mechanisms are still unknown (Lewis, 2016). "Explanations" of the mechanism of the drugs are lacking, even to pharmaceutical researchers. Discovering the mechanism for these drug is, in fact, a proposed application of machine learning (Keiser et al., 2009). Nonetheless, the FDA has gained confidence in the drugs through experimental trials.

7.1 Regulating Inputs and Algorithm Architecture

Most policies aimed at correcting algorithmic bias effectively regulates inputs, including regulation around the underlying training data and how particular variables are used.

Including/Excluding "Sensitive" Variables As we discussed in Section 2.4, directly regulating the content of preferences or statistical reasoning is often infeasible for human bias. No regulator could tell a human recruiter, "Pretend you cannot see this variable and that you do not care about it." By contrast, regulating the input variables in computer algorithms is both technologically possible, and has political and legal support.

Statistical discrimination theory shows how this form of direct regulation may be ineffective, even when technologically feasible. If regulators could force decision-makers to ignore demographic labels, decision-makers could shift weights to other variables correlated with group membership. The shift may not affect the demographic composition of hires, only superficial appearances about how decisions were arrived.

Agan and Starr (2017) provides field experimental evidence of how statistical discriminators shift in response to omitting variables. Their evidence comes from "ban the box" policies forbidding employers from using criminal histories variables in hiring decisions. The authors show that employers responded to the ban by discriminating on other variables that were correlated with criminal history – including race. Removing information about the presence or absence of criminal history made hiring more race-sensitive.

Similar phenomena could arise from forbidding sensitive variables in algorithms. Miller and Tucker (2017) show examples of algorithms attempting to infer ethnic affinity based on affection for cultural products. However, the authors show these algorithms ultimately conflate income with ethnic affinity because income is also correlated with tastes for these products.

Permitting demographic variables in algorithms may be particularly useful in settings such as health applications, where the effectiveness of medications could vary across groups. As a practical

matter, even accurate statistical discrimination on the basis of protected demographic variables is illegal, although it may be difficult for regulators to detect and police.[71]

Pope and Sydnor (2011) suggest a method for using sensitive variables in algorithms. The authors suggest using demographic variables, but only in the *statistical learning process* that uses the historical record. For decision-making about new candidates, these variables would be held *constant or excluded*. The intuition behind the idea is that many useful non-sensitive characteristics of workers – for example education levels – may be correlated with demographic variables that are unacceptable to use. Including these "socially unacceptable predictors" in statistical learning therefore helps ensure that the coefficients attached to acceptable predictors are accurate (and do not reflect correlations with demographic variables).

Excluding these variables creates classic omitted variable bias that affects the weight placed on other, seemingly benign variables. If "years of education" is correlated with a demographic variable, then excluding that demographic variable from predictions changes the weight placed on education. The weight on education will now reflect the predictive content both of education as well demographics.

While the authors suggest *including* demographic variables in statistical learning, they suggest *excluding* them from decision-making about new candidates. This would proceed by holding sensitive variables constant across all scored individuals and instead using only the acceptable variables (with the coefficients derived from the learning process above). Yang and Dobbie (2018) incorporate these and related ideas into a legal framework.

7.2 Transparency

Transparency is a popular policy response. Many requests for transparency require developers to publish an algorithm's functional form, input variables and numeric weights. The most frequent goal is to attempt human interpretation of coefficients. This is so common and such a bad idea, that we have a separate section about it next. As we elaborate upon shortly, transparency through interpreting coefficients can yield highly misleading conclusions.

Even without interpreting coefficients, bias issues in algorithms may not be detectable from code and numeric weights. In many settings we have discussed, the effects of algorithmic selections depend on how they interact with external factors such as a set of outside offers, pre-existing practices or competition from other providers.

Lambrecht and Tucker (2016), discussed in Section 3.5, show that gender-neutral ad campaigns for science jobs reach fewer women. This appeared at first to be algorithmic bias coming from biased historical data, but the true culprit turned out to be the high demand for advertising to women. This makes the price of reaching women with any message – STEM jobs or otherwise – more expensive. This source of the distinction would be unclear from examining the algorithm's

[71] As Agan and Starr (2017) summarize, "Disparate treatment based on race violates employment discrimination laws, whether it is based on accurate statistical generalizations or on inaccurate stereotypes. Title VII of the Civil Rights Act of 1964, which prohibits race and sex discrimination in employment, does not permit otherwise-illegal treatment be based on group generalizations, even if they are empirically supported. For example, in City of Los Angeles Dept of Water and Power v. Manhart, 435 U.S. 702 (1978), the Supreme Court held that an employer could not rely, in designing pension benefits, on the actuarial prediction that women live longer."

code, which simply sought to minimize costs to the advertiser.

Algorithmic transparency has other downsides that have arisen in other economic settings. Transparency helps competitors tacitly collude (Albæk et al., 1997; Luco, 2018; Thomas et al., 2018). Some economists suggests that machine-learning algorithms may achieve collusion more easily than humans even *without* the advantage of transparency because of the machines' faster ability to learn through trial and error (Ezrachi and Stucke, 2016; Calvano et al., 2018).

Legally-mandated transparency is particularly useful for collusion. Competitor firms may perceive voluntary disclosures by rivals as manipulative "cheap talk" whose intent is to mislead (Baliga and Morris, 2002). By contrast, legally-mandated transparency compels truthful revelation. It could therefore offer a more credible focus for collusion.

Earlier results from the labor market (Cowgill, 2017) suggest a role for AI in increasing competition for workers. Allowing greater collusion may undo this. The level of concentration of the US economy has become a concern of economists (De Loecker and Eeckhout, 2017). The collusion argument against transparency may be substantially weaker for naturally monopolistic firms who don't face competition anyway.[72]

Transparency also has consequences for innovation. If innovations can be copied by competitors, the incentives decline (Bhattacharya and Ritter, 1983). The intellectual property system is intended to protect against such copying, but some firms find that secrecy offers better protection (Cohen et al., 2000; Arundel, 2001). Insofar as machine learning is a productive area of private-sector innovation, transparency may depress incentives.

Finally, transparency and explanations have security implications. The request for explanations for AI decisions resemble cybersecurity debate about giving governments special privileges over user data ("backdoor access"). In both cases, the government wants privileged methods to monitor bad behavior in the form of access or "explanations" of decision-making algorithms.

But in both cases, transparency can compromise the security of the larger system. As discussed in Section 4.1, if Van Halen disclosed the explanation for their M&Ms practice, it would allow staff to evade the good practice Van Halen was seeking to encourage. Ederer et al. (2018) formally show that deliberate opacity reduces gaming. A recent computer science paper shows that explanations can be used to reconstruct the underlying model, thus enabling hacking (Milli et al., 2018b). The robustness of a system has its own fairness issues, particularly if skill at manipulation is unevenly distributed (Hu et al., 2018). Milli et al. (2018a) explore the tradeoffs between robustness to manipulation and fairness and other social welfare considerations. In the case of cybersecurity, computer scientists have a strong negative consensus against backdoor access. Although transparency and explanations also feature security risks, they appear to be more popular among computer scientists at the current writing.

Interpreting Coefficients A common complaint about machine learning algorithms is their "black box" nature. Statisticians have long known that outside of a few specific settings – for example, well-identified regressions – the coefficients, weights and other internal properties of algorithms

[72]Many companies extensively using machine learning today are Internet platforms with strong network effects. Some have described these as natural monopolies. Bajari et al. (2018) discussed returns to scale in data, and implications for market power and concentration.

do not have clear interpretations. Nonetheless, misguided researchers, policymakers and journalists attach meaning to the internal numeric weights of algorithms.

We begin with an example that is familiar example to economists. Fixed effects estimators require that one fixed effect be dropped. The choice of the omitted fixed effect affects the numeric values and interpretations of all other fixed effects. This is typically not important for econometric estimation. If fixed effects were used to score or rank job candidates, the choice of omitted variable would not affect relative comparisons.

However, in the arms of lawyers or journalists opining about algorithmic discrimination, variables "receiving negative weight in an algorithm" have enormous, undeserved rhetorical heft. Similar issues occur in non-fixed effect settings.[73] Computer scientists typically do not think much about this choice, and it is not clear it is a good use of their time. Either approach could lead to the same ranking and distances between candidates. However, the optics of one versus the other could matter in a policy debate or legal proceedings, and a company, regulator or journalist could actually make the mistake of attributing significance to the approach taken.

In addition, a 2018 report from *Reuters* alleged Amazon built a machine learning application for screening resumes that was biased against women.[74] As evidence of bias, *Reuters* presented coefficient interpretations featuring the issues above. The Amazon hiring algorithm "[D]owngraded graduates of two all-women's colleges, according to people familiar with the matter." The article did not report how the algorithm scored other colleges, including the other 45 women's colleges.[75] Nearly all colleges (male, female and co-ed) may have been downgraded, depending on which "college fixed effect" was omitted.

In addition, Amazon "penalized resumes that included the word *women's*, as in *womens chess club captain*." Even if chess leadership were correlated with programming skill,[76], the question for gender bias researchers is not whether "women's chess captain" was penalized. It's whether "women's chess captain" is penalized more than "men's chess captain."[77]

Reuters' account shows Amazon's executives interpreting numeric coefficients and making social interpretation in an unsound way. The Amazon episode highlights the trouble of regulation through coefficient interpretation and other "input regulation."

[73] For example, an algorithm for scoring job candidates could proceed by giving every candidate a high initial score, and subtracting points depending on qualification. In this framing, even strong candidates would lose some points. Alternatively, an algorithm could start with a low score, and add points for qualification. Even weak candidates could earn a few points for their accomplishments. Other approaches could start with a moderate score, and points could be either added or subtracted.

[74] "Amazon scraps secret AI recruiting tool that showed bias against women," https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[75] According to *Women's College Coalition*, an association of womens colleges and universities, there were 47 all-female colleges in the United States and Canada in 2014 around the time of the development of Amazon's application: https://www.womenscolleges.org/history, accessed Nov 14, 2018.

[76] Research about chess players suggests chess skill is not as correlated with other analytic skills as intuition suggests. Studies of chess players suggest that even among grand masters, chess skill is *not* easily portable to other domains; this is true even in settings closely resembling chess (Palacios-Huerta and Volij, 2009; Levitt et al., 2011). Evidence on the "portability of expertise" (Green et al., 2017) is limited more generally.

[77] Arguably, both should be penalized compared to a the gender-neutral "chess captain," which suggests leadership over larger, gender-inclusive group.

The allegations against Amazon contained no evidence of a productivity test of the algorithm's marginal male and female candidates (the test in Section 2). Nor did it contain any attempt to compare outcomes under the resume-screening AI to an alternative. Even if the Amazon algorithm penalized women meaningfully, Amazon's human screeners may have penalized women applicants even more severely. After all, Amazon's system was trained on historical human recruiter's decisions. As we review in Section 6, empirical tests from several settings containing a clean counterfactual comparison show an increase in diversity, including one in hiring (Cowgill, 2017).

Knowing details about an algorithm's internal characteristics does not provide insights about the difference between two selection regimes. This is formally derived in Cowgill (2018c) although the intuition is simple. Suppose we implement an algorithm that evaluates job applicants to replace human recruiters, and the model places a negative coefficient on being a woman. This may appear to be a bad algorithm that discriminates against being female. But the algorithm could lead to increased female hires, if it replaces human recruiters who are even more sexist and less persuasive to change.

Cowgill (2017) contains a real-world empirical example of this: The resume screening algorithm in the paper appeared to give negative weight to candidates from non-elite schools. However, the candidates *benefiting* from the algorithm included disproportionately *non-elite graduates*. Human evaluators assessed these credentials even more negatively.

The same effect works in the opposite direction. An algorithm that appears to *help* a certain group (based on its numeric weights) might actually have a negative effect on that group. Suppose an algorithm predicts a loan applicant's probability of repaying a loan successfully, and places a strong positive weight (directly or indirectly) on a particular minority demographic variable. If loan officers (or the *status quo* regime) place higher weight on this attribute, the introduction of the algorithm may *reduce* minority lending despite the positive weight.

Weights and other internal characteristics are simply irrelevant for measuring the impact of an algorithm against an alternative. These examples demonstrate how transparency and interpretability provide misleading intuition about the effects of an algorithm. The Amazon story exemplifies how the numeracy bias is used *against* algorithms in ways it can't be used against human recruiters. What "penalty" did Amazon's human recruiters attach to women's colleges? What negative numerical weight did they attached in their calculations? We may never know because the human formula was not codified, probably deliberately as a result of *Dukes* and similar policies (Section 6).

Reuters' Amazon story was based on internal leaks. Amazon provided no comment. There may have been more to the story than published. Nonetheless, the media and the algorithmic fairness community largely embraced the sexism interpretation. To our knowledge, this essay contains the only publicly expressed skepticism about Reuters' interpretation of Amazon's hiring algorithm.

Amazon embodies the scenario where algorithms have their best chance against the numeracy bias of algorithmic fairness rhetoric. Amazon is a wealthy company with world-class expertise in machine learning. It is is precisely the type of company that *should* be able to mount an effective defense. Amazon nonetheless canceled the program – a decision that may have been rational if the company anticipated future lawsuits and/or bad press from a less technical audience that would

not understand the technical issues above.

7.3 Experiments

"Disparate impact" is a major branch of U.S. anti-discrimination law. "Impact" is explicitly causal language that implies a comparison between counterfactual strategies. However, regulations about discrimination do not require any counterfactual comparisons. Major government policies about bias explicitly state that no experiment or counterfactual evaluations are necessary.[78]

Policy is instead focused discovering bias through yes-or-no tests. There are several reasons that firms and policymakers should concern themselves with *reducing* bias below current levels, expecting that some remaining bias may fail yes-or-no tests. Policies aimed at reducing bias would require measurements comparing the relative bias of two strategies, rather than yes-or-no tests.

As discussed in Section 2.4, comparative tests are useful in part because firms who fail tests for discrimination need to learn a specific new strategy to improve. If statistical discrimination is perfectly accurate (as suggested by neoclassical models), then firms guilty of taste-based discrimination already know how to comply with policy. Once a government punishment offsets their tastes, they will know exactly whom to hire to comply efficiently.

Alternatively if firms are imperfect statistical predictors, then firms' responses to policy are unclear. The behavioral economics literature referenced in Section 2.2 show many settings where humans are indeed inaccurate, biased predictors. Even computer scientists cannot guarantee unbiased predictions in many practical settings in which high-quality training data is unavailable. "Fairness in machine learning" is an active research in topic computer science because of these tensions. Formal proofs in Cowgill (2018c) and Friedler et al. (2016) suggest that an absolute removal in bias is exceedingly unlikely using real-world datasets. As computer scientist Arvind Narayanan (2018a; 2018b) writes, "Bias in machine learning is the rule, not the exception." It is unclear how human cognition avoids the problems above – the behavioral economics literature contains ample evidence suggesting humans are not unbiased statistical predictors.

As a result, all firms' choices of strategies may fail yes-or-no tests of bias. The practical question is which contains the least bias and the most acceptable combination of unfairness. It is in this context that field experiments and machine learning may show the most promise: Not for eliminating bias, but for reducing it.

"Datasets of convenience" are used in many machine learning applications because of their ease of acquisition at the expense of representativeness, unbiasedness and other desirable features. They are an often-cited reason for algorithmic bias. However, perfectly representative data may be practically impossible to acquire, even at great expense. A fruitful area of research may be to better understand when and how historical samples can be used productively, even if they exhibit non-zero bias (Section 5.3 reviews early research in this area). The theoretical and field experimental results in Cowgill (2018c, 2017) suggests that even very simple methods may work well at reducing bias.

[78] The *Uniform Guidelines on Employee Selection Procedures*, a set of federal procedures for enforcing the employment discrimination sections of the 1964 Civil Rights Act, state in Section 14B "These guidelines do not require a user to hire or promote persons for the purpose of making it possible to conduct a criterion-related study."

Counterfactuals through Experimental Manipulation Although causal inference methods are widely used in other disciplines (including most experimental sciences), few empirical computer science papers have used these methods to examine algorithmic bias and fairness. Many of the decisions influenced by machine learning in hiring, admissions, lending and/or criminal justice may have enormous consequences, and are therefore taboo as the topic of experiments. Oyer and Schaefer's 2011 review of employment research notes that field experiments in hiring are rare, asking "What manager, after all, would allow an academic economist to experiment with the firm's screening, interviewing or hiring decisions?"

Beyond taboos, such experiments may raise ethical problems for researchers. Experimenting in decisions such as hiring, lending and/or criminal justice presents distinct ethical issues from those in medical trials, where the ethics literature is more established.[79] These are beyond the scope of this article, but deserve further attention. Some ethical considerations may be mitigated through careful experimental design.[80] Researchers can evaluate algorithms using "nudges," or interventions that allow non-compliance by the subjects.[81] As with medical trials, failing to rigorously test potential solutions to problems through experimentation is itself an ethical issue (Hellman and Hellman, 1991; Darrow et al., 2015).

Because of these issues, many researchers lack tools and intuition for measuring the causal effects of new algorithms. Opening a "FDA for Algorithms" is a commonly-suggested policy solution for algorithmic fairness (Tutt, 2016). Randomized controlled trials are a key component of FDA regulation, but few commentators have specified how this aspect of pharma policy would apply to algorithms. A full discussion of experimental design is beyond the scope of this essay, but Cowgill (2018b) contains a practical guide to designing and executing field experiments motivated at identifying bias in machine learning.

Experiments (and other causal-inference strategies) are inherently counterfactual in nature. However, our approach differs from other discussions of "counterfactuals" in the computer science literaturea about fairness (discussed in Section 5.5). Rather than focusing on the effects of changing *individual characteristics* of a person counterfactually, the experiments we have in mind focus on the effects of counterfactually changing *selection processes*, leaving personal characteristics fixed. The evaluation proceeds by measuring how the characteristics and performance outcomes of selected and rejected candidates change in response to changes in screening criteria.

This approach has the advantage of permitting real-world empirical verification of the models. Researchers can setup trials – field experiments and A/B tests – to test the policy by modifying

[79] In medicine, experimentation may help improve outcomes for future patients who are identical to today's research subjects. However, experimentation in hiring could discover that some of today's workers should not have been hired, thus making some subjects worse off. This may be justifiable if others workers are better off and/or if hiring is overall more efficient or equitable. Nardini (2014) reviews related ethical issues in clinical trials in medicine. Mislavsky et al. (2018) discuss the ethics of corporate experiments.

[80] For example, in the hiring context, a key question is about how non-traditional candidates would perform if hired. In some cases, they could be hired as a test without crowding out traditional candidates. This would allow a firm to pay for the knowledge of how certain types of candidates work, without harming other subjects. Alternatively, researchers could use adaptive trial design (Barker et al., 2009; Palmer and Rosenberger, 1999) such as multi-armed bandits (Joseph et al., 2016; Jabbari et al., 2017; Dimakopoulou et al., 2017) to balance exploration and exploitation. Kallus (2017) studies "instrument-armed bandits," i.e. bandits with noncompliance.

[81] These experiments would measure treatment effects only on compliers, but this is useful for assessing situations where algorithms guide decisions rather than force them.

screening policy. By contrast, researchers cannot easily alter candidate characteristics randomly, holding all others constant.

Without an experiment, a company, regulator or researcher could not know how much algorithm-identified candidates *were going to be admitted anyway, by an alternative or pre-existing process*. Presumably, many algorithm-approved decisions would have been made by humans operating independently. This possibility has been documented or alleged in several context ranging from hiring to judicial decision-making.[82] Without experimental variation, nobody could tell how much outcomes were truly *caused* by the use of an algorithm or were present anyway.

Unappreciated Benefits of Experiments We conclude with a few important points about the value of experiments in this field. First, experiments allow easy comparisons between algorithmic and non-algorithmic decision-making processes, and for researchers to examine fairness effects without having access to the underlying code.[83] This is because experiments are entirely outcomes-based.

In addition, the second counterfactual is harder to defeat. An adversarial programmer can change the variables used in the algorithm, thereby avoiding sensitive variables and instead using variables that are correlated with demographics rather than the demographic variables themselves. However, if these adjustments admit the same set of people, they will not address underlying concerns about fairness or bias.

Second, experiments allow feedback loops to be partialled out. We discuss feedback loops in Section 3.3. An algorithm's label may cause a negative outcome, which could then be used as training data and to reinforce labels. However, these negative outcomes may have happened anyway. Experimental variation would use a control group of randomly selected subjects with similar covariates who are not subject to an algorithm's intervention. This would not stop the feedback loop for the treatment group, but would allow researchers to measure the true extent of the loop so that an intervention can be justified.

Third, experiments can be used to examine the value of incorporating specific variables into screening algorithms. We mentioned this earlier in Section 4.1. To assess if GPA is a useful screening tool, a company could randomly screen some job applicants using GPA and others ignoring GPA. If the firm was already using many other measures of academic ability, adding the GPA variable may change very little. Insofar as it changes the pool of selected candidates, a firm could assess its effects by randomly hiring (or interviewing) candidates affected by the change.

Fourth, experiments also offer some of the benefits of explanation, transparency and interpretability. If a company were considering a new resume screening algorithm, a well-designed experiment would identify which subset of candidates' admissions outcomes would hinge on the new algorithm (this may be a small group, because presumably the new and old methods would agree on many candidates). Once these candidates were identified, an experiment could measure the

[82] Many judges and courtroom observers allege that algorithmic recommendations are ignored by judges, or that judges independently ask for the same variables that are fed into bail-setting algorithms. Cowgill (2018a) contains a collection of real-world examples and allegations.

[83] A recent NSF call for proposals asks, "What are ways we can study data-driven algorithms when researchers don't always have access to the algorithms or to the data, and when the data is constantly changing?" http://trustworthy-algorithms.org/cfw.html.

performance outcomes of these candidates. This would permit a statement such as: "The new algorithm selects candidates who are 5% more productive in their first quarter of employment than the alternative system. It also changes the composition of incoming admits by increasing the proportion of women and candidates with economics degrees."

One could perform this analysis for every sub-group in the applicant pool, enabling reports about every group such as, "The marginal black applicant admitted by the algorithm (but rejected by the human) has a 65% success rate once hired." Such reports may give practitioners or policy-makers fine-grained explanations of why the algorithm was being adopted, and what mechanisms were responsible for the improved outcomes, without many of the burdens of transparency, coefficient interpretation and other "input regulations" discussed above in Section 7.1. These results may also be useful for defensive purposes in lawsuits.

8 Conclusion

As algorithms' impact increases, engineers, policy-makers and businesses will need guidance about how to deploy and regulate them in ways that minimize harmful side effects. While engineers have responded to these needs with computer science research, many of the underlying issues are economic.

Economic theory has a long tradition in which human decisions are modeled "as if" they are mathematically optimizing (Friedman, 1953).[84] The formalism of economics and the underlying assumption of rational prediction has drawn criticism from other social scientists and humanists, who argue that much of human behavior cannot be formalized (Morson and Schapiro, 2018). The same critique is also levied at algorithmic decision-making (Powles and Nissenbaum, 2018). Either way, formalism makes economics theory natural complement to designing and evaluating decision-making algorithms. In addition, economics utilizes a relatively sophisticated empirical toolkit centered around causal inference. These empirical methods have several applications in algorithmic bias that complement existing approaches from computer science.

This essay has developed an economic perspective on algorithmic fairness. Algorithms are not only growing in their scope and potential impact; they also help researchers overcome measurement and research challenges, and may facilitate new insights into centuries-old topics economic research in familiar areas such as education, labor and crime. As we have discussed throughout the paper, the measurability of algorithms presents both research opportunities and policy challenges. These issues will make bias and fairness in algorithms increasingly relevant to a variety of economic literatures. A well-designed algorithm can be an enormous force for social good and scientific progress.

[84] Although economics is often associated with profit-maximization, inherently economics focuses on *utility-maximizing* more generally, where utility can come from many sources besides profit (friendships, votes, others' wellbeing). In this approach, utility functions must be formally specified, but doing allows a formal constrained optimization approach which often characterizes each human's decision as a formula.

References

Aaronson, Scott, "Why philosophers should care about computational complexity," *Computability: Turing, Gödel, Church, and Beyond*, 2013, pp. 261–328.

Adams, ID, M Chan, PC Clifford, WM Cooke, V Dallos, FT De Dombal, MH Edwards, DM Hancock, DJ Hewett, and N McIntyre, "Computer aided diagnosis of acute abdominal pain: a multicentre study.," *Br Med J (Clin Res Ed)*, 1986, *293* (6550), 800–804.

Agan, Amanda and Sonja Starr, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *The Quarterly Journal of Economics*, 2017, *133* (1), 191–235.

_ , Bo Cowgill, and Laura Gee, "The Effects of Salary History Bans: Evidence from a Field Experiment," *Working paper*, 2018.

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach, "A reductions approach to fair classification," *arXiv preprint arXiv:1803.02453*, 2018.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb, "Exploring the Impact of Artificial Intelligence: Prediction versus Judgment," 2017.

Aigner, Dennis J and Glen G Cain, "Statistical theories of discrimination in labor markets," *ILR Review*, 1977, *30* (2), 175–187.

Albæk, Svend, Peter Møllgaard, and Per B Overgaard, "Government-assisted oligopoly coordination? A concrete case," *The Journal of Industrial Economics*, 1997, *45* (4), 429–443.

Aliverti, Emanuele, Kristian Lum, James E Johndrow, and David B Dunson, "Removing the influence of a group variable in high-dimensional predictive modelling," *arXiv preprint arXiv:1810.08255*, 2018.

Alonso, Ricardo, Wouter Dessein, and Niko Matouschek, "When does coordination require centralization?," *The American economic review*, 2008, *98* (1), 145–179.

Amare, Nicole and Alan Manning, "Writing for the robot: How employer search tools have influenced résumé rhetoric and ethics," *Business Communication Quarterly*, 2009, *72* (1), 35–60.

Angrist, Joshua D and Jörn-Steffen Pischke, *Mostly harmless econometrics: An empiricist's companion*, Princeton university press, 2008.

Archibald, C., A. Altman, M. Greenspan, and Y. Shoham, "Computational Pool: A new Challenge for Game Theory Pragmatics," *AI Magazine*, 2010.

Archibald, Christopher and Yoav Shoham, "Modeling billiards games," in "AAMAS" 2009, pp. 193–199.

Arnold, David, Will Dobbie, and Crystal S Yang, "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 2018, *133* (4), 1885–1932.

Arrow, Kenneth J., *THE THEORY OF DISCRIMINATION*, Princeton University Press,

Arundel, Anthony, "The relative effectiveness of patents and secrecy for appropriation," *Research policy*, 2001, *30* (4), 611–624.

Austin, James, "Evaluation of Broward County Jail Population: Current Trends and Recommended Options," 2014.

Autor, David H and David Scarborough, "Does job testing harm minority workers? Evidence from retail establishments," *The Quarterly Journal of Economics*, 2008, pp. 219–277.

_ and Susan N Houseman, "Do Temporary-Help Jobs Improve Labor Market Outcomes for Low-Skilled Workers? Evidence from "Work First"," *American Economic Journal: Applied Economics*, 2010, pp. 96–128.

Ayres, Ian, "Outcome tests of racial disparities in police practices," *Justice research and Policy*, 2002, *4* (1-2), 131–142.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki, "The impact of big data on firm performance: An empirical investigation," Technical Report, National Bureau of Economic Research 2018.

Baker, George P, "Incentive contracts and performance measurement," *Journal of political Economy*, 1992, *100* (3), 598–614.

Baliga, Sandeep and Stephen Morris, "Co-ordination, spillovers, and cheap talk," *Journal of Economic Theory*, 2002, *105* (2), 450–468.

Barker, AD, CC Sigman, GJ Kelloff, NM Hylton, DA Berry, and LJs Esserman, "I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy," *Clinical Pharmacology & Therapeutics*, 2009, *86* (1), 97–100.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan, "Fairness in machine learning," in "Conference on Neural Information Processing Systems, Long Beach, CA" 2017.

Bartik, Alexander and Scott Nelson, "Credit reports as resumes: The incidence of pre-employment credit screening," 2016.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace, "Consumer-Lending Discrimination in the Era of FinTech," *Unpublished working paper. University of California, Berkeley*, 2018.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka, "Attention discrimination: Theory and field experiments with monitoring information acquisition," *American Economic Review*, 2016, *106* (6), 1437–75.

Beck, Andrew H, Ankur R Sangoi, Samuel Leung, Robert J Marinelli, Torsten O Nielsen, Marc J Van De Vijver, Robert B West, Matt Van De Rijn, and Daphne Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science translational medicine*, 2011, *3* (108), 108ra113–108ra113.

Becker, Gary S, "The economics of discrimination Chicago," *University of Chicago*, 1957.

_ , "Nobel lecture: The economic way of looking at behavior," *Journal of political economy*, 1993, *101* (3), 385–409.

Ben-David, Shai, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff, "Learnability can be undecidable," *Nature Machine Intelligence*, 2019, *1* (1), 44.

Berk, Richard, "An impact assessment of machine learning risk forecasts on parole board decisions and recidivism," *Journal of Experimental Criminology*, 2017, *13* (2), 193–216.

_ , Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, "Fairness in Criminal Justice Risk Assessments: The State of the Art," *arXiv preprint arXiv:1703.09207*, 2017.

Bhattacharya, Sudipto and Jay R Ritter, "Innovation and communication: Signalling with partial disclosure," *The Review of Economic Studies*, 1983, *50* (2), 331–346.

Bigman, Yochanan E and Kurt Gray, "People are averse to machines making moral decisions," *Cognition*, 2018, *181*, 21–34.

Blodgett, Su Lin and Brendan O'Connor, "Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English," *arXiv preprint arXiv:1707.00061*, 2017.

Bolino, Mark C and Adam M Grant, "The bright side of being prosocial at work, and the dark side, too: A review and agenda for research on other-oriented motives, behavior, and impact in organizations," *The Academy of Management Annals*, 2016, *10* (1), 599–670.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *The Quarterly Journal of Economics*, 2016, *131* (4), 1753–1794.

Botelho, Tristan L and Mabel Abraham, "Pursuing Quality: How Search Costs and Uncertainty Magnify Gender-based Double Standards in a Multistage Evaluation Process," *Administrative Science Quarterly*, 2017, *62* (4), 698–730.

Brown, Charles, "Education and Jobs: An Interpretation," *The Journal of Human Resources*, 1978, *13* (3), 416–421.

Bryan, Laura Koppes and Jerry K Palmer, "Do job applicant credit histories predict performance appraisal ratings or termination decisions?," *The Psychologist-Manager Journal*, 2012, *15* (2), 106–127.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?," in "AEA Papers and Proceedings," Vol. 108 2018, pp. 43–47.

Busse, Meghan R, Devin G Pope, Jaren C Pope, and Jorge Silva-Risso, "The psychological effect of weather on car purchases," *The Quarterly Journal of Economics*, 2015, *130* (1), 371–414.

Byrnes, Nanette, "Why we should expect algorithms to be biased," 2016.

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy, "Building classifiers with independency constraints," in "Data mining workshops, 2009. ICDMW'09. IEEE international conference on" IEEE 2009, pp. 13–18.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello, "Artificial intelligence, algorithmic pricing and collusion," 2018.

Caplan, Bryan, *The case against education: Why the education system is a waste of time and money*, Princeton University Press, 2018.

Cappelli, Peter, Prasanna Tambe, and Valery Yakubovich, "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward," *Available at SSRN 3263878*, 2018.

Card, David and Gordon B Dahl, "Family violence and football: The effect of unexpected emotional cues on violent behavior," *The Quarterly Journal of Economics*, 2011, *126* (1), 103–143.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in "Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" ACM 2015, pp. 1721–1730.

Chang, Tom and Antoinette Schoar, "Judge specific differences in Chapter 11 and firm outcomes," *Unpublished working paper, National Bureau of Economic Research Cambridge*, 2013.

Chen, Irene, Fredrik D Johansson, and David Sontag, "Why Is My Classifier Discriminatory?," *arXiv preprint arXiv:1805.12002*, 2018.

Chen, Jiahao, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell, "Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved," *arXiv preprint arXiv:1811.11154*, 2018.

Chen, M Keith and Jesse M Shapiro, "Do harsher prison conditions reduce recidivism? A discontinuity-based approach," *American Law and Economics Review*, 2007, *9* (1), 1–29.

Chen, Xi, Xiaotie Deng, and Shang-Hua Teng, "Settling the complexity of computing two-player Nash equilibria," *Journal of the ACM (JACM)*, 2009, *56* (3), 14.

Chiappa, Silvia and Thomas PS Gillam, "Path-specific counterfactual fairness," *arXiv preprint arXiv:1802.08139*, 2018.

Chouldechova, Alexandra, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, 2017, *5* (2), 153–163.

__ **and Aaron Roth**, "The Frontiers of Fairness in Machine Learning," *arXiv preprint arXiv:1810.08810*, 2018.

__ **and Max G'Sell**, "Fairer and more accurate, but for whom?," *arXiv preprint arXiv:1707.00046*, 2017.

Clifford, Robert and Daniel Shoag, "'No More Credit Score': Employer Credit Check Bans and Signal Substitution," 2016.

Coate, Stephen and Glenn C Loury, "Will affirmative-action policies eliminate negative stereotypes?," *The American Economic Review*, 1993, pp. 1220–1240.

Cohen, Wesley M, Richard R Nelson, and John P Walsh, "Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not)," Technical Report, National Bureau of Economic Research 2000.

Commission, Equal Employment Opportunity et al., "Uniform Guidelines on Employee Selection Procedures," *Federal register*, 1978, *43* (166), 38295–38309.

Corbett-Davies, Sam and Sharad Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

_ , Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, "Algorithmic decision making and the cost of fairness," *arXiv preprint arXiv:1701.08230*, 2017.

Cortes, Kristle Romero, Andrew S Glover, and Murat Tasci, "The unintended consequences of employer credit check bans on labor and credit markets," 2018.

Cowgill, Bo, "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening," *Working Paper*, 2017.

_ , "The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities," *Working paper*, 2018.

_ , "The Econometrics of Gatekeeping Experiments," *Working paper*, 2018.

_ , "Bias and Productivity in Humans and Algorithms," *Working Paper*, 2018.

_ **and Cosmina L Dorobantu**, "Competition and Specificity in Market Design: Evidence from Geotargeted Advertising," *Available at SSRN 3267053*, 2018.

_ **and Eric Zitzewitz**, "Mood Swings at Work: Stock Price Movements, Effort and Decision Making," 2008.

_ **and** _ , "Corporate prediction markets: Evidence from google, ford, and firm x," *The Review of Economic Studies*, 2015, *82* (4), 1309–1341.

_ **and** _ , "Incentive Effects of Equity Compensation: Employee Level Evidence from Google," *Dartmouth Department of Economics working paper*, 2015.

_ **and Fabrizio Dell'Acqua**, "Biased Programmers? Or Biased Training Data? A Field Experiment about Algorithmic Bias," *AEA RCT Registry*, 2018.

Crawford, Vincent P and Joel Sobel, "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, 1982, pp. 1431–1451.

Crews, Aaron, "The Big Move Toward Big Data in Employment," in "Data-Driven Law," Auerbach Publications, 2018, pp. 59–102.

Daniely, Amit, "Complexity theoretic limitations on learning halfspaces," in "Proceedings of the forty-eighth annual ACM symposium on Theory of Computing" ACM 2016, pp. 105–117.

Darrow, Jonathan J, Ameet Sarpatwari, Jerry Avorn, and Aaron S Kesselheim, "Practical, legal, and ethical issues in expanded access to investigational drugs," 2015.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta, "Automated experiments on ad privacy settings," *Proceedings on Privacy Enhancing Technologies*, 2015, *2015* (1), 92–112.

Dawes, Robyn M, "A case study of graduate admissions: Application of three principles of human decision making.," *American psychologist*, 1971, *26* (2), 180.

_ , "The robust beauty of improper linear models in decision making.," *American psychologist*, 1979, *34* (7), 571.

_ , David Faust, and Paul E Meehl, "Clinical versus actuarial judgment," *Science*, 1989, *243* (4899), 1668–1674.

Dessein, Wouter, "Authority and communication in organizations," *The Review of Economic Studies*, 2002, *69* (4), 811–838.

Devenow, Andrea and Ivo Welch, "Rational herding in financial economics," *European Economic Review*, 1996, *40* (3-5), 603–615.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114.

_ , _ , and _ , "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, 2016, *64* (3), 1155–1170.

Dimakopoulou, Maria, Susan Athey, and Guido Imbens, "Estimation Considerations in Contextual Bandits," *arXiv preprint arXiv:1711.07077*, 2017.

Dobbie, Will and Jae Song, "Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection," *The American Economic Review*, 2015, *105* (3), 1272–1311.

_ , Andres Liberman, Daniel Paravisini, and Vikram Pathania, "Measuring Bias in Consumer Lending," Technical Report, National Bureau of Economic Research 2018.

_ , Jacob Goldin, and Crystal S Yang, "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 2018, *108* (2), 201–40.

Doleac, Jennifer L and Megan Stevenson, "Are Criminal Risk Assessment Scores Racist?," *Brookings, August*, 2016, *22*.

Doshi-Velez, Finale and Been Kim, "Towards a rigorous science of interpretable machine learning," *Working paper*, 2017.

Dressel, Julia and Hany Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, 2018, *4* (1), eaao5580.

Durlauf, Steven N and Daniel S Nagin, "Imprisonment and crime," *Criminology & Public Policy*, 2011, *10* (1), 13–54.

Dwork, Cynthia and Christina Ilvento, "Fairness Under Composition," *arXiv preprint arXiv:1806.06122*, 2018.

Edelman, Benjamin, Michael Luca, and Dan Svirsky, "Racial discrimination in the sharing economy: Evidence from a field experiment," *American Economic Journal: Applied Economics*, 2017, *9* (2), 1–22.

Ederer, Florian, Richard Holden, and Margaret Meyer, "Gaming and strategic opacity in incentive provision," *The RAND Journal of Economics*, 2018, *49* (4), 819–854.

Edmans, Alex, Diego Garcia, and Øyvind Norli, "Sports sentiment and stock returns," *The Journal of Finance*, 2007, *62* (4), 1967–1998.

Engelberg, Joseph and Christopher A Parsons, "Worrying about the stock market: Evidence from hospital admissions," *The Journal of Finance*, 2016.

Ensign, Danielle, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian, "Runaway feedback loops in predictive policing," *arXiv preprint arXiv:1706.09847*, 2017.

Ezrachi, Ariel and Maurice E Stucke, "Virtual competition," 2016.

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist, "What is a Patent Worth? Evidence from the US Patent Lottery," Technical Report, National Bureau of Economic Research 2017.

Farrell, Joseph and Matthew Rabin, "Cheap talk," *Journal of Economic perspectives*, 1996, *10* (3), 103–118.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, "Certifying and removing disparate impact," in "Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" ACM 2015, pp. 259–268.

Flores, Anthony W, Kristin Bechtel, and Christopher T Lowenkamp, "False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks," *Fed. Probation*, 2016, *80*, 38.

Friedberg, Leora, Richard Hynes, and Nathaniel Pattison, "Who Benefits from Credit Report Bans?," Technical Report, Working Paper, December 7. 1, 2, 5, 9 2016.

Friedler, Sorelle A, Carlos Scheidegger, and Suresh Venkatasubramanian, "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236*, 2016.

Friedman, Batya and Helen Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, 1996, *14* (3), 330–347.

Friedman, Milton, *Essays in positive economics*, University of Chicago Press, 1953.

Fu, Hu, Patrick Jordan, Mohammad Mahdian, Uri Nadav, Inbal Talgam-Cohen, and Sergei Vassilvitskii, "Ad auctions with data," in "Algorithmic Game Theory," Springer, 2012, pp. 168–179.

Fudenberg, Drew and David K Levine, "Self-confirming equilibrium," *Econometrica: Journal of the Econometric Society*, 1993, pp. 523–545.

Furman, Jason and Robert Seamans, "AI and the Economy," *Innovation Policy and the Economy*, 2019, *19* (1), 161–191.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably unequal? the effects of machine learning on credit markets," 2017.

Gennaioli, Nicola and Andrei Shleifer, "What Comes to Mind*," *The Quarterly Journal of Economics*, 2010, *125* (4), 1399–1433.

Gentzkow, Matthew and Jesse M Shapiro, "Code and data for the social sciences: A practitioners guide," *Chicago, IL: University of Chicago*, 2014.

Gibbons, Robert, "Incentives in organizations," *Journal of economic perspectives*, 1998, *12* (4), 115–132.

Gillen, Benjamin J, Charles R Plott, and Matthew Shum, "A Pari-Mutuel-Like Mechanism for Information Aggregation: A Field Test inside Intel," *Journal of Political Economy*, 2017, *125* (4), 1075–1099.

Gilovich, Thomas, Kenneth Savitsky, and Victoria Husted Medvec, "The illusion of transparency: biased assessments of others' ability to read one's emotional states.," *Journal of personality and social psychology*, 1998, *75* (2), 332.

Goodman, Bryce and Seth R Flaxman, "European Union regulations on algorithmic decision-making and a right to explanation," 2017.

Green, Etan, Justin M Rao, and David M Rothschild, "A Sharp Test of the Portability of Expertise," 2017.

Greenwald, Anthony G, Debbie E McGhee, and Jordan LK Schwartz, "Measuring individual differences in implicit cognition: the implicit association test.," *Journal of personality and social psychology*, 1998, *74* (6), 1464.

Greiner, D James and Donald B Rubin, "Causal effects of perceived immutable characteristics," *Review of Economics and Statistics*, 2011, *93* (3), 775–785.

Grove, William M, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson, "Clinical versus mechanical prediction: a meta-analysis.," *Psychological assessment*, 2000, *12* (1), 19.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys (CSUR)*, 2018, *51* (5), 93.

Gupta, Arpit, Christopher Hansman, and Ethan Frenchman, "The heavy costs of high bail: Evidence from judge randomization," *The Journal of Legal Studies*, 2016, *45* (2), 471–505.

Guruswami, Venkatesan and Prasad Raghavendra, "Hardness of learning halfspaces with noise," *SIAM Journal on Computing*, 2009, *39* (2), 742–765.

Guryan, Jonathan and Kerwin Kofi Charles, "Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots," *The Economic Journal*, 2013, *123* (572), F417–F432.

Hadfield-Menell, Dylan and Gillian K Hadfield, "Incomplete Contracting and AI Alignment," 2018.

Halac, Marina and Pierre Yared, "Commitment vs. flexibility with costly verification," Technical Report, National Bureau of Economic Research 2016.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning through noticing: Theory and evidence from a field experiment," *The Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Hardt, Moritz, Eric Price, Nati Srebro et al., "Equality of opportunity in supervised learning," in "Advances in Neural Information Processing Systems" 2016, pp. 3315–3323.

Heckman, James, "Sample Selection Bias as a Specification Error.," *Econometrica*, 1979.

Hellman, Samuel and Deborah S Hellman, "Of mice but not men: problems of the randomized clinical trial," 1991.

Highhouse, Scott, "Stubborn reliance on intuition and subjectivity in employee selection," *Industrial and Organizational Psychology*, 2008, *1* (3), 333–342.

Hirshleifer, David and Tyler Shumway, "Good day sunshine: Stock returns and the weather," *The Journal of Finance*, 2003, *58* (3), 1009–1032.

Hoffman, Mitch, Lisa B Kahn, and Danielle Li, "Discretion in Hiring," 2016.

Holland, Paul W, "Statistics and causal inference," *Journal of the American statistical Association*, 1986, *81* (396), 945–960.

Holmstrom, Bengt and Paul Milgrom, "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design," *JL Econ. & Org.*, 1991, *7*, 24.

Hossenfelder, Sabine, "Lost in Math: How beauty leads physics astray," 2018.

Hu, Lily, Nicole Immorlica, and Jennifer Wortman Vaughan, "The Disparate Effects of Strategic Manipulation," *arXiv preprint arXiv:1808.08646*, 2018.

Hummel, Patrick and R Preston McAfee, "When does improved targeting increase revenue?," in "Proceedings of the 24th International Conference on World Wide Web" International World Wide Web Conferences Steering Committee 2015, pp. 462–472.

Jabbari, Shahin, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth, "Fairness in reinforcement learning," in "International Conference on Machine Learning" 2017, pp. 1617–1626.

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu, "Political language in economics," 2015.

Johndrow, James E and Kristian Lum, "An algorithm for removing sensitive information: application to race-independent recidivism prediction," *arXiv preprint arXiv:1703.04957*, 2017.

Joseph, Matthew, Michael Kearns, Jamie H Morgenstern, and Aaron Roth, "Fairness in learning: Classic and contextual bandits," in "Advances in Neural Information Processing Systems" 2016, pp. 325–333.

Jr, Joseph J Doyle, "Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care," *Journal of political Economy*, 2008, *116* (4), 746–770.

_ et al., "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *American Economic Review*, 2007, *97* (5), 1583–1610.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein, "Simple rules for complex decisions," 2017.

Kahneman, Daniel, "Remarks by Daniel Kahneman," *NBER Economics of AI Conference*, 2017.

_ , Barbara L Fredrickson, Charles A Schreiber, and Donald A Redelmeier, "When more pain is preferred to less: Adding a better end," *Psychological science*, 1993, *4* (6), 401–405.

_ , M Rosenfield, Linnea Gandhi, and Tom Blaser, "Noise: How to overcome the high, hidden cost of inconsistent decision making," *Harvard Business Review*, 2016, *10*, 38–46.

Kallus, Nathan, "Instrument-armed bandits," *arXiv preprint arXiv:1705.07377*, 2017.

_ and Angela Zhou, "Residual Unfairness in Fair Machine Learning from Prejudiced Data," *arXiv preprint arXiv:1806.02887*, 2018.

Kamiran, Faisal and Toon Calders, "Classifying without discriminating," in "Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on" IEEE 2009, pp. 1–6.

Kaplan, A., *The conduct of inquiry: methodology for behavioral science* Chandler publications in anthropology and sociology, Chandler Pub. Co., 1964.

Kartik, Navin, "Strategic communication with lying costs," *The Review of Economic Studies*, 2009, *76* (4), 1359–1395.

_ , Marco Ottaviani, and Francesco Squintani, "Credulity, lies, and costly talk," *Journal of Economic theory*, 2007, *134* (1), 93–116.

Kearns, Michael J, *The computational complexity of machine learning*, MIT press, 1990.

Keiser, Michael J, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran et al., "Predicting new molecular targets for known drugs," *Nature*, 2009, *462* (7270), 175.

Kerr, Steven, "On the folly of rewarding A, while hoping for B," *Academy of Management journal*, 1975, *18* (4), 769–783.

Kleinberg, Jon and Manish Raghavan, "Selection Problems in the Presence of Implicit Bias," *arXiv preprint arXiv:1801.03533*, 2018.

_ and Sendhil Mullainathan, "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability," *arXiv preprint arXiv:1809.04578*, 2018.

_ , Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, 2017, *133* (1), 237–293.

_ , **Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan**, "Algorithmic Fairness," in "AEA Papers and Proceedings," Vol. 108 2018, pp. 22–27.

_ , _ , _ , **and Cass R Sunstein**, "Discrimination in the Age of Algorithms," *Available at SSRN 3329669*, 2019.

_ , **Sendhil Mullainathan, and Manish Raghavan**, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

Kling, Jeffrey R, "Incarceration length, employment, and earnings," *The American economic review*, 2006, *96* (3), 863–876.

Knowles, John, Nicola Persico, and Petra Todd, "Racial bias in motor vehicle searches: Theory and evidence," *Journal of Political Economy*, 2001, *109* (1), 203–229.

Koh, Pang Wei and Percy Liang, "Understanding black-box predictions via influence functions," *arXiv preprint arXiv:1703.04730*, 2017.

Kohler-Hausmann, Issa, "Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination," *Available at SSRN 3050650*, 2018.

Korin, Joel B and Madelyn S Quattrone, "Electronic health records raise new risks of malpractice liability," *New Jersey Law Journal. June*, 2007, *19*.

Krause, Josua, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini, "A workflow for visual diagnostics of binary classifiers using instance-level explanations," *arXiv preprint arXiv:1705.01968*, 2017.

Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva, "Counterfactual fairness," in "Advances in Neural Information Processing Systems" 2017, pp. 4066–4076.

Lahey, Joanna N, "Age, women, and hiring an experimental study," *Journal of Human resources*, 2008, *43* (1), 30–56.

_ **and Douglas R Oxley**, "Discrimination at the intersection of age, race, and gender: Evidence from a lab-in-the-field experiment," Technical Report, National Bureau of Economic Research 2018.

Lambrecht, Anja and Catherine E Tucker, "Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads," 2016.

Larson, Jeff, Surya Mattu, and Julia Angwin, "Unintended Consequences of Geographic Targeting," *Technology Science*, 2015.

_ , _ , **Lauren Kirchner, and Julia Angwin**, "How we analyzed the COMPAS recidivism algorithm," *ProPublica (5 2016)*, 2016.

Lazear, Edward P, "Pay equality and industrial politics," *Journal of political economy*, 1989, *97* (3), 561–580.

Lee, David S and Thomas Lemieux, "Regression discontinuity designs in economics," *Journal of economic literature*, 2010, *48* (2), 281–355.

Levitt, Steven D, John A List, and Sally E Sadoff, "Checkmate: Exploring backward induction among chess players," *American Economic Review*, 2011, *101* (2), 975–90.

Lewis, Tanya, "Mystery Mechanisms," *The Scientist Magazine*, 2016.

Li, Danielle, "Expertise versus Bias in Evaluation: Evidence from the NIH," *American Economic Journal: Applied Economics*, 2017, *9* (2), 60–92.

Loecker, Jan De and Jan Eeckhout, "The rise of market power and the macroeconomic implications," Technical Report, National Bureau of Economic Research 2017.

Logg, Jennifer, Julia Minson, and Don A Moore, "Algorithm Appreciation: People Prefer Algorithmic To Human Judgment," 2018.

Lou, Yin, Rich Caruana, and Johannes Gehrke, "Intelligible models for classification and regression," in "Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining" ACM 2012, pp. 150–158.

_ , _ , _ , and Giles Hooker, "Accurate intelligible models with pairwise interactions," in "Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining" ACM 2013, pp. 623–631.

Luco, Fernando, "Who benefits from information disclosure? The case of retail gasoline," 2018.

Ludwig, Jens, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan, "Should Algorithms Be Regulated?," 2018.

Lum, Kristian and James Johndrow, "A statistical framework for fair predictive algorithms," *arXiv preprint arXiv:1610.08077*, 2016.

_ and William Isaac, "To predict and serve?," *Significance*, 2016, *13* (5), 14–19.

Lundberg, Scott M and Su-In Lee, "A unified approach to interpreting model predictions," in "Advances in Neural Information Processing Systems" 2017, pp. 4765–4774.

Lundberg, Shelly J and Richard Startz, "Private discrimination and social intervention in competitive labor market," *The American Economic Review*, 1983, *73* (3), 340–347.

Maestas, Nicole, Kathleen J Mullen, and Alexander Strand, "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt," *The American Economic Review*, 2013, *103* (5), 1797–1829.

Malioutov, Dmitry M, Kush R Varshney, Amin Emad, and Sanjeeb Dash, "Learning interpretable classification rules with boolean compressed sensing," in "Transparent Data Mining for Big and Small Data," Springer, 2017, pp. 95–121.

Malmendier, Ulrike and Geoffrey Tate, "Who makes acquisitions? CEO overconfidence and the market's reaction," *Journal of financial Economics*, 2008, *89* (1), 20–43.

Martens, David and Foster Provost, "Explaining data-driven document classifications," *MIS Quarterly*, 2014, *38* (1), 73–100.

Mearian, Lucas, "Lawyers smell blood in electronic medical records," *Computerworld*, 2015.

Meehl, Paul E, "Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.," 1954.

Meier, Stephan, "A survey of economic theories and field evidence on pro-social behavior," 2006.

Miller, Alex P., "Want Less-Biased Decisions? Use Algorithms.," *Harvard business review*, 2018.

Miller, Amalia and Catherine Tucker, "Algorithms and Historical Racial Bias," *Mimeo, MIT*, 2017.

Miller, Amalia R and Catherine E Tucker, "Electronic discovery and the adoption of information technology," *The Journal of Law, Economics, and Organization*, 2012, *30* (2), 217–243.

Miller, Janet and Judith Glusko, "Standing up to the scrutiny of medical malpractice," *Nursing management*, 2003, *34* (10), 20–22.

Milli, Smitha, John Miller, Anca D Dragan, and Moritz Hardt, "The Social Cost of Strategic Classification," *arXiv preprint arXiv:1808.08460*, 2018.

_ , Ludwig Schmidt, Anca D Dragan, and Moritz Hardt, "Model Reconstruction from Model Explanations," *arXiv preprint arXiv:1807.05185*, 2018.

Mislavsky, Robert, Berkeley Dietvorst, and Uri Simonsohn, "Critical Condition: People Only Object to Corporate Experiments If They Object to a Condition," *Available at SSRN*, 2018.

Mitchell, Shira, Eric Potash, and Solon Barocas, "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions," *arXiv preprint arXiv:1811.07867*, 2018.

Morgan, John and Felix Várdy, "Diversity in the Workplace," *American Economic Review*, 2009, *99* (1), 472–85.

Morson, Gary Saul and Morton Schapiro, *Cents and sensibility: What economics can learn from the humanities*, Princeton University Press, 2018.

Mullainathan, Sendhil and Jann Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Nakamura, Emi and Jón Steinsson, "Identification in macroeconomics," *Journal of Economic Perspectives*, 2018, *32* (3), 59–86.

Narayanan, Arvind, "Algorithmic bias: the hard problems," *available at https://www.youtube.com/watch?v=QBv8reoyc40*, 2018.

_ , "Data as a mirror of society: Lessons from the emerging science of fairness in machine learning," *available at https://q-aps.princeton.edu/sites/default/files/q-aps/files/qssc-abstract-narayanan.pdf*, 2018.

_ , "FAT* tutorial: 21 fairness definitions and their politics," *New York, NY, USA*, 2018.

Nardini, Cecilia, "The ethics of clinical trials," *Ecancermedicalscience*, 2014, *8*.

Nisan, Noam and Amir Ronen, "Algorithmic Mechanism Design," *Games and Economic Behavior*, 2001, *35* (1), 166 – 196.

Obermeyer, Ziad and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People," in "Proceedings of the Conference on Fairness, Accountability, and Transparency" ACM 2019, pp. 89–89.

O'Neil, Cathy, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Broadway Books, 2017.

Oyer, Paul and Scott Schaefer, "Personnel Economics: Hiring and Incentives," *Handbook of Labor Economics*, 2011, *4*, 1769–1823.

Palacios-Huerta, Ignacio and Oscar Volij, "Field centipedes," *American Economic Review*, 2009, *99* (4), 1619–35.

Palmer, Christopher R and William F Rosenberger, "Ethics and practice: alternative designs for phase III randomized clinical trials," *Controlled clinical trials*, 1999, *20* (2), 172–186.

Pathak, Parag A and Tayfun Sönmez, "Leveling the playing field: Sincere and sophisticated players in the Boston mechanism," *American Economic Review*, 2008, *98* (4), 1636–52.

Phelps, Edmund S, "The statistical theory of racism and sexism," *The american economic review*, 1972, pp. 659–661.

Pope, Devin G and Justin R Sydnor, "Implementing anti-discrimination policies in statistical profiling models," *American Economic Journal: Economic Policy*, 2011, *3* (3), 206–31.

Poursabzi-Sangdeh, Forough, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach, "Manipulating and measuring model interpretability," *arXiv preprint arXiv:1802.07810*, 2018.

Powles, Julia and Helen Nissenbaum, "Seductive Diversion of 'Solving' Bias in Artificial Intelligence," *Medium.com*, 2018, *11*.

Quinn, Mariah A, Allyson M Kats, Ken Kleinman, David W Bates, and Steven R Simon, "The relationship between electronic health records and malpractice claims," *Archives of internal medicine*, 2012, *172* (15), 1187–1189.

Rabin, Matthew, "A perspective on psychology and economics," *European economic review*, 2002, *46* (4-5), 657–685.

_ **and Joel L Schrag**, "First impressions matter: A model of confirmatory bias," *The Quarterly Journal of Economics*, 1999, *114* (1), 37–82.

Ransbotham, Sam and Eric Overby, "Does Information technology increase or decrease hospitals' risk? An empirical examination of computerized physician order entry and malpractice claims," Technical Report, ICIS 2010 PROCEEDINGS 2010.

_ **, Eric M Overby, and Michael C Jernigan**, "Electronic Trace Data and Legal Outcomes: The Effect of Electronic Medical Records on Malpractice Claim Resolution Time," 2019.

Rao, Justin M and David H Reiley, "The economics of spam," *Journal of Economic Perspectives*, 2012, *26* (3), 87–110.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in "Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining" ACM 2016, pp. 1135–1144.

Rind, Bruce, "Effect of beliefs about weather conditions on tipping," *Journal of Applied Social Psychology*, 1996, *26* (2), 137–147.

Romei, Andrea and Salvatore Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, 2014, *29* (5), 582–638.

Rozenblit, Leonid and Frank Keil, "The misunderstood limits of folk science: An illusion of explanatory depth," *Cognitive science*, 2002, *26* (5), 521–562.

Sampat, Bhaven and Heidi L Williams, "How do patents affect follow-on innovation? Evidence from the human genome," *available at http://economics.mit.edu/files/9778*, 2014.

Scalia, Justice Antonin, "Wal-Mart Stores, Inc vs Dukes et al.," *131 Supreme Court*, 2011, *2541*.

Schwartzstein, Joshua, "Selective attention and learning," *Journal of the European Economic Association*, 2014, *12* (6), 1423–1452.

Schweitzer, Maurice E and Gérard P Cachon, "Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence," *Management Science*, 2000, *46* (3), 404–420.

Siemroth, Christoph, "The informational content of prices when policy makers react to financial markets," *Journal of Economic Theory*, 2019, *179*, 240–274.

Silberzahn, Raphael, Eric Luis Uhlmann, Dan Martin, Pasquale Anselmi, Frederik Aust, Eli C Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier et al., "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results," *Advances in Methods and Practices in Psychological Science*, 2018, *1* (3), 337–356.

Simoiu, Camelia, Sam Corbett-Davies, Sharad Goel et al., "The problem of infra-marginality in outcome tests for discrimination," *The Annals of Applied Statistics*, 2017, *11* (3), 1193–1216.

Smith, M, D Patil, and C Muñoz, "Big risks, big opportunities: The intersection of big data and civil rights," *White House, Washington, DC*, 2016, *4*.

Spangher, Alexander and Berk Ustun, "Actionable Recourse in Linear Classification," in "Proceedings of the 5th Workshop on Fairness, Accountability and Transparency in Machine Learning. https://econcs. seas. harvard. edu/files/econcs/files/spangher_fatml18. pdf" 2018.

Spence, Michael, "Job Market Signaling," *The Quarterly Journal of Economics*, 1973, *87* (3), 355–374.

Stevenson, Megan, "Assessing Risk Assessment in Action," *George Mason Law & Economics Research Paper*, 2017, (17-36), 4.

__ **and Jennifer Doleac**, "Algorithmic Risk Assessment Tools in the Hands of Humans," 2018.

Stevenson, Megan T, "Distortion of justice: How the inability to pay bail affects case outcomes," *The Journal of Law, Economics, and Organization*, 2018, *34* (4), 511–542.

Studdert, David M, Michelle M Mello, Atul A Gawande, Tejal K Gandhi, Allen Kachalia, Catherine Yoon, Ann Louise Puopolo, and Troyen A Brennan, "Claims, errors, and compensation payments in medical malpractice litigation," *New England journal of medicine*, 2006, *354* (19), 2024–2033.

Tan, Sarah, Julius Adebayo, Kori Inkpen, and Ece Kamar, "Investigating Human+ Machine Complementarity for Recidivism Predictions," *arXiv preprint arXiv:1808.09123*, 2018.

Thomas, Bourveau, Guoman She, and Alminas Zaldokas, "Corporate Disclosure as a Tacit Coordination Mechanism: Evidence from Cartel Enforcement Regulations," *Working paper*, 2018.

Thompson, Nicholas, "Playing With Numbers.," *Washington Monthly*, 2000, *32* (9), 16–23.

Thompson, Richard E, "A validation of the Glueck Social Prediction Scale for proneness to delinquency," *The Journal of Criminal Law, Criminology, and Police Science*, 1952, *43* (4), 451–470.

Tutt, Andrew, "An FDA for algorithms," 2016.

Tversky, A and D Kahneman, "The framing of decisions and the psychology of choice," *Science*, 1981, *211* (4481), 453–458.

Ustun, Berk and Cynthia Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, 2016, *102* (3), 349–391.

Varian, Hal, "Artificial intelligence, economics, and industrial organization," in "The Economics of Artificial Intelligence: An Agenda," University of Chicago Press, 2018.

Varian, Hal R, "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, 2014, *28* (2), 3–28.

Victoroff, Michael S, Barbara M Drury, Elizabeth J Campagna, and Elaine H Morrato, "Impact of electronic health records on malpractice claims in a sample of physician offices in Colorado: a retrospective cohort study," *Journal of general internal medicine*, 2013, *28* (5), 637–644.

Virapongse, Anunta, David W Bates, Ping Shi, Chelsea A Jenter, Lynn A Volk, Ken Kleinman, Luke Sato, and Steven R Simon, "Electronic health records and malpractice claims in office practice," *Archives of internal medicine*, 2008, *168* (21), 2362–2367.

Wattenberg, Martin, Fernanda Viégas, and Moritz Hardt, "Attacking discrimination with smarter machine learning," *Google Research Website*, 2016.

Weaver, Joshua D, "Predicting Employee Performance Using Text Data from Resumes." PhD dissertation, Seattle Pacific University 2017.

Weinstein, Jack B, "The Democratization of Mass Actions in the Internet Age," *Colum. JL & Soc. Probs.*, 2011, *45*, 451.

Wormith, J Stephen and Colin S Goldstone, "The clinical and statistical prediction of recidivism," *Criminal Justice and Behavior*, 1984, *11* (1), 3–34.

Yang, Crystal and Will Dobbie, "Equal Protection under Algorithms: A New Statistical and Legal Framework," *Working paper*, 2018.

Yeomans, Michael, A Shah, Sendhil Mullainathan, and Jon Kleinberg, "Making sense of recommendations," 2017.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in "Proceedings of the 26th International Conference on World Wide Web" International World Wide Web Conferences Steering Committee 2017, pp. 1171–1180.

Zhang, Baobao and Allan Dafoe, "Artificial Intelligence: American Attitudes and Trends," Technical Report, Center for the Governance of AI, Future of Humanity Institute, University of Oxford 2019.

Žliobaitė, Indrė, "Measuring discrimination in algorithmic decision making," *Data Mining and Knowledge Discovery*, 2017, pp. 1–30.