8-6-2019

# Bias and Productivity in Humans and Machines

Bo Cowgill
*Columbia University*

Upjohn Institute working paper ; 19-309

W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH
Researching the causes and consequences of unemployment

# Bias and Productivity in Humans and Machines

**Upjohn Institute Working Paper 19-309**

Bo Cowgill
*Columbia University*
Email: bo.cowgill@gmail.com

August 2019

## ABSTRACT

Where should better learning technology (such as machine learning or AI) improve decisions? I develop a model of decision-making in which better learning technology is complementary with experimentation. Noisy, inconsistent decision-making introduces quasi-experimental variation into training datasets, which complements learning. The model makes heterogeneous predictions about when machine learning algorithms can improve human biases. These algorithms can remove human biases exhibited in historical training data, but only if the human training decisions are sufficiently noisy; otherwise, the algorithms will codify or exacerbate existing biases. Algorithms need only a small amount of noise to correct biases that cause large productivity distortions. As the amount of noise increases, the machine learning can correct both large and increasingly small productivity distortions. The theoretical conditions necessary to completely eliminate bias are extreme and unlikely to appear in real datasets. The model provides theoretical microfoundations for why learning from biased historical datasets may lead to a decrease (if not a full elimination) of bias, as has been documented in several empirical settings. The model makes heterogeneous predictions about the use of human expertise in machine learning. Expert-labeled training datasets may be suboptimal if experts are insufficiently noisy, as prior research suggests. I discuss implications for regulation, labor markets, and business strategy.

**JEL Classification Codes**: C44; C45; D80; O31; O33

**Key Words**: machine learning; training data; decision algorithm; decision-making; human biases

**Acknowledgments**:

The author thanks Dan Gross, John Horton, Daniel Kahneman, John Morgan, Paul Oyer, Olivier Sibony and seminar participants at Columbia, Harvard, MIT, Stanford, and the University of Pennsylvania for valuable feedback. I thank the W.E. Upjohn Institute for supporting this research.

# 1 Introduction

Where should better learning technology improve decisions? Many observers suggest that the better use of data in decisions will result in more empirically grounded, less-biased outcomes. However, predictive algorithms trained using historical data could codify and amplify historical bias. Scholars concerned about algorithmic bias have pointed to a number of troubling examples involving judicial decision-making (Angwin et al., 2016), hiring (Datta et al., 2015; Lambrecht and Tucker, 2016), and targeted advertising (Sweeney, 2013). Policymakers ranging from German chancellor Angela Merkel[1] to the U.S. Equal Employment Opportunity Commission[2] have reacted with public statements and policy guidance. The European Union has adopted sweeping regulations targeting algorithmic bias.[3]

Despite these worries, counterfactual comparisons to other decision-making methods are rare. Where they exist, machine judgment often appears to less biased than human judgment, even when trained on historical data (Kleinberg et al., 2017; Cowgill, 2017; Stern et al., 2018); How can algorithms trained on biased historical data ultimately decrease bias, rather than entrench it?

In this paper, I develop a formal model of the effects of improving learning technology on decision-making. I then apply the model to machine learning, algorithmic bias and their connections to human judgments. The key feature of this model is that improved learning technology is complementary with greater experimentation. In the model, human decision makers generate a historical dataset containing biased decisions, which can arise either from taste-based discrimination or poorly calibrated statistical discrimination that updates slowly.

The human decisions, however, are not only biased but also noisy and inconsistent. Noise in human decisions plays a critical role in how effectively new learning technology can utilize the historical record. The noisiness of human decision making provides experimental variation that is complementary with learning technology. Noisy human judgments creates experimental variation in the training data that facilitates debiasing, rather than codification of preexisting bias. Without noise, new learning technology has no information about counterfactual decisions and their outcomes.

Noisy human judgments reduce the sample selection problem driving algorithmic bias. With sufficient noise, superior learning technology can overcome not only sample selection bias, but also biases in how outcomes are evaluated among the selected applicants. Depending on the level of noise, an algorithm can either replicate historical human bias or completely correct it. The requirements for completely eliminating bias are extreme, and a more plausible scenario is that this type of algorithm will reduce, if not fully eliminate, bias.

The model has several implications for machine learning practitioners, business strategists, and policymakers. The main implication is that settings most ripe for productivity enhancements from

[1] In October 2016, German chancellor Angela Merkel told an audience that "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, accessed July 30, 2019.

[2] In October 2016, the US 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, accessed July 30, 2019.

[3] See the EU General Data Protection Regulation https://www.eugdpr.org/, adopted in 2016 and enforceable as of May 25, 2018.

supervised machine learning are those in which human decision-makers exhibit both bias and inconsistency.

Settings with these features are where machine learning entrepreneurs and venture capitalists may find the most promising – and where human-decision makers may see their jobs most likely to to be displaced. Settings without these features are where machine learning applications are more likely to entrench, rather than reduce, bias. Regulators and policy-makers concerned about algorithmic bias should seek out settings featuring high bias and low noise.

The model also addresses the much-criticized use of "datasets of convenience." This term refers to datasets used in machine learning because they are easy to acquire, even though they do not exhibit theoretically ideal properties for learning. Often these datasets were collected for a separate reason and are repurposed for machine learning by opportunistic computer scientists *ex-post*. They may feature some combination of noise, bias, and unrepresentativeness.

Despite their flaws, "datasets of convenience" are not easily substitutable. Ideal data are expensive or impossible to gather. Machine learning theory instead needs better methods to assess the characteristics of datasets of convenience from the beginning and determine the settings in which an imperfect dataset can be used to improve decisions and reduce errors.

This paper contains an example of such a model. The model may not apply to all settings. However, it demonstrates the theoretical mechanism through which a dataset with heavy bias can yield *less biased* predictions. That mechanism is noisy human decisions act as inadvertent A/B tests – ones that even simple learning algorithms naturally exploit to correct biases. Throughout the paper, I justify the noisiness of human decisions with extensive citations into the psychology and behavioral economics literature.

### 1.1 Benchmarking

One reason the model in this paper performs well is that it is compared to a benchmark of human decision-making. Most assessments of a machine learning performance – metrics such as precision or recall – are based on a model's fit to historical data, rather than relative performance against a counterfactual decision method. The use of counterfactuals in this paper refers to benchmarking a focal algorithm against an alternative decision-making procedure. As I discuss in the literature review below (Section 1.2), this use of "counterfactual" differs from others used in the fairness literature (Kusner et al., 2017; Chiappa and Gillam, 2018).

Counterfactual comparisons are often necessary to measure the practical impact of new decision-making techniques (such as the use of machine learning algorithms). In many cases, counterfactual comparisons may be the ultimate arbiter of whether businesspeople or policymakers adopt new technology. For example: Opening a "FDA for Algorithms" is a commonly suggested policy solution for algorithmic fairness (Tutt, 2016). Counterfactual evaluation through randomized controlled trials – including placebos – are a key component of FDA regulation of pharmaceuticals.

The practical goal of most applied machine learning is *not* to maximize precision or recall, but to maximize the performance improvement between their model and some prefThe existing or alternative decision-making process. However, many computer scientists struggle to optimize this performance difference. There is no *ex-ante* "target variable" to directly model. The assessment

can only be performed after a model has been deployed and tested.

Machine learning theory should give practitioners guidance about when to expect relativistic performance gains, based on inferences about the training data. This paper – which makes predictions about relative performance depending on the bias and noisiness of the training data generated by the status quo – is one attempt to do this. The data-generating process above both is both the source of training data for machine learning, as well as the counterfactual benchmark the algorithm is compared against. These are key features of the model. I show that bias and noise in training data may drive performance improvements in bias, despite inevitably leading to lower precision, recall, and other goodness-of-fit metrics.

The remainder of this paper proceeds as follows. In Section 2, I outline a theoretical setup of human and machine decision-making. In Section 3, I solve and assess the equilibrium outcomes given the setup. I discuss extensions in Section 4, and conclude with a discussion in Section 5. Appendix A contains formal proofs of all propositions.

### 1.2 Related Literature

The model in this paper is related to the emerging fairness literature in computer science (Friedler and Wilson, eds, 2018), and particularly to the usefulness for randomness in learning. Within the fairness literature, several papers explore the application multi-armed banditdts, active- and online learning (Joseph et al., 2016; Dimakopoulou et al., 2017). These papers emphasize the benefit of deliberate, targeted exploration through randomization.

However, some settings give researchers the bandit-like benefits of random exploration for free because of noise in the environment (particularly noise in human decision-making). This may be particularly useful when multi-armed bandits aren't allowed or feasible.[4] However, the experiments arising from environmental noise (described in this paper) are inefficient and poorly targeted.

Methods from the bandit literature are far more statistically efficient because they utilize noise more effectively than human psychology's behavioral quirks. In addition, many bandit-methods eventually (asymptotically) converge to unbiasedness. However as I discuss in Proposition 4, the approach in this paper may not ever converge if the environment isn't sufficiently noisy.

This paper also builds on an early formal models of the effects of machine learning and algorithmic in decision-making, particularly in a strategic environment. This theory model is related to Mullainathan and Obermeyer (2017); Chouldechova and G'Sell (2017); Hardt et al. (2016); Kleinberg et al. (2016). In addition, Hoffman et al. (2016) contains a theoretical model of decision-making by humans and algorithms and evaluates differences.

This paper is related to Agrawal et al. (2017), which models the economic consequences of improved prediction. The paper concludes by raising "the interesting question of whether improved machine prediction can counter such biases or might possibly end up exacerbating them."

This paper aims to advance this question by characterizing and decomposing the nature of prediction improvements. Prediction improvements may come about from improvements in bias

[4] Algorithms requiring deliberate randomization are sometimes viewed as taboo or unethical.

or variance, which may have differing economic effects. As Agrawal et al. (2017) allude, some changes that superficially resemble "prediction improvements" may in fact reinforce deeply held biases.

The model in this paper separately integrates prediction errors from bias and variance – and the possibility of each improving – into a single model that makes heterogeneous predictions about the effects of AI. It also develops microfoundations for how these changes arise endogenously – from the creation of training data through its use by machine learning engineers.

Although the paper addresses counterfactual comparisons, the approach here differs from other discussions of "counterfactuals" used in the fairness literature (Kusner et al., 2017; Chiappa and Gillam, 2018). For example, Kusner et al. (2017) assesses fairness in algorithms by asking: If a candidate's characteristics were (counterfactually) different, would an algorithm's suggestions change? If the suggestions would change in response to a sensitive variable changing, this is interpreted as unfair.[5]

This paper examines a different counterfactual comparison: Suppose that candidates' characteristics could *not* change (which they often cannot), but screening mechanisms *could* counterfactually change. How would a change in screening policy impact the types of candidates selected? How would overall productivity of hired workers be affected?

How would this effect the overall quality of candidates? This approach has the advantage of permitting real-world empirical verification of the models. Researchers can organize trials – field experiments and A/B tests – to test the policy by modifying screening policy. By contrast, researchers cannot easily randomly alter candidate characteristics in the real world. The idea of using model features (weights, coefficients, or derivatives of an algorithm) to measure the impact of an algorithm – which is implicit in Kusner et al. (2017) and related papers – is formally analyzed in Proposition 9 of this paper.

# 2 Theoretical Framework

In this section, I develop a simple decision-theoretic model that aims to help identify settings where algorithmic decision-making will improve outcomes. Although the setting is motivated by hiring, it can be applied to other decision-making settings.

## 2.1 Setup

### 2.1.1 Players

The model features two players. The first is a human decision-maker, who is employed to review resumes and select candidates for a job test or interview. The second is a machine learning engineer, who takes historical data from the human's decisions and creates a predictive model of test outcomes based on the candidates' observable characteristics. This model will later be deployed on new candidates drawn from the same distribution.

[5] A related paper by Chiappa and Gillam (2018) pursues a similar approach.

Job candidates in this framework are not strategic players. Candidates come in two types, $\theta = 1$ and $\theta = 0$ in equal proportion.[6] Both have probabilities $p_0$ and $p_1$ of passing the test. Because this paper is focused on bias, we will focuse on the choice between Type 0 and 1 candidates. In reality, would consider many other factors including qualifications. These variables in our model could be represented in estimates of the $p$ variables.

Type 1 is more likely to pass ($p_1 > p_0$). After testing is completed, each tested candidate has an outcome $y$, a binary variable representing whether the candidate was accepted or not.

### 2.1.2 Utilities

The screener is paid a utility reward of $r \geq 0$ for a candidate who passes the test. We can think of $r$ as the payoffs to correct decisions. In addition, the agent exhibits taste-based bias in favor of Type 0, and receives a taste-based payoff $b \geq 0$ for choosing Type 0.

The human judge also receives random net utility shocks $\eta \sim F$ for picking Type 1. Suppose $F$ is continuous, symmetric, and has continuous and infinite support. $F$ could be a normal distribution (which may be plausible based on the central limit theorem) but can assume other shapes as well.

These utility shocks add random noise and inconsistency to the screener's judgement. This formulation of noise – a utility function featuring a random component – is used in other models and settings, beginning as as early as (Marschak, 1959) and in more recent discrete choice research. The mean of $F$ is zero – if there are average non-zero payoffs to picking either type, this would be included in the bias term $b$.

The noise shocks are motivated by the extensive psychology and behavioral economics literature, showing the influence of random extraneous factors in human decision-making. For example, the noise shocks may come from exogenous factors such as weather (Schwarz and Clore, 1983; Rind, 1996; Hirshleifer and Shumway, 2003; Busse et al., 2015), sports victories (Edmans et al., 2007; Card and Dahl, 2011), stock prices (Cowgill and Zitzewitz, 2008; Engelberg and Parsons, 2016), or other sources of environmental variance that affect decision-makers' mindset or mood, but are unrelated to the focal decision.[7]

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).[8]

The engineers are also strategic players in the model. The firm's management can create incentives for ML engineers to build the hiring algorithm in a variety of ways. This paper will not develop a theoretically optional incentive scheme for the ML engineers, but will instead examine

[6] Proportions of candidates are not a critical part of this theory, and the conclusions of the paper do not depend on a particular proportion. For simplicity, I have used equal proportions.

[7] Note that these exogenous factors may alter the payoffs for picking both Type 0 and Type 1 candidates; $F$ is the distribution of the *net* payoff for picking Type 1.

[8] Bookseller.com, https://www.thebookseller.com/news/william-collins-scoops-kahnemans-book-7-figure-pre-empt-752276. Accessed September 10, 2018

how can be reduced under a particular incentive scheme (a loss function). I will examine a set of algorithms arising when engineers are asked to predict $y$ (passing the test) from $\theta$ by approximating $E[y|\theta]$.

ML engineers will be induced to approximate $E[y|\theta]$ under a variety of incentive schemes, for example, if they are rewarded or penalized based on squared-error loss function for $y$. A wide variety of ML algorithms can be used to predict $E[y|\theta]$.

Firms face strong incentives to predict $E[y|\theta]$. Although other predictions may be useful in various circumstances, developing and maintaining such an algorithm may be expensive, particularly as new data arrive. As a result, firms face incentives to develop prediction technology that can be repurposed for a variety of different tasks. Estimating $E[y|\theta]$ could be useful for a many tasks. Because this may not be the theoretically optimal incentive scheme for all ML engineering tasks, and these results should be interpreted as a lower bound. Other incentive schemes may have better performance. However, the approach above is used in many real world settings and requires only historical data that real-world practitioners may plausibly utilize without additional data gathering.

The machine learning engineer's job in this model is similar to a research econometrician's. However, the machine learning engineer in this setup is not required to produce a model of the human screening process that can be interpreted in light of economic theory of human decision-making. The machine learning engineers are simply required to predict outcomes for candidates in a way that's useful for their firms' selection process.

In the sections that follow, I will show conditions under which algorithms trained in the above manner and used in decision-making will reduce bias. As previously discussed, there are several ways that the algorithm above can be improved to further decrease bias. An emerging literature in computer science (Friedler and Wilson, eds, 2018) develops these improvements, although it does not discuss the usefulness of noise in decreasing bias, and there are few empirical evaluations of how well these methods work compared to counterfactual methods.

### 2.1.3 Sequence

In the first part of the game, the screeners make choices. The historical record of this data-generating process is recorded into a *training data*, which is then given to algorithm developers who are tasked with creating an algorithm to rank candidates for interviews. Then the data are handed over to engineers, who estimate a prediction model. The model is then put into production.

Note that in this setup, the human labeling process is both the source of training data for machine learning, as well as the counterfactual benchmark against which the machine learning is assessed.

### 2.1.4 Information

Screeners in the model are able to see the $\theta$ variable (1 or 0) and the $\eta$ noise realizations. The machine learning engineers can also see the $\theta$ variable for types, and whether they were eventually hired. However, the machine learning engineers do not know the values $p_1$, $p_0$, $q$, $b$ or the $\eta$

realizations. The data-generating process does not encode the source of "noise," and thus it cannot be exploited for econometric identification in a statistical model by these engineers.

The machine learning engineers thus face a limited ability to infer information about candidate quality from the choice to interview. Candidates may be interviewed because of taste-based bias, because of the rewards of performance, or because of a random shock.

This paper will study an algorithm in which knowing why candidates were interviewed – or whether they were interviewed at all – is not necessary. I will assume that all the ML engineers can see is $\theta$ and an outcome variable $y$ for each candidate. $y$ will equal 1 if the candidate was tested and passed and equal zero otherwise.

### 2.1.5 Modeling Choices

In the next section, I analyze the equilibrium behavior for the setup above, beginning with a discussion of a few modeling choices. Although the setup may apply to many real world settings, there are a few limitations of the model worth discussing.

First, although human screeners are able to observe and react to the $\eta$ realizations, they do not recognize them as noise and thus do not learn from the experimentation they induce. This assumption naturally fits settings featuring taste-based discrimination, as I modeled above. In Section 4.1, I discuss alternative microfoundations for the model, including statistical discrimination. From the perspective of this theory, the most important feature of the screeners' bias is that it is stubborn and is *not* self-correcting through learning. Insofar as agents are statistical discriminators, the experimentation is not deliberate and they do not learn from the exogenous variation generated by the noise.

Second, the human screeners and machine learning engineers do not strategically interact in the above model. For example, the human screeners do not attempt to avoid job displacement by feeding the algorithm deliberately sabotaged training data. This may happen if the screeners' direct immediate costs and rewards from picking candidates outweigh the possible effects of displacement costs in the future of automation (perhaps because of present bias).

In addition, there is no role for "unobservables" in this model besides noise. In other words, the only variables privately observed by the human decision-maker (and not in the training data) are noise realizations $\eta$. These noise realizations are not predictive of the candidate's underlying quality, and serve only to facilitate accidental experimentation and exploration of the candidate space. By contrast, in other models (Hoffman et al., 2016), humans are able to see predictive variables that the ML algorithm cannot, and this can be the source of comparative advantage for the humans, depending on how predictive the variable is.

For the theory in this paper to apply, the noise realizations $\eta$ must be truly random – uncorrelated with other observed or unobserved variables, as well as the final productivity outcome. If these conditions are violated, the algorithm may nonetheless have a positive effect on reducing bias. However, this would have to come about through a different mechanism than outlined in the proofs below.

Lastly, this paper makes assumptions about the asymptotic properties of algorithmic predictions.

In particular, I assume that the algorithm convergest to $E[Y|\theta]$, but without specifying a functional form. This is similar to Bajari et al.'s 2018 "agnostic empirical specification." The convergence property is met by a variety of prediction algorithms, including OLS. However, asymptotic properties of many machine learning algorithms are often still unknown. Wager and Athey (2017) shows that the predictions of random forests are asymptotically unbiased. I do not directly model the convergence or its speed. The paper is motivated by applications of "big data," in which sample sizes are large. However, it is possible that for some machine learning algorithms, convergence to this mean may be either slow or nonexistent, even when trained on large amounts of data.

# 3 Equilibrium Choices

## 3.1 Screener's Choices

A risk-neutral human screener will make the "right" decision (Type 1) if $rp_1 + \eta > rp_0 + b$. In other words, the screener makes the right decision if the random utility shocks are enough to offset the taste-based bias ($b$) favoring Type 0.

Let $\underline{\eta} = r(p_0 - p_1) + b$ be the minimum $\eta$ necessary to offset the bias, given the other rewards involved. Such an $\eta$ (or greater) happens with probability of $\Pr(\eta > r(p_0 - p_1) + b) = 1 - F(r(p_0 - p_1) + b)) = q$.

Because this paper is motivated by settings in which the training data are biased, we will restrict attention to the set of distributions $F$ for which $q \in [0, \frac{1}{2}]$. In other words, there will be variation in how often the screener chooses the right decision, but she does not make the right decision in a majority of cases.

The probability $q$ of picking the right candidate changes as a function of the other parameters of this model. The partial derivatives of $q$ are the basis for Proposition 1 and the comparative statics of the human screener selecting Type 1.

**Proposition 1.** *The screener's probability of picking Type 1 candidates ($q$) is decreasing in $b$, increasing in $r$, increasing in the quality difference in Type 1 and Type 0 ($p_1 - p_0$), and increasing in the variance of $F$.* Proof*: See Appendix A.1.*

Proposition 1 makes four statements that can be interpreted as follows. First, as the bias $b$ is greater, the shock necessary to offset this bias must be larger. If $F$ is held constant, these will be more rare.

Second, as the reward for successful decisions $r$ increases, the human screener is equally (or more) likely to make the right decision to pick Type 1. This is because the rewards benefit from picking Type 1 will increasingly outweigh his/her taste-based bias. The $\eta$s necessary to offset this bias are smaller and more common.

Third: Proposition 1 states that as the difference between Type 1 and Type 0 ($p_1 - p_0$) is larger, the screener is more likely to choose Type 1 despite her bias. This is because the taste-based bias against Type 1 is offset by a greater possibility of earning the reward $r$. The minimum $\eta$ necessary for the Type 1 candidate to be hired is thus smaller and more probable.

Finally, $q$ can be higher or lower depending on the characteristics of $F$, the random utility shocks function with mean of zero. For any $b$ and $r$, I will refer to the *default* decision as the type the screener would choose without any noise. Given this default, $F$ is "noisier" if increases the probability mass necessary to flip the decision from the default. This is similar to the screener "trembling" (Selten, 1975) and picking a different type than she would without noise.

Where Type 0 is the default, a *default* $F$ will place greater probability mass above $\underline{\eta}$. This corresponds to a greater $\eta$ realizations above $\underline{\eta}$ favoring Type 1 candidates. In these situations, $q$ is increasing in the level of noise in $F$. For a continuous, symmetric distribution such as the normal distribution, greater variance in $F$ places is noisier regardless of $r$ and $b$, since it increases the probability of a $\eta$ that flips the decision.

### 3.2 ML Engineer's Choices

As previously discussed in Section 2.1.2, this paper examines a set of algorithms in which the engineer is asked to predict $y$ (passing the test) from $\theta$ by approximating $E[y|\theta]$. For Type 0 candidates, this converges to $(1-q)p_0$. For Type 1 candidates, this convergees to $qp_1$.

The ML engineers then use the algorithm to pick the type with a higher $E[Y|\theta]$. It then implements this decision consistently, without any noise. I will now compare the performance of the algorithm's selected candidate to that of the human decision process.

### 3.3 Effects of Shift from Human Screener to Algorithm

**Proposition 2.** *If screeners exhibit bias but zero noise, the algorithm will perfectly codify the humans' historical bias. The algorithm's perfomance will precisely equal that of the biased screeners and exhibit high goodness-of-fit measures on historical human decision data.* Proof*: See Appendix A.2.*

Proposition 2 formalizes a notion of algorithmic bias. In the setting above – featuring biased screeners $b > 0$ with no noise – there is no difference in the decision outcomes. The candidates approved (or rejected) by the humans would face the same outcomes in the machine learning algorithm.

The intuition behind Proposition 2 is that machine learning cannot learn to improve upon the existing historical process without a source of variation and outcomes. Without a source of clean variation – exposing alternative outcomes under different choices – the algorithms will simply repeat what has happened in the past rather than improve upon it.

Because the model will perfectly replicate historical bias, it will exhibit strong goodness-of-fit measures on the training data. The problems with this algorithm will not be apparent from cross-validation, or from additional observations from the data-generating process.

Thus there are no decision-making benefits to using the algorithm. However it is possible that the decision-maker receives other benefits, such as lower costs. Using an algorithm to make a decision may be cheaper than employing a human to make the same decision.

**Proposition 3.** *If screeners exhibit zero bias but non-zero amounts of noise, the algorithm will improve upon the performance of the screeners by removing noise. The amount of performance improvement is increasing in the amount of noise and the quality difference between Type 1 and Type 0 candidates.* Proof*: See Appendix A.3.*

Proposition 3 shows that performance improvements from the algorithm can partly come from improving consistency. Even when human decisions are not biased, noise may be a source of their poor performance. Although noise is useful in some settings for learning – which is the main theme of this paper – the noise harms performance if the decision process is already free of bias.

**Proposition 4.** *If biased screeners are NOT sufficiently noisy, the algorithm will codify the human bias. The reduction in noise will actually make outcomes worse.* Proof*: See Appendix A.4.*

Proposition 4 describes a setting in which screeners are biased and noisy. This generates some observations about Type 1's superior productivity, but not enough for the algorithm to correct for the bias. In the proof for Proposition 4 in Appendix A.4, I formalize the threshold level of noise below which the algorithm is biased.

Beneath this threshold, the algorithm ends up codifying the bias, similarly to Proposition 2 (which featured bias, but no noise). However, the adoption of machine learning actually worsens decisions in the setting of Proposition 4 (whereas it simply made no difference in the setting of Proposition 2). In a biased human regime, any amount of noise actually helps the right candidates gain employment.

The adoption of the machine learning removes this noise by implementing the decision consistently. Without sufficient experimentation in the underlying human process, this algorithm cannot correct the bias. The reduction in noise in this setting actually makes outcomes worse than if we trusted the biased, slightly noisy humans.

**Proposition 5.** *If screeners are biased and sufficiently noisy, the algorithm will reduce the human bias.* Proof*: See Appendix A.5.*

Proposition 5 shows the value of noise for debasing – one of the main results of the paper. If the level of noise is above the threshold in the previous Proposition 4, then the resulting algorithm will feature lower bias than the original screeners' data. This is because the random variation in the human process has acted as a randomized controlled trial, randomly exposing the learning algorithm to Type 1's quality, so that this productivity can be fully incorporated into the algorithm.

In this sense, experimentation and machine learning are compliments. The greater experimentation, the greater ability the machine learning to remove bias. However, this experimentation does not need to be deliberate. Random, accidental noise in decision-making is enough to induce the debiasing if the noise is a large enough influence on decision-making.

Taken together, Propositions 4 and 5 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 nonexperts (they are biased toward their area of expertise, Li, 2017). If this is true, then Proposition 4 suggests that using expert-provided labels for training data in machine learning will codify bias. Furthermore, the performance improvement coming from lower noise will be small, because counterfactual expert was consistent.

Even if experts' evaluations are (on average) better than nonexperts, experts' historical data are not necessarily more useful if the experts fail to explore.

**Proposition 6.** *If the algorithms' human data contain non-zero bias, then "algorithmic bias" cannot be reduced to zero unless the humans in the training data were perfectly noisy (i.e., picking at random).* Proof*: See Appendix A.6.*

Even if screeners are sufficiently noisy to reduce bias (as in Proposition 5), the algorithm's predictions still underestimate the advantage of Type 1 above Type 0.

In particular, the algorithm predicts a $y$ of $qp_1$ for Type 1 and $(1-q)p_0$ for Type 0. The algorithm's implicit quality ratio of Type 1 over Type 0 is $qp_1/(1-q)p_0$. This is less than the quality ratio of Type 1 over Type 0 ($p_1/p_0$) – unless noise is maximized by increasing the variance of $F$ until $q = 1/2$. This would make the training data perfectly representative (i.e., humans were picking workers at random). Despite the reduction in bias, the algorithm will remain handicapped and exhibit some bias because of its training on biased training data.

Picking at random is extremely unlikely to appear in any real-world setting, since the purpose of most hiring is to select workers who are better than average and thus undersample sections of the applicant pool perceived to be weaker. A complete removal of bias therefore appears infeasible from training datasets from real-world observations, particularly observations of agents who are *not* optimizing labels for *ex-post* learning.

It is possible for an algorithm to achieve a total elimination of bias without using perfectly representative training data. This may happen if a procedure manages to "guess" the a totally unbiased algorithm from some other heuristic. Some of the algorithmic innovations suggested by the emerging fairness literature may achieve this. However, in order to achieve certainty that this is algorithm is unbiased, one would need a perfectly representative training dataset (i.e., one where the screeners were picking at random).

**Proposition 7.** *As the amount of noise in human decisions increases, the machine learning can correct increasingly small productivity distortions. The algorithm needs only a small amount of noise to correct errors with large productivity consequences.* Proof*: See Appendix A.7.*

The proposition means that if the screeners' bias displays a large amount of bias, only a small amount of noise is necessary for the algorithm to correct the bias. Similarly if screeners display a small amount of bias, then high amounts of noise are necessary for the algorithm to correct the bias. Large amounts of noise permit debiasing for both large and smaller biases, where as small noise permits only correction of large biases only.[9] Because we want all biases corrected, lots of noise is necessary to remove both large and small biases. However, Proposition 7 suggests that even a small amount of noise is necessary to reduce the most extreme biases.

The intuition behind Proposition 7 is as follows: Suppose that screeners were highly biased against Type 1 workers; this would conceal the large productivity differences between Type 1 and Type 0 candidates. The machine learning algorithm would need to see only a few realizations – a small amount of noise – in order to reduce the bias. Because each "experiment" on Type 1 workers

[9] "Large" and "small" biases are used here in a relative sense – "small" biases in this model could be very harmful on a human scale, but are labeled "small" in this model only in comparison to still even greater biases.

shows so much greater productivity, few such experiments would be necessary for the algorithm to learn the improvement. By contrast, if the bias against Type 1 is small, large amounts of noise would be necessary for the algorithm to learn its way out of it. This is because each "experiment" yields a smaller average productivity gain. As a result, the algorithm requires more observations in order to understand the gains from picking Type 1 candidates.

A recent paper by Azevedo et al. (2018) makes a similar point about $A/B$ testing. A company whose innovation policy is focused on large productivity innovations will need only a small test of each experiment. If the experiments produce large effects, they will be detectable in small sample sizes.

Proposition 7 effectively says there are declining marginal returns to noise. Although there may be increasing *cumulative* returns to noise, the marginal returns are decreasing. As Proposition 2 states, no corrections are possible if screeners are biased and feature no noise. The very first unit of noise – moving from zero noise to positive noise – allows for correction of any large productivity distortions. As additional noise is added, the productivity improvements from machine learning become smaller.

**Proposition 8.** *In settings featuring bias and sufficiently high noise, the algorithm's improvement in bias will be positive and increasing in the level of noise and bias. However, metrics of goodness-of-fit on the training data (and on additional observations from the data-generating process) have an upper bound that is low compared to settings with lower noise and/or lower bias.* Proof*: See Appendix A.8.*

The proof in Appendix A.8 comprares the algorithm's goodness-of-fit metrics on the training data in the setting of Proposition 5 (where debiasing happens) to Propositions 2 and 4, which codify bias. In the setting that facilitates debiasing, goodness-of-fit measures are not only low relative to the others, but also in absolute numbers (compared to values commonly seen in practice).

The implication of Proposition 8 is: If engineers avoid settings where models exhibit poor goodness-of-fit on the training data (and future samples), they will avoid the settings where machine learning has the greatest potential to reduce bias.

**Proposition 9.** *The "coefficient" or "weight" the machine learning algorithm places on $\theta = 1$ when ranking candidates does not equal the treatment effect of using the algorithm rather than human discretion for $\theta = 1$ candidates.* Proof*: See Appendix A.9.*

Proposition 9 discusses how observers should interpret the coefficients and/or weights of the machine learning algorithm. It shows that these weights may be highly misleading about the impact of the algorithm. For example: It's possible for an algorithm that places negative weight on $\theta = 1$ when ranking candidates could nonetheless have a strong positive benefit for $\theta = 1$ candidates and their selection outcomes. This would happen if the human penalized these characteristics even more than the algorithm did.

The internal weights of these algorithms are completely unrelated to which candidates benefit from the algorithm compared to a status quo alternative. The latter comparison requires a comparison to a counterfactual method of selecting candidates.

# 4 Extensions

## 4.1 Other Microfoundations for Noise and Bias

In the setup above, I model bias as taste-based discrimination, and noise coming from utility shocks within the same screener over time. However, both the noise and bias in the model can arise from different microfoundations. These do not affect conclusions of the model. I show these alternative microfoundations formally in Appendix A.10.

The formulation above models the bias against Type 1 candidates as "taste-based" (Becker, 1957), meaning that screeners receive direct negative payoffs for selecting one type of worker. A taste-based discriminator may be conscious of his/her taste-based bias (as would a self-declared racist) or unconscious (as would someone who feels worse hiring a minority, but can't say why). Either way, taste-based discrimination comes directly from the utility function.

Biased outcomes can also arise from statistical discrimination (Phelps, 1972; Arrow, 1973). Screeners exhibiting statistical discrimination (and no other type of bias) experience no direct utility preferences for attributes such as gender or race. "Statistical discrimination" refers to the process of making educated guesses about an unobservable candidate characteristic, such as which applicants' perform well as employees. If applicants performance is (on average) even slightly correlated with observable characteristics such as gender or race, employers may be tempted to use these variables as imperfect proxies for unobservable abilities. If worker quality became easily observable, screeners exhibiting statistical discrimination would be indifferent between races or genders.

The framework in this paper can be reformulated so that the bias comes from statistical discrimination. This simply requires one additional provision: That the "educated guesses" are wrong and are slow to update. Again, the psychology and behavioral economics literature provides ample examples of decision-makers having wrong, overprecise prior beliefs that are slow to update.

Similarly, the noise variable $\eta$ can also have alternative microfoundations. The formulation beginning in Section 2 proposes that $\eta$ represents time-varying noise shocks within a single screener (or set of screeners). However, $\eta$ can also represent noise coming from between-screener variation. If a firm employs multiple screeners and randomly assigns applications to screeners, then noise can arise from idiosyncrasies in the each screener's tastes.

The judgment and decision-making literature contains many examples of this between-screener variation as a source of noise.[10] This literature uses "noise" to refer to within-screener and between-screener random variations interchangeably. Kahneman et al. (2016) simply writes, "We call the chance variability of judgments *noise*. It is an invisible tax on the bottom line of many companies."

Similarly, the empirical economics literature has often exploited this source of random variation for causal identification.[11] This includes many papers in an important empirical setting for the CS

[10] For example, this literature has shown extensive between-screener variation in valuing stocks (Slovic, 1969), evaluating real-estate (Adair et al., 1996), sentencing criminals (Anderson et al., 1999), evaluating job performance (Taylor and Wilsted, 1974), auditing financies (Colbert, 1988), examining patents (Cockburn et al., 2002) and estimating task-completion times Grimstad and Jørgensen (2007).

[11] For example, assignment of criminal cases to judges (Kling, 2006), patents applications to patent examiners (Sampat

literature on algorithmic fairness: Judicial decision-making. As algorithmic risk-assessment tools have grown in popularity in U.S. courts, a series of academic papers and expose-style journalism allege these risk assessment tools are biased. However, these allegations typically do not compare the alleged bias to what a counterfactual human judge would have done without algorithmic guidance.

A series of economics papers examine the random assignment of court cases to judges. Because human judges' approaches are idiosyncratic, random assignment creates substantial noisiness in how cases are decided. These researchers have documented and exploited this noise for all kinds of analysis and inference. The randomness documented in these papers suggests that courts exhibit the noisiness I argue is the key prerequisite for debiasing human judgement through algorithms.

However, clean comparisons with nonalgorithmic judicial decision-making are rare. One paper that does this is Kleinberg et al. (2017). It utilizes random assignment in judges for evaluating a machine learning algorithm for sentencing and finds promising results on reducing demographic bias.

### 4.2 Additional Bias: How Outcomes are Codified

Until now, the model in this paper has featured selection bias in which a lower-quality candidate joins the training data because of bias. This is a realistic portrayal of many fields, where performance is accurately measured for workers in the field, but entry into the field may contain bias. For example: In jobs in finance, sales, and some manual labor industries, performance can be measured objectively and accurately for workers in these jobs. However, entry into these labor markets may feature unjust discrimination.

In other settings, bias may also appear within the training data in the way outcomes are evaluated for workers who have successfully entered. For example: Suppose that every positive outcome by a Type 1 candidate is scored at only 90% as valuable as those by Type 0. In this extension, I will evaluate the model's impact when $\theta = 1$ candidates are affected by both types of bias.

Let $\delta \in [0, 1]$ represent the discount that Type 1's victories are given in the training data. High $\delta$s represent strong bias in the way Type 1's outcomes are evaluated. If $\delta = 0.9$, then Type 1's victories are *codified* as only 10% as valuable as Type 0's even if they are equally valuable in an objective sense. This could happen if (say) the test evaluators were biased against Type 1 and subtracted points unfairly.[12]

In Appendix B, I provide microfoundations for $\delta$ and update the propositions above to incorporate both types of bias. Again, noise is useful for debasing in many settings (Appendix Proposition 15). The introduction of the second type of bias actually increases the usefulness of noise. However, the existence of the second type of bias also creates limitations. For a threshold level of $\delta$, the algorithm under this procedure will not decrease bias and can only entrench it (Appendix

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).

[12] As with the earlier bias in hiring ($b$), the evaluation bias here ($\delta$) could itself be the result of tastes or statistical inferences about the underlying quality of work.

Proposition 11) no matter how much noise in selection.

These conclusions assume that evaluations could be biased ($\delta$), but these evaluations are not themselves noisy (in the same way that selection decisions were). Future research will add a parameter for noisy posthire evaluations.

## 5 Discussion and Conclusion

This paper contains a model of how human judges make decisions, how these decisions are codified into training data, and how this training data is incorporated into a decision-making algorithm by engineers under mild assumptions.

I show how characteristics of the underlying human decision process propagate the later codification into training data and an algorithm, under circumstances common in practice.

The key feature of the model is that improvements to the human process are made possible only through experimental variation. This experimentation need not be deliberate and can come through random noise in historical decision-making.

Although the model was motivated by hiring, it could be applied to a wide variety of other settings in which bias and noise may be a factor. For example: Many researchers wonder if machine learning or AI will find natural applications decreasing behavioral economics biases (loss aversion, hindsight bias, risk-aversion, etc). This model predicts that this is a natural application, but only these biases are realized in a noisy and inconsistent way.

Similarly, one can also use this model to assess why certain machine learning applications have been successful and which ones may be next. For example: Early, successful models of computer chess utilized supervised machine learning based on historical human data. The underlying humans most likely played chess with behavioral biases, and also featured within-player and between-player sources of noise. The model suggests that the plausibly high amounts of bias and noise in human chess moves make it a natural application for supervised AI that would reduce both the inconsistency and bias in human players.

Recent work by (Brynjolfsson and Mitchell, 2017; Brynjolfsson et al., 2018) attempts to classify jobs tasks in the Bureau of Labor Statistics' O*NET database for their suitability for machine learning applications. The authors create "a 21 question rubric for assessing the suitability of tasks for machine learning, particularly supervised learning." A similar paper by Frey and Osborne (2017) attempts to classify job tasks easily automated by machine learning.

In these papers, the level of noise, random variation, or experimentation in the training data is not a criteria for "suitability for machine learning." Noisiness or quasi-experimental variation is not a major component of the theoretical aspects in either paper. In Brynjolfsson and Mitchell (2017); Brynjolfsson et al. (2018), noise is a *negative* predictor of "suitability for machine learning."

Instead, both analyses appear to focus mostly on settings in which human decision-making process can easily be mimicked rather than improved upon through learning. This is consistent with the goal of maximizing goodness of fit to historical data (as characterized above) rather than reducing bias. Proposition 2 suggests that applying machine learning in low noise environments

will yield mimicking, cost savings, high goodness-of-fit measures and possible entrenchment of bias – rather than better, less-biased decision-making.

If the goal is learning and improving, noisiness should be positively correlated with adoption. Future empirical work in the spirit of (Frey and Osborne, 2017; Brynjolfsson and Mitchell, 2017; Brynjolfsson et al., 2018) may be able to separately characterize jobs or tasks where machine learning yields cost-reduction, mimicking benefits from those where benefits arise from learning and optimizing using experiments.

Researchers in some areas of machine learning – particularly active learning, online learning, and multi-armed bandits – embrace randomization as a tool for learning. In many settings researchers enjoy the benefits of randomization for free because of noise in the environment.

By promoting consistent decision-making, adopting algorithms may actually eliminate useful experimental variation this paper has argued is so useful. Ongoing experimentation can be particularly valuable if the data-generating environment is changing. However, the experiments arising from environmental noise are inefficient and poorly targeted. Methods from the bandit and online learning contain much more statistically efficient use of noise than human psychology's behavioral quirks.

The setup described above may not apply well to all settings. In particular, there may be settings in which variables observable to humans (but unobservable in training data) could play a larger role as they do *not* in the model above. We may live biased world featuring lots of noise – but still not enough to use in debiasing. For some variables (such as college major or GPA) we may have enough noise to facilitate debiasing, while for others (such as race or gender), historical bias may be too entrenched and consistent for algorithms to learn their way out. In addition, there are other ways that machine learning could reduce bias besides the mechanisms in this paper.

In many settings, however, these and other assumptions of the model are realistic. The small number of empirical papers featuring clean comparisons between human and algorithmic judgment (Kleinberg et al., 2017; Cowgill, 2017; Stern et al., 2018) demonstrate reduction of bias.

One interpretation of this paper is that it makes *optimistic* predictions about the impact of machine learning on bias, even without extensive adjustments for fairness. Prior research cited throughout this paper suggests that noise and bias are abundant in human decision-making, and thus ripe for learning and debiasing through the theoretical mechanisms in the model. Proposition 7 may have particularly optimistic implications – if we are in a world with lots of bias, we need only a little bit of noise for simple machines to correct it. If we are in a world with lots of noise (as psychology researchers suggest), simple algorithms should be able to correct even small biases.

Given this, why have so many commentators raised alarms about algorithmic bias? One possible reason is the choice of benchmark. The results of this paper suggest that completely eliminating bias – a benchmark of zero-bias algorithmic perfection – may be extremely difficult to realize from naturally occurring datasets (Proposition 6). However, reducing bias of an existing noisy process may be more feasible. Clean, well-identified comparisons of human and algorithmic judgment are rare in this literature, but the few available (Kleinberg et al., 2017; Cowgill, 2017; Stern et al., 2018) suggest a reduction of bias. These results may come perhaps for the theoretical reasons motivated by this model.

The impact of algorithms compared to a counterfactual decision process may be an important component of how algorithms are evaluated for adoption and legal/social compliance. However, standard machine learning quality metrics – goodness of fit on historical outcomes – do not capture these counterfactual comparisons. This paper suggests that to maximize counterfactual impact, researchers should pick settings in which traditional goodness-of-fit measures may be lower (i.e., those featuring lots of bias and noise).

Relative comparisons are sometimes feasible only after a model has been deployed and tested. One attractive property of the model in this paper is that many of the pivotal features could plausibly be measured in advance – at the beginning of a project, before the deployment and model building – to estimate the eventual comparative effect vs a status quo. (Kahneman et al., 2016) described simple methods for "noise audits" to estimate the extent of noise in a decision process. Levels of bias could be estimated or calibrated through historical observation data, which may suggest an upper or lower bound for bias.

Eliminating bias may be difficult or impossible using "datasets of convenience." Machine learning theory should give practitioners guidance about when to expect practical, relative performance gains, based on observable inferences about the training data. This paper – which makes predictions about relative performance depending on the bias and noisiness of the training data generated by the status quo – is one attempt to do this.

## References


**Adair, Alastair, Norman Hutchison, Bryan MacGregor, Stanley McGreal, and Nanda Nanthakumaran**, "An analysis of valuation variation in the UK commercial property market: Hager and Lord revisited," *Journal of Property Valuation and Investment*, 1996, *14* (5), 34–47.

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

**Anderson, James M, Jeffrey R Kling, and Kate Stith**, "Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines," *The Journal of Law and Economics*, 1999, *42* (S1), 271–308.

**Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner**, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," *ProPublica, May*, 2016, *23*.

**Arrow, Kenneth J.**, *The Theory of Discrimination*, Princeton University Press, 1973.

**Autor, David H 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.

**Azevedo, Eduardo M, Deng Alex, Jose Montiel Olea, Justin M Rao, and E Glen Weyl**, "A/b testing," 2018.

**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.

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

**Brynjolfsson, Erik and Tom Mitchell**, "What can machine learning do? Workforce implications," *Science*, 2017, *358* (6370), 1530–1534.

**_ , _ , 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.

**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.

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

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

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

**Cockburn, Iain M, Samuel Kortum, and Scott Stern**, "Are all patent examiners equal? The impact of examiner characteristics," Technical Report, National Bureau of Economic Research 2002.

**Colbert, Janet L**, "Inherent risk: An investigation of auditors' judgments," *Accounting, Organizations and society*, 1988, *13* (2), 111–121.

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

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

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

**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.

**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.

**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.

**Frey, Carl Benedikt and Michael A Osborne**, "The future of employment: how susceptible are jobs to computerisation?," *Technological forecasting and social change*, 2017, *114*, 254–280.

**Friedler, Sorelle A. and Christo Wilson, eds**, *Conference on Fairness, Accountability and Transparency, 23-24 February 2018*, Vol. 81 of *Proceedings of Machine Learning Research* PMLR 2018.

**Grimstad, Stein and Magne Jørgensen**, "Inconsistency of expert judgment-based estimates of software development effort," *Journal of Systems and Software*, 2007, *80* (11), 1770–1777.

**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.

**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.

**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.

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

**_ , 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.

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

**_ , 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.

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

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

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

**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.

**Marschak, Jacob**, "Binary choice constraints and random utility indicators," Technical Report, YALE UNIV NEW HAVEN CT COWLES FOUNDATION FOR RESEARCH IN ECONOMICS 1959.

**Mullainathan, Sendhil and Ziad Obermeyer**, "Does Machine Learning Automate Moral Hazard and Error?," *American Economic Review*, 2017, *107* (5), 476–480.

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

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

**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.

**Schwarz, Norbert and Gerald L Clore**, "Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states.," *Journal of personality and social psychology*, 1983, *45* (3), 513.

**Selten, Reinhard**, "Reexamination of the perfectness concept for equilibrium points in extensive games," *International journal of game theory*, 1975, *4* (1), 25–55.

**Slovic, Paul**, "Analyzing the expert judge: A descriptive study of a stockbroker's decision process.," *Journal of Applied Psychology*, 1969, *53* (4), 255.

**Stern, Léa H, Isil Erel, Chenhao Tan, and Michael S Weisbach**, "Selecting Directors Using Machine Learning," 2018.

**Sweeney, Latanya**, "Discrimination in online ad delivery," *Queue*, 2013, *11* (3), 10.

**Taylor, Robert L and William D Wilsted**, "Capturing judgment policies: A field study of performance appraisal," *Academy of Management Journal*, 1974, *17* (3), 440–449.

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

**Wager, Stefan and Susan Athey**, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 2017, (just-accepted).

# A Proofs of Main Propositions

## A.1 Proof of Proposition 1

Note that $q = 1 - F(x)$, which is decreasing in the argument $x$. In our setting, $x = r(p_0 - p_1) + b$, which is increasing in $b$ and so $1 - F(x)$ is decreasing in $b$. Note that $p_1 > p_0$, so $x$ is also decreasing in $r \implies 1 - F(x)$ is increasing in $r$. Also: $x$ is decreasing in $p_1 - p_0$, so $1 - F(x)$ is increasing in $p_1 - p_0$. Finally, if the variance of $F$ is large, then $F(x)$ is smaller for any given $x$ given the earlier restrictions on $F$. As a result, $1 - F(x)$ is larger (increasing).

## A.2 Proof of Proposition 2

If screeners are biased but not noisy, then $b > 0$ and the variance of $F$ is zero. As a result, the probability that the human will pick the correct (Type 1) candidate is $q = 0$, because $\eta$ will have to be positive to offset the bias. If $F$ has mean and variance of zero, $\eta$ will never be positive.

The machine learning engineer will receive a training dataset consisting of half Type 0, half Type 1s. The Type 1 candidates will all have an outcome $y$ equal to zero. The Type 0 candidates will have outcomes $y$ that are 1 with probability $p_0$, and 0 with probability $1 - p_0$. Thus the machine predictions of $E[Y|\theta = 0]$ will converge to $p_0$, and $E[Y|\theta = 1]$ will converge to zero.

The engineers will then pick the Type with the greatest estimated $E[Y|\theta]$, and will implement the decision consistently. Because $p_0 > 0$, the machine will prefer the Type 0 candidates. The original, underlying human process will yield $p_0$ successful candidates. The machine learning algorithm will yield the same output, creating zero difference in the quality of decison-making.

## A.3 Proof of Proposition 3

If screeners exhibit zero bias, then $b = 0$, they prefer Type 1 but pick Type 0 only if a noise realization $\eta$ is sufficiently low. This will happen with probability $1 - q$. If $b = 0$, noisier $F$s increase the probability of switching the decision-maker away from his/her default. In this setting, the default is to pick Type 1 (the better candidate), and so noisier $F$s will decrease $q$ (as discussed in Proposition 1).

The machine learning engineer will receive a training dataset consisting of half Type 0, half Type 1s. The Type 1 candidates will have outcomes $y$ that are 1 with probability $qp_1$, and 0 with the remaining probability. The Type 0 candidates will have outcomes $y$ that are 1 with probability $(1 - q)p_0$, and 0 with the remaining probability.

Thus the machine predictions of $E[Y|\theta = 1]$ will converge to $qp_1$, and $E[Y|\theta = 0]$ will converge to $(1 - q)p_0$. The engineers will then pick the Type with the greatest estimated $E[Y|\theta]$, and will implement the decision consistently. Because $qp_1 > (1 - q)p_0$, the machine will prefer the Type 1 candidates.

The original, underlying human process will yield $qp_1 + (1 - q)p_0$ successful candidates. The machine learning algorithm will yield $p_1$ successful candidates, for a output difference in output

of $(p_1 - p_0)(1 - q)$. Note that this is increasing in the quality difference between Type 1 and Type 0 candidates $(p_1 - p_0)$ and in the amount of noise. Because $b = 0$, as the variance of $F$ goes up, $q$ goes down.

### A.4 Proof of Proposition 4

If screeners are biased, then $b > 0$ and greater variance in $F$ increases $q$, the probability that Type 1 will be selected. The machine's predictions will converge to $E[Y|\theta = 0] = (1 - q)p_0$ for Type 0 and $E[Y|\theta = 0] = qp_1$ for Type 1.

The ML engineers will select Type 1 if $qp_1 > (1 - q)p_0$, and will otherwise select Type 0. It will select Type 0 if $q < p_0/(p_1 + p_0)$. For a given $r$, $b$, $p_1$ and $p_0$, this happens only if the noise function $F$ does not have sufficiently large variance.

If the variance in $F$ is not sufficiently large, the algorithm will implement Type 0 (the wrong choice) consistently. Performance under this algorithm will yield $p_0$ candidates. This is worse for the performance than if the original human process had been used, which would yield $qp_1 + (1 - q)p_0$.

### A.5 Proof of Proposition 5

If screeners are biased, then $b > 0$ and greater variance in $F$ increases $q$, the probability that Type 1 will be selected. The machine's predictions will converge to $E[Y|\theta = 0] = (1 - q)p_0$ for Type 0 and $E[Y|\theta = 0] = qp_1$ for Type 1.

The ML engineers will select Type 1 if $qp_1 > (1 - q)p_0$, and will otherwise select Type 0. It will select Type 1 if $q > p_0/(p_1 + p_0)$. For a given $r$, $b$, $p_1$ and $p_0$, this happens only if the noise function $F$ has sufficiently large variance.

### A.6 Proof of Proposition 6

Even if screeners are biased and sufficiently noisy to reduce bias ($q > p_0/(p_1 + p_0)$, see Proposition 5), the algorithm's predictions still underestimate the advantage of Type 1 above Type 0. In particular, the algorithm predicts a $y$ of $qp_1$ for Type 1 and $(1 - q)p_0$ for Type 0. The algorithm's implicit quality ratio of Type 1 over Type 0 is $qp_1/(1 - q)p_0$. This is less than the quality ratio of Type 1 over Type 0 $(p_1/p_0)$ – unless noise is maximized by increasing the variance of $F$ until $q = 1/2$.

### A.7 Proof of Proposition 7

In the presence of bias, the algorithm can reduce the bias if $q > p_0/(p_1 + p_0) = \underline{q}$. Recall from Proposition 1 that $q$ is decreasing in $b$ and increasing in the variance of $F$. To achieve any arbitrary $q > \underline{q}$ (for a fixed $r$, $p_1$ and $p_0$), either the variance of $F$ can stay fixed and bias $b$ can go down, or the bias $b$ can stay fixed and the variance can increase.

## A.8 Proof of Proposition 8

In the presence of bias, the algorithm can reduce the bias if $q > p_0/(p_1 + p_0) = \underline{q}$. This proof will examine the false positive and false negative rates in the training data of this setting. In this setting, Proposition 5 shows that the machine learning algorithm will select Type 1. Proposition 7 shows that as the variance of $F$ increases, the algorithm will be able to reduce smaller and smaller biases.

I will study two particular measures of goodness-of-fit: precision and recall, as these are commonly used in the machine learning literature. In this setting, recall measures: "If a candidate is selected by a machine, what's the probability that a human screener would have picked it as well?" Precision measures, "If the candidate is selected by the human, what's the probability that the machine would have selected it?" Precision and recall can vary between zero and one, with higher values corresponding to higher goodness-of-fit.

In our setting, recall is $q$ (the probability that the human would pick a machine approved candidate of Type 1). Precision is $q/2$. Note that these move in the same direction as a function of the primitives ($p_1$, $p_0$, $b$ and $r$) as discussed in Proposition 1. Both precision and recall are increasing in the amount of noise.

However, recall that in our setup, $q \in [0, 1/2]$. This means that precision cannot go above 0.5 and cannot go above 0.25. These are relatively low benchmarks. By comparison, both precision and recall on the training data are 1 from Proposition 2, where screeners exhibit bias but no noise.

## A.9 Proof of Proposition 9

The "coefficient" or "weight" the machine learning algorithm places on feature $\theta = 0$ is $E[Y|\theta = 1] - E[Y|\theta = 0]$. This is the difference in the algorithm's expected score for Type 1 and Type 0 candidates. The treatment effect of the machine learning algorithm on Type 1 candidates equals the change in the probability of Type 1 candidates being selected between the human decision-makers and algorithm.

I will show that these two quantities are not generically equal or even the same sign. I will present two counterexamples:

- Suppose screeners are biased, but NOT sufficiently noisy. As a result the algorithm codifies bias rather than reduces it ($q < p_0/(p_1 + p_0)$). The coefficient or weight in the machine learning algorithm will equal $qp_1 - (1 - q)p_0$.

  In the human regime, Type 1's probability of being selected is $q < 1$, and using the algorithm the probability is 0. The treatment effect is equal to $-q$, which is not generally equal to the coefficient or weight featured in the algorithm ($qp_1 - (1 - q)p_0$) and not even the same sign.

- Now suppose screeners are biased, and sufficiently noisy. As a result the algorithm improves bias ($q > p_0/(p_1 + p_0)$). The coefficient or weight in the machine learning algorithm will again equal $qp_1 - (1 - q)p_0$.

  In the human regime this value is $q < 1$, and using the algorithm the probability is 1. The treatment effect is equal to $1 - q$, which is not generally equal to the coefficient or weight featured in the algorithm ($qp_1 - (1 - q)p_0$).

### A.10 Alternative Microfoundations to Noise and Bias

# B Propositions with Additional Bias

## B.1 Microfoundations of $\delta$

As discussed in Section 4.2, bias may come both in choices to hire as well as how to evaluate candidates who have been hired. In this model, $b$ represents the amount of bias in hiring, and $\delta$ represents how much Type 1 candidates are discriminated against in evaluation.

Like $b$, $\delta$ can have several microfoundations. One possibility, presented above, is that taste-based discrimination is responsible for $\delta$. In this microfoundation, an evaluator experiences direct utility gains for underreporting Type 1's success.

However, $\delta$ can also represent statistical discrimination by an evaluator. An evaluator may not be able to directly measure a worker's productivity. This is a realistic assumption in many settings, where a worker's output cannot be fully monitored.

For example: A salesperson may be hired to invest in relationships with clients. Her manager may not be able to observe all aspects of relationship-building, and the outcome of this activity may take years to realize in firm revenue. The manager may therefore use a worker's observable characteristics to make educated guesses about the quality of the work that's hard to directly observe. If these inferences are wrong and slow to update, this will lead to a $\delta \in [0, 1]$ based on statistical discrimination.

## B.2 Additional Propositions

**Proposition 10.** *The screener's probability of picking Type 1 candidates ($q$), is decreasing in $b$, increasing in $r$, increasing in the quality difference in Type 1 and Type 0 ($p_1 - p_0$), decreasing in $\delta$ and increasing in the variance of $F$.*

*Proof.* A risk-neutral human screener will make the "right" decision (Type 1) if $rp_1(1-\delta) + \eta > rp_0 + b$. In other words, the screener makes the right decision if the random utility shocks are enough to offset the taste-based bias ($b$) favoring Type 0.

Let $\underline{\eta} = r(p_0 - p_1(1-\delta)) + b$ be the minimum $\eta$ necessary to offset the bias, given the other rewards involved. Such an $\eta$ (or greater) happens with probability of $\Pr(\eta > r(p_0 - p_1(1-\delta)) + b) = 1 - F(r(p_0 - p_1(1-\delta)) + b)) = q$.

Note that $q = 1 - F(x)$, which is decreasing in the argument $x$. In our setting, $x = r(p_0 - p_1(1-\delta)) + b$, which is increasing in $b$ and so $1 - F(x)$ is decreasing in $b$. Note that $p_1 > p_0$, so $x$ is also decreasing in $r \implies 1 - F(x)$ is increasing in $r$. Also: $x$ is decreasing in $p_1 - p_0$, so $1 - F(x)$ is increasing in $p_1 - p_0$. $x$ is increasing in $\delta$, so $1 - F(x)$ is decreasing in $\delta$. Note that the addition of $\delta$ to the model will make $q$ smaller than in the original Proposition 1, unless $\delta = 0$. Finally, if the variance of $F$ is large, then $F(x)$ is smaller for any given $x$ given the earlier restrictions on $F$. As a result, $1 - F(x)$ is larger (increasing). □

**Proposition 11.** *If bias in the evaluation $\delta$ is above a threshold, the machine learning approach will entrench bias irrespective of the amount of noise. This threshold value is decreasing in the quality of Type 1.*

Thus the machine predictions of $E[Y|\theta = 1]$ will converge to $qp_1(1-\delta)$, and $E[Y|\theta = 0]$ will converge to $(1-q)p_0$. The algorithm picks Type 1 if $qp_1(1-\delta) > (1-q)p_0$, which happens only if $\delta < (p_1 + p_0)/p_1$, which is decreasing in $p_1$, increasing in $p_0$ and decreasing in the quality difference $p_1 - p_0$.

**Proposition 12.** *If screeners exhibit bias but zero noise, the algorithm will perfectly codify the humans historical bias. The algorithm's perfomance will precisely equal that of the biased screeners.*

*Proof.* If screeners are biased but not noisy, then $b > 0$ and the variance of $F$ is zero. In other words, the screener makes the right decision if the random utility shocks are enough to offset the taste-based bias ($b$) against hiring Type 1 and against scoring Type 1. As a result, the probability that the human will pick the correct (Type 1) candidate is $q = 0$, because $\eta$ will have to be positive to offset the bias. If $F$ has mean and variance of zero, $\eta$ will never be positive. This is essentially unchanged from Proposition 12.

The machine learning engineer will receive a training dataset consisting of half Type 0, half Type 1s. The Type 1 candidates will all have an outcome $y$ equal to zero. The Type 0 candidates will have outcomes $y$ that are 1 with probability $p_0$, and 0 with probability $1 - p_0$. Thus the machine predictions of $E[Y|\theta = 0]$ will converge to $p_0$, and $E[Y|\theta = 1]$ will converge to zero.

The engineers will then pick the Type with the greatest estimated $E[Y|\theta]$, and will implement the decision consistently. Because $p_0 > 0$, the machine will prefer the Type 0 candidates. The original, underlying human process will yield $p_0$ successful candidates. The machine learning algorithm will yield the same output, creating zero difference in the quality of decison-making. □

**Proposition 13.** *If screeners exhibit i. zero bias in hiring ($b = 0$), ii. zero or sufficiently low bias in scoring hired workers ($\delta < (p_1 - p_0)/p_1$), but iii. non-zero amounts of noise, the algorithm will improve upon the performance of the screeners by removing noise. The amount of performance improvement is increasing in the amount of noise and the quality difference between Type 1 and Type 0 candidates.*

If screeners exhibit zero bias then $b = 0$, they prefer Type 1 but pick Type 0 only if a noise realization $\eta$ is sufficiently low. This will happen with probability $1 - q$. If $b = 0$, noisier $F$s increase the probability of switching the decision-maker away from his/her default. In this setting, the default is to pick Type 1 (the better candidate), and so noisier $F$s will decrease $q$ (as discussed in Proposition 1).

The machine learning engineer will receive a training dataset consisting of half Type 0, half Type 1s. The Type 1 candidates will have outcomes $y$ that are 1 with probability $qp_1(1-\delta)$, and 0 with the remaining probability. The Type 0 candidates will have outcomes $y$ that are 1 with probability $(1-q)p_0$, and 0 with the remaining probability.

Thus the machine predictions of $E[Y|\theta = 1]$ will converge to $qp_1(1-\delta)$, and $E[Y|\theta = 0]$ will converge to $(1-q)p_0$. The engineers will then pick the Type with the greatest estimated $E[Y|\theta]$, and will implement the decision consistently. Because $qp_1(1-\delta) > (1-q)p_0$, the machine will prefer the Type 1 candidates.

The original, underlying human process will yield $qp_1(1-\delta)+(1-q)p_0$ successful candidates. The machine learning algorithm will yield $p_1(1-\delta)$ successful candidates, for a output difference in output of $(p_1(1-\delta)-p_0)(1-q)$. Note that this is decreasing in $\delta$, increasing in the quality difference between Type 1 and Type 0 candidates $(p_1-p_0)$, and in the amount of noise. Because $b=0$ and $(\delta<(p_1-p_0)/p_1)$, as the variance of $F$ goes up, $q$ goes down.

**Proposition 14.** *If screeners and evaluators are biased ($b>0$ and $\delta>0$) are NOT sufficiently noisy, the algorithm will codify bias. The reduction in noise will actually make outcomes worse.*

If screeners are biased, then $b>0$ and greater variance in $F$ increases $q$, the probability that Type 1 will be selected. The machine's predictions will converge to $E[Y|\theta=0]=(1-q)p_0$ for Type 0 and $E[Y|\theta=0]=qp_1(1-\delta)$ for Type 1.

The ML engineers will select Type 1 if $qp_1(1-\delta)>(1-q)p_0$, and will otherwise select Type 0. It will select Type 0 if $q<p_0/(p_1(1-\delta)+p_0)$. For a given $r$, $b$, $p_1$, $\delta$ and $p_0$, this happens only if the noise function $F$ does not have sufficiently large variance.

If the variance in $F$ is not sufficiently large, the algorithm will implement Type 0 (the wrong choice) consistently. Performance under this algorithm will yield $p_0$ candidates. This is worse for the performance than if the original human process had been used, which would yield $qp_1(1-\delta)+(1-q)p_0$.

**Proposition 15.** *If screeners are biased and sufficiently noisy, the algorithm will reduce the humans' bias.*

If screeners are biased, then $b>0$ and greater variance in $F$ increases $q$, the probability that Type 1 will be selected. The machine's predictions will converge to $E[Y|\theta=0]=(1-q)p_0$ for Type 0 and $E[Y|\theta=0]=qp_1(1-\delta)$ for Type 1.

The ML engineers will select Type 1 if $qp_1(1-\delta)>(1-q)p_0$, and will otherwise select Type 0. It will select Type 1 if $q>p_0/(p_1(1-\delta)+p_0)$. For a given $r$, $b$, $p_1$, $\delta$ and $p_0$, this happens only if the noise function $F$ has sufficiently large variance.

**Proposition 16.** *If the algorithms' human data contains non-zero bias then, "algorithmic bias" cannot be reduced to zero unless the humans in the training data were perfectly noisy (ie, picking at random).*

Even if screeners are sufficiently noisy to reduce bias (as in Proposition 15), the algorithm's predictions still underestimate the advantage of Type 1 above Type 0 – unless the training data were perfectly representative (ie, humans were picking workers at random). Despite the reduction in bias, the algorithm will remain handicapped and exhibit some bias because of its training on biased training data.

Picking at random is extremely unlikely to appear in any real-world setting, since the purpose of most hiring is to select workers who are better than average and thus undersample sections of the applicant pool perceived to be weaker. A complete removal of bias therefore appears infeasible from training datasets from real-world observations, particularly observations of agents who are *not* optimizing labels for *ex-post* learning.

Even if screeners are biased and sufficiently noisy to reduce bias ($q>p_0/(p_1(1-\delta)+p_0)$, see Proposition 15), the algorithm's predictions still underestimate the advantage of Type 1 above

Type 0. In particular, the algorithm predicts a $y$ of $qp_1$ for Type 1 and $(1-q)p_0$ for Type 0. The algorithm's implicit quality ratio of Type 1 over Type 0 is $qp_1/(1-q)p_0$. This is less than the quality ratio of Type 1 over Type 0 ($p_1/p_0$) – unless noise is maximized by increasing the variance of $F$ until $q = 1/2$.

**Proposition 17.** *The minimum amount of noise necessary for the machine learning to reduce bias is a* decreasing *function of the amount of bias.*

Proposition 17 means that if the screeners display a large amount of bias, only a small amount of noise is necessary for the machines to correct the bias. Similarly if screeners display a small amount of bias, then high amounts of noise are necessary for the algorithm to correct the bias.

The intuition behind Proposition 17 is as follows: Suppose that screeners were highly biased against Type 1 workers, this would conceal the large productivity differences between Type 1 and Type 0 candidates. The machine learning algorithm would need to see only a few realizations – a small amount of noise – in order to reduce the bias. Because each "experiment" on Type 1 workers shows so much greater productivity, few such experiments would be necessary for the algorithm to learn the improvement.

By contrast, if the bias against Type 1 is small – large amounts of noise would be necessary for the algorithm to learn its way out of it. This is because each "experiment" yields a smaller average productivity gain. As a result, the algorithm requires more observations in order to understand the gains from picking Type 1 candidates.

In the presence of bias, the algorithm can reduce the bias if $q > p_0/(p_1 + p_0) = \underline{q}$. Recall from Proposition 10 that $q$ is decreasing in $b$ and increasing in the variance of $F$. To achieve any arbitrary $q > \underline{q}$ (for a fixed $r$, $p_1$ and $p_0$), either i) the variance of $F$ can stay fixed and bias $b$ can go down, or ii) the bias $b$ can stay fixed and the variance can increase.

**Proposition 18.** *In settings featuring bias sufficiently high noise, the algorithm's improvement in bias will be positive and increasing in the level of noise and bias. However, metrics of goodness-of-fit on the training data (and on additional observations from the data-generating process) are* decreasing *in the amount of noise and bias.*

The implication of Proposition 18 is: If engineers avoid settings where models exhibit poor goodness-of-fit on the training data (and future samples), they will avoid the settings where machine learning has the greatest potential to reduce bias.

**Proposition 19.** *The "coefficient" or "weight" the machine learning algorithm places on $\theta = 1$ when ranking candidates does not equal the treatment effect of using the algorithm rather than human discretion for $\theta = 1$ candidates.*

Proposition 19 discusses how observers should interpret the coefficients and/or weights of the machine learning algorithm. It shows that these weights may be highly misleading about the impact of the algorithm. For example: It's possible for an algorithm that places negative weight on $\theta = 1$ when ranking candidates could nonetheless have a strong positive benefit for $\theta = 1$ candidates and their selection outcomes. This would happen if the human penalized these characteristics even more than the algorithm did.

The internal weights of these algorithms are completely unrelated to which candidates benefit from the algorithm compared to a status quo alternative. The latter comparison requires a comparison to a counterfactual.

The "coefficient" or "weight" the machine learning algorithm places on feature $\theta = 0$ is $E[Y|\theta = 1] - E[Y|\theta = 0]$. This is the difference in the algorithm's expected score for Type 1 and Type 0 candidates. The treatment effect of the machine learning algorithm on Type 1 candidates equals the change in the probability of Type 1 candidates being selected between the human decision-makers and algorithm.

I will show that these two quantities are not generically equal or even the same sign. I will present two counterexamples:

- Suppose screeners are biased, but NOT sufficiently noisy. As a result the algorithm codifies bias rather than reduces it ($q < p_0/(p_1 + p_0)$). The coefficient or weight in the machine learning algorithm will equal $qp_1 - (1 - q)p_0$.

  In the human regime, Type 1's probability of being selected is $q < 1$, and using the algorithm the probability is 0. The treatment effect is equal to $-q$, which is not generally equal to the coefficient or weight featured in the algorithm ($qp_1 - (1 - q)p_0$) and not even the same sign.

- Now suppose screeners are biased, and sufficiently noisy. As a result the algorithm improves bias ($q > p_0/(p_1 + p_0)$). The coefficient or weight in the machine learning algorithm will again equal $qp_1 - (1 - q)p_0$.

  In the human regime this value is $q < 1$, and using the algorithm the probability is 1. The treatment effect is equal to $1 - q$, which is not generally equal to the coefficient or weight featured in the algorithm ($qp_1 - (1 - q)p_0$).