Algorithmic Bias is Data Bias

How data shapes the predictive power of machine learning algorithms

Ellorine Carle

Independent Study Research Paper
Barnard College
Spring 2018

INTRODUCTION

Over the past decade, machine learning algorithms have reshaped information exchange, prediction, and decision-making. They facilitate how we find information on the internet,[1] enable revolutionary technology such as self-driving cars,[2] and influence crucial verdicts, such as for how long we go to prison, for what jobs we are hired, and whether or not we qualify for social benefits.[3] Because of their flexible framework, machine learning techniques have proved widely successful in a number of applications ranging from translation,[4] to medical diagnoses,[5] and résumé screening.[6] However, concerns about bias, privacy, and the overall effectiveness of machine learning algorithms have materialized as well. Academics have uncovered "algorithmic bias" in models used in judicial decision-making,[7] hiring,[8] and targeted advertising.[9] Policymakers in the E.U. have carved legislation aimed to check the power of algorithmic decision-making, passing sweeping regulation which comes into effect next month.[10] A number of popular books[11] and news articles on the topic suggest that unease proliferates beyond academia and politics. If the last decade was about realizing the power of machine learning, the next may be about grappling with its limitations.

There are apparent dichotomies in how well machine learning algorithms perform: sometimes, they demonstrate incredible predictive power, and other times,

[1] Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.
[2] Wing, Jeannette. "Data for Good". Lecture delivered for Columbia ADI. Columbia University, New York, April 3, 2018.
[3] O'Neil, Cathy. *Weapons of Math Destruction.* S.l.: PENGUIN BOOKS, 2017.
[4] Lewis-Kraus, Gideon. "The Great A.I. Awakening." *New York Times Magazine*, December 14, 2016. Accessed April 29, 2018.
[5] Mukherjee, Siddhartha. "A.I. versus M.D." *The New Yorker*, April 3, 2017. Accessed April 29, 2018.
[6] Cowgill, Bo. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening." *Working Paper*. March 16, 2018.
[7] Dressel, Julia and Hany Farid. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4, no. 1. January 17, 2018.
[8] Mann, Gideon and Cathy O'Neil. "Hiring Algorithms Are Not Neutral." *Harvard Business Review*, December 9, 2016.
[9] Datta, Amit, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings." *Proceedings on Privacy Enhancing Technologies* 2015, no. 1 April 18, 2015.
[10] Garcia, Megan. "Racist in the Machine: The Disturbing Implications of Algorithmic Bias." *World Policy Journal* 33, no. 4 (Winter 2016).
[11] Several recently published books include: Cathy O'Neil, *Weapons of Math Destruction*. Virginia Eubanks, *Automating Inequality*. Frank Pasquale, *The Black Box Society*. Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*.

they hardly do better than random chance; sometimes, they perpetuate human biases, and other times, they alleviate them. In order to understand these disparities, it suffices to start with a simple question: when are machine learning algorithms useful? This paper argues that the data used to build machine learning algorithms is the foremost determinant of their predictive power and capacity for mitigating human biases. Thus, it is not algorithmic bias, but data biases that should be of primary concern. This conclusion motivates further research in techniques to generate richer data, alleviate biases in existing data, and produce better-defined guidelines for how the technology should fit into decision-making processes.

The paper is divided into three parts. Part I will provide a general overview of machine learning algorithms: what they are, what it means for an algorithm to be useful, and scenarios in which this usefulness is realized or not. Part II will look at the particular application of machine learning algorithms to employee selection processes. Part III will examine potential solutions to the issues outlined in Parts I and II.

PART I: BACKGROUND

1. What is a machine learning algorithm?

An algorithm is no more than a set of instructions to be followed in order to solve a problem. A recipe for making an apple pie, GPS directions to the grocery store, and the process by which a neural network classifies an image as either a cat or a dog all fit under the algorithmic umbrella. The most useful features of algorithms are their replicability and reliability: they get from point A to point B every time and as many times as specified. In the abstract sense, algorithms have existed for all of human history as codes or rules, though computer technology has harnessed the power of algorithms in unprecedented ways.

When researchers began first thinking about computerized learning, they relied on rule-based learning, essentially a large collection of algorithms structured as "IF…THEN" statements defining how a machine would deduce an output from a given input.[12] For instance, in early computerized language translation, programmers hard-coded the vocabulary and syntax rules of each language.[13] This approach to computer learning proved quite limited in complex problems, which are susceptible to "combinatorial explosion"—as more inputs are added to the problem, the number of combinations that one has to examine grows exponentially, requiring an intolerable amount of computing power.[14] The issues with rule-based learning motivated a new approach called machine learning.

Machine learning relies on learning "by experience" rather than learning by formula and has been crucial in advancing artificial intelligence capabilities.[15] The machine learning approach has fewer defined rules for how the algorithm must deduce an output from a given input. In supervised learning, the machine learning technique involves showing an algorithm a set of observations with pre-labeled outcomes called the training dataset. The algorithm then determines how the characteristics of the observations are related to the outcome so that when it is shown a new observation, it can use its experience with similar observations to predict an unknown outcome. The driving force behind the success of machine learning algorithms has been the rise of "Big Data"—large amounts of usable complex, unstructured, time-sensitive granular information[16]—which provides the algorithms with rich and expansive combinations of observations and outcomes to learn from. Analogous to the fact that humans get

[12] Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16, no. 3 (2001): 199-215.

[13] Lewis-Kraus, Gideon. "The Great A.I. Awakening."

[14] Nilsson, Nils J. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge ; New York: Cambridge University Press, 2010.

[15] *Ibid*.

[16] Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt, 2013.

smarter as they acquire new information about the world, it is generally the case that as machine learning algorithms are exposed to more data, they generate better predictions.

Machine learning algorithms are not only a replacement for rule-based learning in the realm of artificial intelligence, but also an alternative to statistical modeling techniques used for prediction tasks.[17] Unlike statistical models which require assumptions about the underlying distributions of the data and specification of a functional form, machine learning algorithms only require choosing the complexity parameter of the model. Furthermore, machine learning algorithms are better at handling "wide" datasets than statistical modeling techniques because they can incorporate more right-hand-side variables.[18] For instance, it is unfeasible to add a large number of variables and interaction terms to a logistic regression because it leads to more regressors than data points.[19] In contrast, a machine learning algorithm searches for complex interactions between variables by design. As Mullainathan and Spiess observe in a comparison of the two approaches, "The very appeal of machine learning is high dimensionality: flexible functional forms allow us to fit varied structures of the data."[20]

The flexibility sanctioned by machine learning techniques poses their biggest tradeoff. A model perfectly fitted to the training dataset means that each observation's outcome are precisely predicted. However, the model will have over-fitted to the noise and errors in the training data and will therefore fare poorly when making predictions on new observations. In machine learning, this phenomenon is called the bias-variance tradeoff: models fitted tightly to the training data accurately predict individual data points but change dramatically if the training set is altered (high variance, low bias),

[17] Breiman, Leo. "Statistical Modeling: The Two Cultures."

[18] Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives.* 31, no. 2 (Spring 2017): 87-106.

[19] *Ibid.*

[20] *Ibid.*

whereas models capturing very general trends in the training data inaccurately predict individual data points but change much less if the training set changes (low variance, high bias).[21] Because the goal of machine learning is to generate accurate predictions about new data, procedures of regularization and empirical tuning address these tradeoffs to optimize the algorithm's out-of-sample predictive power. Regularization involves imposing limits on the model's complexity so that only a set number of relationships are allowed in the model.[22] For example, with decision trees, an elementary machine learning method, limits can be placed on the depth of the tree, number of features used, and number of observations at the end of each node. Empirical tuning involves evaluating the performance of models with varying degrees of complexity on new observations to choose the model which makes the most accurate out-of-sample predictions.[23] Overall, the goal of these procedures is to find a model which accurately captures the regularities of the training data but generalizes well to new data.

2. What does it mean for a machine learning algorithm to be useful?

The goal of machine learning algorithms is to make predictions, so the most basic notion of an algorithm's usefulness is whether or not its predictions are accurate. For example, an algorithm used to classify pictures as cats or dogs is useful if it correctly identifies cats as cats and dogs as dogs. But usefulness is a relative concept that will vary depending on the risks of misclassifying, and the efficacy of alternative methods of generating the predictions. For instance, a cat/dog classifier with only 80% accuracy might still be useful if the time and money saved by using the machine outweighs the greater accuracy that human classifiers may achieve. In contrast, a self-driving vehicle

[21] Munro, Paul. "Bias Variance Decomposition." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Web, 100-01. Springer, 2011.
[22] Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach."
[23] *Ibid*.

algorithm may only be implemented if the image recognition system accurately identifies a stop sign more than 99.99% of the time because of the high cost of misclassification.[24] If we know that the decisions made by human résumé screeners are no more accurate than random chance, a machine learning algorithm may still produce valuable insights even if it identifies the best job applicants only 60% of the time because of the significant financial impact of marginally improving predictive power.

3. What factors regulate the predictive power of a machine learning algorithm?

An algorithm's predictive power is determined by the quantity of data it is exposed to, but more importantly, to what extent there is variation in the data. As the size of the training dataset increases, the likelihood of the algorithm learning greater nuances, more complex relationships, and variation in outcomes increases, enhancing the algorithm's understanding of the relationships between the explanatory variables and the outcome. However, this is not always the case if a larger quantity of data does not provide any new information. For instance, in the cat/dog classifier, an algorithm exposed to 1,000 images of many types of dogs will probably have a better understanding of what differentiates a cat from a dog than an algorithm which has seen 1,000,000 Golden Retrievers.

The quantity, variation, and nature of the observations in the training dataset reflect how much bias and noise—the two error types—are present in the data. Bias refers to when predictions systematically deviate from the true values and encompasses modeling bias and data bias. Data bias denotes the predictions already encoded in the raw training data, and whether or not they are representative of the truth. Modeling bias denotes the erroneous assumptions made by the learning algorithm when it

[24] Watzenig, Daniel, and Martin Horn. *Automated Driving: Safer and More Efficient Future Driving*. Cham: Springer International Publishing, 2017.

generates predictions.[25] A model with no bias would perfectly predict each outcome in the training data, at the expense of its generalizability and predictive power. "Biased algorithms" are a growing concern in the popular media, academia, and politics. However, the accusations of bias in algorithms are misleading and imprecise. A machine learning algorithm represents no more than a framework for making a prediction based on a massive quantity of similar past outcomes. When the data fed into the algorithm are biased in some way, the algorithm will generate biased outcomes because it is designed to draw conclusions from the examples it is given. Most journalists and consumers are not concerned with how tightly the algorithm was fit to the training data. Rather, their claims of algorithmic bias primarily refer to data biases.

The other form of error which may arise in the data is noise. Noise is the random variability of a measurement around its true value. Noise occurs when taking objective measurements, like taking one's temperature with a cheap thermometer, as well as in subjective decision-making, such as valuing stocks, appraising real estate, and evaluating job performance. Noise does not exhibit a systematic pattern in how it deviates from the true value and therefore cannot be formally accounted for in the model.[26]

Bias and noise in the training data can affect the machine learning algorithm's predictions in several different ways. Cowgill shows three relationships between errors in the training data and the algorithm's performance relative to alternate prediction methods (such as a human decision-maker):[27]

1) *Machine learning algorithms will produce better predictions when the training data are noisy but unbiased.*

[25] Munro, Paul. "Bias Variance Decomposition."

[26] Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach."

[27] Cowgill, Bo. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening."

2) *When the training data exhibits bias but insufficient noise, the algorithm will perpetuate and possibly intensify bias.*
3) *When the training data exhibits bias but sufficient noise, the algorithm will reduce bias and improve predictions.*

The first proposition illustrates how machine learning algorithms can reduce noise in predictions. The algorithm optimizes the weights placed on the available right-hand-side variables and applies them consistently to new observations to maximize the number of correct predictions. When given an identical observation, the machine learning algorithm will always generate the same result. The second proposition shows that when the data is biased, the algorithm does not have enough freedom to understand that the biased outcomes are actually erroneous. The third proposition motivates the most opportunity for machine learning algorithms. It illustrates that algorithms can still prove useful even when the data they are trained on exhibit some bias—an inevitability in many practical applications of machine learning algorithms today.

4. When are machine learning algorithms useful?

This section presents instances in which machine learning algorithms succeed, or at least prove more effective than alternate prediction methods.

A. Algorithms outperform human decision makers when human decisions are noisy, and potentially exhibit some bias.

Research by Kleinberg et al.,[28] Dawes et al.,[29] and Kahneman et al.[30] show that human judgement calls—decisions which do not adhere to rigid rules—are often quite noisy. In some cases, humans may not reach the same conclusion upon seeing the same data twice.[31] Professionals from fields such as medicine, consulting, real estate, hiring, and computing, often make decisions that "deviate from those of their peers, from their own prior decisions, and from rules they themselves claim to follow."[32] Machine learning technology has already demonstrated superior predictive capabilities in many of these areas, precisely because of this fact. For instance, in medicine, an algorithm proved significantly better at diagnosing skin conditions than board certified dermatologists.[33] Kleinberg et al. find that machine learning algorithms used to approve or deny criminal defendants bail are particularly effective because judges who usually make the decisions tend to respond to noise as if it were a signal.[34] The researchers show that "internal states, such as [the judge's] mood" and "specific features of the case that are salient and over-weighted" are sources of noise in human decision-making.[35] Replacing the judges by an algorithm leads to reductions in all categories of crime in addition to reducing the percentage of incarcerated African Americans and Latinos.[36]

B. *Algorithms are effective when decision-making involves a large number of factors and frequently updating beliefs in response to new data.*

[28] Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human Decisions and Machine Predictions." *NBER Working Paper*, February 2017.

[29] Dawes, Robyn M., David Faust, and Paul E. Meehl. "Clinical versus Actuarial Judgment." *Heuristics and Biases*, 1989, 716-29. Accessed April 29, 2018.

[30] Kahneman, Daniel, Andrew Rosenfield, Linnea Gandhi, and Tom Blaser. "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making." *Harvard Business Review*, October 2016.

[31] *Ibid*.

[32] *Ibid*.

[33] Esteva, Andre, Brett Kuprel, Roberto Novoa, and Justin Ko. "Dermatologist-level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542 (January 25, 2017): 115-18.

[34] Kleinberg, Jon et al. "Human Decisions and Machine Predictions."

[35] *Ibid*.

[36] *Ibid*.

Algorithms provide a framework to understand how a large collection of variables produce specific outcomes. In contrast, humans have difficultly analyzing more than one thing at a time,[37] so it is no surprise that when bombarded with lots of information, they only pay attention to small fractions of it. In the psychology literature, the study of multiple cue probability learning reveals that humans are unable to predict an outcome once it is related to more than two or three cues in a non-linear way.[38] Brehmer argues further that people do not have the cognitive mechanisms to solve probabilistic tasks.[39] Experimental studies that ask humans to identify relationships between cues and an outcome reveal that humans tend to generate deterministic hypotheses about the underlying relationships (i.e. A leads to B) as opposed to probabilistic ones (i.e. A may lead to B or C).[40] Machine learning algorithms operate under a probabilistic framework and can easily handle hundreds of factors and identify patterns that are incomprehensible to humans.

Besides providing a mechanism for analyzing large swaths of inputs, machine learning algorithms are also capable of updating their beliefs about the relations between factors in a systematic and consistent fashion. As the algorithms are exposed to a greater amount of data, they are constantly adjusting the weights given to different factors in order to maximize predictive power. One example of an algorithmic solution of this nature is a model developed by Chouldechova et al. used to determine whether children are at risk of abuse at home.[41] Cases involve a complex set of time-dynamic factors which are difficult for understaffed child services agencies to track and make sense of. Over a 16-month period of implementation, the algorithm led to fewer

[37] Hamilton, Jon. "Think You're Multitasking? Think Again." *NPR*, October 2, 2008. Accessed April 29, 2018.

[38] Harvey, N., and Fischer, I. Development of experience-based judgment and decision making: The role of outcome feedback. In T. Betsch & S. Haberstroh (Eds.), *The routines of decision making* (pp. 119-137). 2005.

[39] Brehmer, B. "In one word: Not from experience." Acta Psychologica, 45, 223-241. 1980.

[40] Ibid.

[41] Chouldechova, Alexandra, Emily Putnam-Hornstein, and Diana Benavides-Prado. "A Case Study of Algorithm-assisted Decision Making in Child Maltreatment Hotline Screening Decisions." *Proceedings of Machine Learning Research*81 (2018): 1-15.

investigations of low-risk cases, giving time-strapped case workers more flexibility to deal with high-risk cases.[42]

C. Defining precise goals for prediction.

Algorithms demonstrate usefulness when the outcome the algorithm is predicting is narrow and well-defined. Within highly specific scenarios, the number of factors which cannot be accounted for by the model are minimized, augmenting the algorithm's predictive power. For illustration, consider the algorithms used in baseball which are designed to predict very specific outcomes within the larger game. The models simulate precise scenarios, such as the likelihood of getting a specific player out if they face a specific pitcher, or whether is it favorable to have a certain batter bunt to get a runner from first to second base in the ninth inning.[43] All of these factors ultimately contribute to the primarily prediction goal of determining who will win the game (or the World Series) but are never built to address that question on its own. Recognizing that a win is made up of many smaller "wins," the algorithms optimize on smaller decisions made throughout the game which compound into the team's ultimate success.

D. Effective algorithms are able to learn from large, rich, relevant data sources.

The rise of big data has vastly increased the potential data sources from which machine learning algorithms can learn. Algorithms built using data is that relevant to the outcome tend to prove most successful, as in the case of the baseball algorithms above, which utilize a rich number of game and player metrics clearly linked to performance.[44] Another example of a successful algorithmic application is the FICO

[42] Hurley, Dan. "Can an Algorithm Tell When Kids Are in Danger?" *New York Times Magazine*, January 2, 2018. Accessed April 29, 2018.

[43] O'Neil, Cathy. *Weapons of Math Destruction*.

[44] *Ibid*.

credit score, a mathematical formula developed in the 1950s to predict credit-worthiness.[45] The algorithm uses information about an individual's financial history—payments, amounts owed, length of credit history, types of credits used, and new credit—all of which provide information that directly affect a person's demonstrated ability to pay off a loan.[46]

E. The algorithm is useful if it is better than alternatives.

Machine learning algorithms are one of many techniques for making predictions. Other ways of formulating predictions could be via the human brain, a flip of a coin, a mechanical process, or with a logistic regression. In evaluating the effectiveness of a machine learning technique, it is pertinent to examine the alternatives, including other machine learning methods. Even if an algorithm's predictions are not very accurate, a different technique may be worse, which means the machine learning algorithm may still be useful. Kleinberg et al. estimate a substantial welfare gain from implementing the algorithm used for bail eligibility in the courts, showing that "crime can be reduced by up to 24.8% with no change in jailing rates, or jail populations can be reduced by 42.0% with no increase in crime rates," precisely because the algorithm outperforms the noisy decisions made by judges.[47]

5. When are algorithms not useful?

This section outlines some of the reasons why machine learning algorithms produce unfavorable outcomes.

A. The training data are biased, but do not exhibit enough variation.

[45] *Ibid.*

[46] *Ibid.*

[47] Kleinberg, Jon et al. "Human Decisions and Machine Predictions."

Because algorithmic learning is limited by the information in the training dataset, it is impossible to expect an algorithm trained only on biased decisions to then make unbiased decisions on new data. More alarmingly, if the data have high bias but a small amount of variation, the machine will further codify the bias because it will essentially remove the noise which may have produced unbiased decisions.[48] Consistently biased data is often a result of systematic issues with the data generating process. For example, a recent audit by Angwin et al. of an algorithm built to predict the likelihood that criminal defendants would commit another crime found an alarming disparity between how the model treated different racial groups.[49] The algorithm systematically overestimated that previously incarcerated African American would reenter the prison system, while underestimating that Whites would return to prison. The algorithm integrates a 137-question survey and existing arrest records to calculate the score.[50] Because of the use of arrest records as a proxy for criminal behavior, the training data are likely biased since in general, white collar crime is underreported,[51] while black men in impoverished inner cities are stopped by police even when they have not committed a crime.[52]

B. *The data sources lack richness and are unrelated to the outcomes.*

One of the most powerful features of machine learning algorithms is their ability to handle as many input variables as are available. While this might help alleviate issues of omitted variable bias, it also encourages prediction based on spurious correlations. Because big data processes have created many sources of usable data and

[48] Cowgill, Bo. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening."

[49] Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." *Pro Publica*, May 23, 2016. Accessed April 29, 2018.

[50] *Ibid.*

[51] Martinez, Joseph P. *Unpunished Criminals: The Social Acceptablity of White Collar Crimes in America*. Master's thesis, Eastern Michigan University, 2014.

[52] "Stop-and-Frisk Data." New York Civil Liberties Union. April 04, 2018. Accessed April 29, 2018. https://www.nyclu.org/en/stop-and-frisk-data.

the machine learning technique is focused solely on prediction, it is common for machine learning algorithms to incorporate any information that improves the algorithm's performance, regardless of its relatedness to the outcomes. In 2014, Google developed a machine learning algorithm to predict where the flu virus would spread using search query data.[53] The algorithm proved highly predictive at first; however, over time it became increasingly inaccurate, even as the inputs to the algorithm were updated. It performed especially poorly during non-seasonal breakouts because the machine was primarily tracking seasonality, not the spread of the flu itself.[54] Search volume for things like "high school basketball" picked up in the winter, and the algorithm was erroneously associating that with increased prevalence of the flu.[55] This example illustrates that machine learning algorithms lack robustness when learning primarily from proxies and spurious correlations. If the true causes of an outcome change, models which lack some understanding of these causal relationships are left in the dust.

C. *Feedback given to models is imbalanced or nonexistent.*

A pitfall of prediction in certain settings is that the decision-makers do not necessarily receive feedback on the efficacy of their projections. In college admissions, officers will never know how well applicants they reject would have fared at the university. In marketing, advertisers will never know how effective a version of an advertisement not shown to a consumer would have been. Algorithms are not immune to this problem either, since they cannot update their beliefs if they do not have access to the true outcomes of all of their decisions. An application of machine learning algorithms where this issue is likely to arise is in recommendation systems. For example, if a Netflix algorithm determines that a user overwhelmingly enjoys comedies

[53] Lazer, D., R. Kennedy, G. King, and A. Vespignani. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343, no. 6176 (2014): 1203-205.

[54] *Ibid*.

[55] *Ibid*.

based on her viewing history, it will display comedies on her screen. The algorithm will then learn which of the displayed comedies the user likes but will never learn whether the user would have liked other movies, perhaps ones from the horror genre, that it did not display. This creates a feedback loop because the user only sees comedies, so she continues to only watch comedies, and may never uncover her unrealized affinity for horror movies. Though this example is simplistic, it illustrates the gaps in the recommendation system's reliance on a user's historical preferences in predicting—and likely shaping—her future preferences. Chaney et al. show that feedback loops created by recommendation systems results in homogenization of user behavior because the "algorithms encourage similar users to interact with the same set of items."[56] Most importantly, homogenization does not result in increased utility for individual users because "based on their true preferences, they would enjoy a broader range of items."[57]

D. The algorithm's predictions do not precisely address the decision being made.

Machine learning algorithms can only predict one outcome variable, so the decision being made with the model should be equally limited in scope. Problems arise when algorithms predict a certain variable but then are used to make a decision embodying a collection of variables. For instance, in college admissions, someone reviewing an application will likely base their decisions on many predictions, such as the applicant's expected fit within the school's culture, their future academic success, and their contributions to the diversity of perspectives on campus. Somehow combining these factors into a single left-hand-side variable so that a machine learning algorithm could perform the task would prove exceedingly difficult, especially since all of the factors do not hold equal weight from candidate to candidate. The consequence might

[56] Chaney, Allison, Brandon Stewart, and Barbara Engelhardt. "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility." *Working Paper*, October 30, 2017.
[57] *Ibid*.

be that an algorithm is built to predict only one metric, such as GPA. It would be inappropriate for the algorithm's predictions to replace an admissions officer altogether, as the information contained in the algorithm's predictions only represents a fraction of what determines an applicant's fit at a university.

E. The algorithm is used to predict far off outcomes

When machine learning algorithms are used to predict an outcome occurring far into the future, variables that come into play after the prediction has been made gain greater influence over the outcome. This is analogous to the poor performance of long-term weather forecasting. As the number of days into the future a weather prediction is made for increases, a greater number of unknown future variables are unaccounted for by the model, compromising the forecast. It is therefore important to design machine learning algorithms that make predictions that are as specific as possible and occur in a reasonably small amount of time into the future. For instance, in building an algorithm to predict an applicant's expected college GPA, it might be more appropriate to limit the scope of the prediction to only their expected freshman GPA to minimize potential confounders.

F. The algorithm cannot solve root problems.

In their quest for perfect predictions, machine learning algorithms largely ignore questions of causation.[58]A machine learning technique can make a prediction, but it is unable to say why it made that prediction or identify which factors are significantly related to the outcome. For example, in the dermatologist's office, algorithms can identify a normal mole from a cancerous one, but they lack the capability to inform anyone why the cancerous one appears in the shape it does and with the coloring it has,

[58] Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach."

both of which could help understand underlying immunological processes to advance scientific knowledge.[59] Mullainathan and Speiss explain that when building a machine learning algorithm, there is rarely stability in which right-hand-side variables are used as the model is fit to different partitions of the training set.[60] As a result, there is "greater chance that two functions with very different coefficients can produce similar prediction quality" and therefore there is "uncertainty in attributing effects to one variable over another."[61] The practical translation of this shortcoming is that machine learning algorithms cannot uncover the root of problems, such as what truly causes health issues, crime, or student success.

G. *Algorithms are not useful if their decisions are ignored.*

Even if machine learning algorithms exhibit excellent decision-making capabilities, they are completely useless if their decisions are not utilized or outright defied. Dietvorst et al. reveal that humans exhibit tremendous algorithmic aversion in their reluctance to concede decision-making privileges to machines.[62] Across the five prediction tasks studied by the researchers, subjects preferred a human forecaster over an algorithm, even with knowledge of the algorithm's superior performance.[63] However Cowgill finds that humans are equally capable of deferring to algorithmic decisions, demonstrating that human résumé screeners with knowledge of an algorithm's decision overwhelmingly choose to agree with it.[64] This variation in behavior toward algorithms suggests that one of the crucial steps in designing an effective machine learning tool is

[59] Mukherjee, Siddhartha. "A.I. versus M.D."

[60] Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach."

[61] *Ibid*.

[62] Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology*, 2014.

[63] *Ibid*.

[64] Cowgill, Bo. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening."

to anticipate whether humans are likely to use it and how the decisions will be translated into real world actions.

6. Conclusion

Part I provided an overview of machine learning algorithms in a general context and outlined factors which bolster and impair their usefulness. A common thread across the majority of the factors is that the data is the primary driver in determining whether an algorithm will generate sound predictions without biases. The variation and richness of the right-hand-side variables, the narrowness and timeliness of the outcome variable, and how feedback is integrated into the model are all data concerns at their core. Even determining the relative performance of the algorithms compared to other decision makers is quintessentially a data question of who is able to best identify relationships within data to make predictions. These data-centric conclusions motivate the similarly data-driven solutions discussed in Part III.

PART II: ALGORITHMS IN HIRING

1. Introduction

Choosing whether or not to hire someone for a job includes some kind of data evaluation. Whether the information is a candidate assessment score used company-wide, or an interviewer's informal prior beliefs about similar candidates, the decision-making process is tied to data, regardless of whether it comes from an Excel table or an interviewer's brain. However, as existing records are increasingly digitized,[65] data storage made cheaper,[66] and a plethora of new insights comes out of "big data-enabled"

[65] "The Digitization of Human Resources." Accenture. Accessed April 29, 2018. https://www.accenture.com/us-en/insight-digitization-human-resources.

[66] Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*.

technologies,[67] companies are increasingly favoring the data stored in Excel and elsewhere when making decisions during the employee selection process. Machine learning algorithms have tremendous potential to alleviate problems with traditional hiring processes, which are resource intensive, potentially discriminatory, and often ineffective. However, new issues arise when hiring decisions are handed over to an algorithm, including replication of biases from the training data, false assumptions made within the statistical tool, and ambiguities in how to interpret the results. The potential benefits and equally viable dangers of applying machine learning algorithms to the employee selection process illustrate that these new solutions are far from catch-all.

2. Definitions

A brief search of the web for a company offering "human resource analytics" will produce a wide spectrum of results, ranging from data visualization tools to algorithmic predictive scores. This section will focus specifically on analytics in which machine learning algorithms are used to make predictions about future employment outcomes of potential applicants. To build a machine learning algorithm in this context requires collecting a mass of historical data about a company's workforce, constructing a model based on characteristics of the historical sample to predict a selected outcome, and then applying that model to the same characteristics of new applicants. The algorithm may tell hiring managers who is likely to be a top performer,[68] idea generator,[69] or simply last the longest at the company.[70]

[67] "People Analytics: Recalculating the Route." Deloitte Insights. Accessed April 30, 2018. https://www2.deloitte.com/insights/us/en/focus/human-capital-trends/2017/people-analytics-in-hr.html.

[68] Fleck, Carole. "Using Algorithms to Build a Better Workforce." SHRM. June 1, 2016. Accessed April 30, 2018. https://www.shrm.org/hr-today/news/hr-magazine/0616/pages/using-algorithms-to-build-a-better-workforce.aspx.

[69] Savage, David and Richard Bales, "Video Games in Job Interviews: Using Algorithms to Minimize Discrimination and Unconscious Bias" *ABA Journal of Labor & Employment Law*, vol. 32, 2017. December 19, 2016.

[70] Gautier, Kathryn, and Lalith Munasinghe. "Is Human Capital Analytics a Worthwhile Investment for Your Organization? Linking Data-driven HR Policies to the Bottom-line." *White Paper, TalenTeck LLC*. March 2018. Accessed April 29, 2018.

In hiring, algorithmic bias primarily refers to discrimination. Discrimination is the treatment of a particular group of people in a way that is worse than the treatment of other groups.[71] Discrimination is unlawful in employment if it applies to protected classes, which include gender, race, religion, color, national origin, people over 40, and disability.[72] In hiring, discrimination implies the systematic denial of employment, though there are a number of other ways that employees of a protected class could encounter discrimination in the workplace, such as disparities in compensation, scheduling, or treatment by coworkers, bosses and customers. Under current law, hiring discrimination is generally said to have occurred if the hiring rate of a minority group is no less than four-fifths of the hiring rate of the majority group.[73]

Though bias in decision making is most concerned with discrimination of certain demographic groups, a more expansive definition would be one which encompasses any erroneous assumptions made by a human screener or algorithm which lead to an inaccurate estimation of a specific outcome for an applicant, such as their likelihood of accepting the job, future performance, or expected tenure. The simplest way to illustrate this concept is by assuming that all job applicants are either good performers or bad performers. The decisions made by the algorithm or human screener are biased if the decision-maker mislabels bad performers as good performers and vice versa.

Labor theory explains employment discrimination using two general models: taste-based and statistical discrimination.[74] Taste-based models assume that a disamenity value is accrued when a firm hires a minority worker, and therefore the worker must compensate by being more productive or accepting a lower wage.[75] Some economists argue that taste-based discrimination by customers or employees is

[71] "Discrimination." *Cambridge Dictionary.* Accessed April 29, 2018. https://dictionary.cambridge.org/us/dictionary/english/discrimination.
[72] "Discrimination by Type." Types of Discrimination. Accessed April 30, 2018. https://www.eeoc.gov/laws/types/.
73 King, Allan G. and Marko Mrkonich, ""Big Data" and the Risk of Employment Discrimination." 68. Oklahoma Law Review. 555, 2016.
[74] Arrow, Kenneth J. "The Theory Of Discrimination." *Discrimination in Labor Markets*, 1973, 1-33.
[75] *Ibid*.

empirically more relevant because it better explains the persistence of bias against certain groups.[76] Statistical discrimination occurs because of a signaling problem in the workplace where employers cannot infer how productive, well-matched, or profitable an applicant will be on the job.[77] As a result, the employer relies on priors about distinguishable characteristics of other workers who appear similar to the candidate to make conclusions about the candidate in question.[78] In a way, machine learning hiring algorithms codify this idea by using characteristics of past employees—which may or may not causally relate to employment outcomes—to predict outcomes of new applicants with similar traits.

3. History

Applying machine learning algorithms to hiring is predominantly made possible by the rise of techniques to handle big data. Workforce-related data exemplifies the characteristics of big data, encompassing a disarray of numerical, textual, and even audio/visual information in varying time scales and formats. Human resource data may include answers to an applicant survey, résumés, video recordings of interviews, promotions/ratings, compensation, and engagement/pulse surveys. Increasingly, employers also have access to external information about job candidates provided by data brokers or mined from public internet profiles.[79] LinkedIn connections,[80] credit histories,[81] and web browser history[82] are among the plethora of information that employers can find about prospective employees if they choose to look. Big data

[76] Becker, Gary S. "Investment in Human Capital: A Theoretical Analysis." *Journal of Political Economy* 70, no. 5 (1962): 9-49.
[77] Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *The American Economic Review* 62, no. 4 (1972): 659-61.
[78] *Ibid*.
[79] O'Neil, Cathy. *Weapons of Math Destruction.*
[80] LinkedIn Talent Insights and other Human Resource Analytics Companies
[81] Rivlin, Gary. "The Long Shadow of Bad Credit in a Job Search." *The New York Times*, May 11, 2013. Accessed April 29, 2018.
[82] Peck, Don. "They're Watching You at Work." *The Atlantic*, December 2013.

technology enables the integration of these “messy,” disparate data sources into practical sources of insight.

How extensively companies use data-driven algorithms to make hiring decisions is unclear, though HR analytics leaders suggest that the practice is becoming increasingly prevalent.[83] In the United States, 60–70 percent of companies employ a personality test in their application process,[84] a practice which may not necessarily feed into a predictive algorithm but certainly could. Machine-learning hiring algorithms are more likely to affect low-skill, low-wage workers in general, as companies employing higher-skilled individuals tend to invest more time in in-person interviews and résumé evaluation.[85] Nevertheless, technology firms including Google, Xerox, and IBM have gradually turned to analytics to assess potential job candidates for well-paying positions as well.[86]

Though the potential for bias in analytics-based hiring has only been widely discussed in the last five years or so, the relationship between algorithmic decision-making and discrimination was realized much earlier on. In the 1980s, St. George’s Hospital Medical School in the UK performed a computerized screening of its initial applicant pool.[87] It was discovered that the program denied interviews to a large number of women and minorities because of their names.[88] Since then, academics and journalists have identified bias in psychometric employment testing, résumé scanning technology, and other algorithmic employee selection methods.

4. Issues with algorithms used in the employee selection process

[83] Bersin, Josh. "People Analytics: Here With A Vengeance." Josh Bersin. December 19, 2017. Accessed April 30, 2018. https://joshbersin.com/2017/12/people-analytics-here-with-a-vengeance.
[84] O'Neil, Cathy. *Weapons of Math Destruction.*
[85] *Ibid.*
[86] "5 Companies Using Big Data to Transform Human Resources." Everwise. September 29, 2016. Accessed April 30, 2018. https://www.geteverwise.com/human-resources/5-companies-using-big-data-to-transform-human-resources/.
[87] Burn-Murdoch, John. "The Problem with Algorithms: Magnifying Misbehaviour." The Guardian. August 14, 2013. Accessed April 30, 2018. https://www.theguardian.com/news/datablog/2013/aug/14/problem-with-algorithms-magnifying-misbehaviour.
[88] *Ibid.*

There is immense flexibility when designing an algorithm for use in the employee selection process, in terms of choosing input data, defining the model framework, and determining when in the hiring process to employ the algorithm. As a result, there are a number of factors which may produce biased outcomes or lead to disparate impact for certain groups. This section will examine some of them.

A. Problematic prediction goals

One application of machine learning algorithms in hiring is to identify applicants who will perform well at the company if hired. This exercise can be challenging because finding an actual performance metric or an appropriate proxy variable to serve as the left-hand-side variable is often difficult. A meta-analysis of employee selection methods by Schmidt and Hunter suggests that dollar value of output and individual output as a percentage of mean output are two ideal proxies for performance.[89] In practical settings though, many companies do not keep detailed enough records of their workforce, nor do they understand how to marry financial data with human resources data to quantify individual productivity.[90] In addition, this measure may not apply in settings where the job functions of individual workers are highly variable within the company, work is team-based, or worker outputs are difficult to quantify, as in many service jobs.

In the absence of sound productivity data, machine learning algorithms may be built using rougher proxies, such as employee tenure, performance reviews, or promotions. Using these proxies for productivity has the potential to reinforce existing biases within the company. For instance, if a firm fostered an environment hostile to women for many years causing them to perform poorly and quit early in their tenure, a model built with workforce data from this period would indicate that women are bad

[89] Schmidt, Frank L., and John E. Hunter. "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings." *Psychological Bulletin* 124, no. 2 (1998): 262-74.
[90] Gautier, Kathryn, and Lalith Munasinghe. "Quantifying employee learning curves and their contributions to store revenue at a large U.S. retailer." *White Paper, TalenTeck LLC*. March 2018. Accessed April 29, 2018.

hires (whether gender appears explicitly in the model or not, other variables certainly serve as proxies). Even if the company culture had radically changed to be more inclusive, the historical dataset still holds large weight in the algorithm, and therefore impedes true change within the company.

Another issue that can arise when defining an outcome variable is choosing a performance metric that occurs far into the future. The model only has access to the applicant's characteristics as of the time they applied; additional factors which may influence the worker's productivity or tenure after the score is produced are not taken into account by the algorithm. These factors may include quality of the worker's manager, scheduling, promotions, and satisfaction with team members—all of which do not come into play until after the algorithmic decision has been made. In effect, the longer the time duration between when an applicant is scored and when the outcome variable is realized, the higher the likelihood of uncontrolled confounding variables influencing the outcome.

Far off outcome variables also mean that the time gap between when an algorithm makes a decision and when it learns whether the decision was accurate or not impairs how quickly the algorithm can update its beliefs. Regardless of the performance metric used as an outcome variable, there is an inevitable time lag until the algorithm will be aware of its mistake, because the measures of performance only come into existence after the employee has worked long enough to warrant them. For example, if whether employees reach the six-month tenure mark is used as the outcome variable, the model will not be able to affirm that it correctly identified high scoring applicants until six months after the score was generated. Therefore, using outcome variables that occur far into the future mean that the algorithm does not learn of its type I errors and correct predictions very quickly.

B. Biased historical employment decisions

Historical datasets used to train algorithms can also carry biases arising during the employment selection process. Without sufficient noise to compensate for the discriminatory nature of earlier human decisions, algorithms which learn from the historical data may inherit and perpetuate past biases. For example, companies whose selection processes prioritize cultural fit may build a homogenous workforce in terms of demographics, life experience, and perspectives. Rivera finds that a number of professional service firms in law, banking, and consulting overwhelmingly emphasize cultural matching between interviewer and interviewee over assessing job relevant skills or qualifications.[91] Interviewers look for candidates similar to them in terms of "leisure pursuits, experiences, and self-presentation styles," resulting in the assessors "using their own experiences as models of merit."[92] Workers at the firms in the study primarily come from upper-middle class backgrounds and have attained Ivy League degrees, and the candidates they select do too. One banker acknowledged that "You are basically hiring yourself. This is not an objective process."[93]

Building an algorithm using data from such a scenario may reinforce this selection process and lead to discrimination of applicants who do not look like the existing Ivy League pedigreed workforce. The algorithm will have greater exposure to employment outcomes for individuals similar to those in the existing workforce than those with different backgrounds, such as applicants from state universities. As a result, a model used in the situation outlined by Rivera may be very good at differentiating between two people from Harvard, resulting in low false positive (bad matches get accepted) and false negative rates (good matches get rejected). But when an applicant from U. Mass Amherst applies to the company, the algorithm may not have sufficient historical data to be as predictive. As a result, there might be a high false negative rate

[91] Rivera, Lauren A. "Hiring as Cultural Matching." *American Sociological Review* 77, no. 6 (2012): 999-1022.

[92] *Ibid*.

[93] *Ibid*.

because the algorithm has not seen good matches with similar characteristics in the past, simply because they never had a chance of being hired at all.

Concern over biases in the training data are legitimate due to the fact that persistent discrimination in hiring is found across industries and on multiple dimensions. A natural experiment in symphony orchestras analyzed by Goldin and Rouse quantifies the effect of gender bias at the interview stage of hiring.[94] Over time it became common practice that symphonies placed a screen between the musician and the evaluator so that the musician's identity was hidden.[95] The effect was a substantial increase in the number of female hires, showing that discrimination was a factor in previous selection decisions.[96] Even prior to meeting with candidates in person, there is evidence that employers are biased on paper. Economists Bertrand and Mullainathan designed an experiment to see if résumés with African American sounding names led to fewer callbacks than Caucasian names.[97] The researchers crafted several versions of a résumé and then randomly assigned names to them to hold skills and experience constant. They sent dozens of the fake résumés out to employers in a variety of low to medium skill industries. The results showed that African American résumés received substantially fewer callbacks.[98] Neumark et al. study both the résumé and interview stages and illustrate gender bias in the restaurant industry, where higher-price restaurants discriminate against women at a much higher rate than lower-priced restaurants.[99] Evidence of discrimination in a number of workplace settings shows that training datasets are not likely to be free of biases. When building a machine learning

[94] Goldin, Claudia, and Cecilia Rouse. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." *American Economic Review* 90, no. 4 (September 2000): 715-41.

[95] *Ibid.*

[96] *Ibid.*

[97] Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94, no. 4 (September 2004): 991-1013.

[98] *Ibid.*

[99] Neumark, David, Roy Bank, and Kyle Van Nort. "Sex Discrimination in Restaurant Hiring: An Audit Study." *The Quarterly Journal of Economics* 111, no. 3 (August 1996): 915-41.

algorithm using past employment outcomes, it is pertinent to examine the extent to which biases are present and whether they negatively impact the predictions made by the algorithm.

C. Completeness, comprehensiveness, and specificity of the training data

Across companies, there are significant disparities in the amount of data available for use in the machine learning algorithms. Some firms may have rich, standardized information about applicants going back many years, while others may have been less adept at digitizing records, implementing a universal application form, or storing data into a centralized database. Additionally, data sources external to the company—public data, social networking data, credit data—may be relevant additions to the predictive models. Imbalances in the information acquired about different candidates, and generalizations inferred about individual candidates from group-level data sources have the potential to produce biased outcomes.

Hiring algorithms often incorporate web-based data, such as activity on social websites or browser history, provoking the question of how the algorithms would handle people with a small, or nonexistent, digital footprint. For example, a technology company found that their employees' performance was correlated with recreational web browsing activity, particularly whether or not they had visited a certain Manga site.[100] While the correlation was certainly useful in the algorithms, it represents a way by which certain groups could experience discrimination. For instance, a qualified applicant who does not browse the web after work because she has family responsibilities may not be identified in the algorithm for a reason completely unrelated to the job. Similarly, imbalances in how much web data is available about each person is influenced by socioeconomic status and age. Because of the lack of transparency in how

[100] Peck, Don. "They're Watching You at Work."

machine learning algorithms weigh certain variables, it is unclear how the absence of certain variables will affect certain individuals' treatment by the algorithm.

There are few, if any, penalties for adding more right-hand-side variables into machine learning algorithms, and additional information tends to improve predictions. Therefore, there is an inclination to draw upon as much data as is available, which may incentivize inclusion of variables aggregated at the group level. For instance, zip code level census data—easily accessed, free, and clean—is one example of "cheap" data used to draw conclusions about individuals based on the communities they reside in.[101] Oftentimes and not surprisingly, Barocas and Selbst write this data fails to capture "significant variation within subpopulations."[102] This practice is known as "redlining," a consequence of the once legal practice of drawing red lines around neighborhoods ineligible for bank loans.[103] Group level aggregates are not problematic in themselves, unless the variables come to significantly influence the algorithm's predictions. A reason this may occur is if the training dataset contains insufficient information about the individuals themselves. Furthermore, aggregate level data can serve as proxies for demographic characteristics such as race or economic status, which may be problematic if one is trying to build a model "blind" to such characteristics. However, as Kurtz surmises: "If you wanted to remove everything correlated with race, you couldn't use anything. That's the reality of life in America."[104]

D. Algorithms cannot learn from type II errors

Machine learning algorithms used in hiring can learn from the bad-matched candidates that mistakenly received high scores by the model and then hired. However,

[101] "Community Facts." American FactFinder. October 05, 2010. Accessed April 30, 2018. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.
[102] Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." 104. *California Law Review*. (2016).
[103] King, Allan G. and Marko Mrkonich, ""Big Data" and the Risk of Employment Discrimination."
[104] Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact."

it is impossible for the model to gain any insight about the candidates that they scored poorly, and subsequently rejected from the job. The inability of the algorithm to learn from its type II errors explains why the model is quite sensitive to the degree of bias and variability in the training data. For example, if a company uniformly chose to hire male candidates in the past, the algorithm will continue to choose male candidates because it lacks the ability to realize that some female candidates are actually highly productive workers. If the algorithm could learn from its type II errors, it would have the ability to mitigate biases of the historical data instead of exacerbating them.

E. Algorithmic aversion in the workplace

High confidence in traditional employee selection methods makes the acclimatization toward empirical decisions aids, such as algorithms, very difficult. Even if company leaders wish to incorporate algorithms, hiring managers may simply refuse to use them, as they inherently undermine human judgement. A survey of 201 HR professionals confirmed this belief, overwhelmingly saying that the unstructured interview was more effective than any paper and pencil assessment.[105] The most credible justification for aversion to algorithms is the "broken-leg" phenomenon: supporters argue that extreme outliers cannot be detected by algorithms, and only humans can pick up on hard-to-detect, but serious, deal-breakers.[106] However, on the aggregate, broken-legged individuals represent a miniscule population, compared with the much larger number of candidates incorrectly accepted or rejected because of the lack of useful information and gain of misleading information gleaned through interviews.

[105] Highhouse, Scott. "Stubborn Reliance on Intuition and Subjectivity in Employee Selection." *Industrial and Organizational Psychology* 1, no. 03 (2008): 333-42.

[106] *Ibid*.

5. Useful applications of algorithms in the employee selection process

Machine learning algorithms have demonstrated significant effectiveness in the employee selection process, outperforming the accuracy of human judgements, bolstering job-relevant testing methods, formalizing decision-making processes, and engendering the potential for less discriminatory outcomes. Hoffman et al. found that in a sample of 300,000, low-skill workers hired by algorithms—measuring cognition, personality, cultural fit, and technical skills—had longer tenure than those hired by humans.[107] Even simpler equations may contain more predictive power than humans, as noted in a Harvard meta-analysis of 17 studies of applicant evaluations showing that the equation outperformed people by at least 25%.[108] In terms of improving diversity, a talent management analytics company reported that using a hiring algorithm increased Hispanic hires by 31 percent, and African American hires by 60 percent in the restaurant industry.[109] This section will look at some of the factors which may contribute to these encouraging results.

A. Noisy Historical Employment Decisions

As discussed in Part I, algorithms outperform human decisions when human decision-makers are noisy. A number of employee selection processes, including résumé screening and the unstructured interview, often poorly predict actual employment outcomes.[110] In some scenarios, faulty predictions may arise from consistent, systemic decision-maker biases; however, evidence shows that the errors in human decisions may also exhibit irregular patterns more akin to random noise.[111]

[107] Hoffman, Mitchell, Lisa Kahn, and Danielle Li. "Discretion in Hiring." *The Quarterly Journal of Economics* 133, no. 2 (May 2018): 765-800. doi:10.3386/w21709.

[108] Kuncel, Nathan, Deniz Ones, and David Klieger. "In Hiring, Algorithms Beat Instinct." *Harvard Business Review*, May 2014.

[109] Lam, Bourree. "For More Workplace Diversity, Should Algorithms Make Hiring Decisions?" *The Atlantic*, June 22, 2015.

[110] Schmidt, Frank L., and John E. Hunter. "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings."

[111] Kahneman, Daniel, Andrew Rosenfield, Linnea Gandhi, and Tom Blaser. "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making."

When human decisions are sufficiently noisy, utilizing algorithms can alleviate minor biases and improve prediction quality.

In a software engineering firm, Cowgill demonstrates that replacing human résumé screeners with a machine learning algorithm increases selection of individuals likely to pass an interview round and then accept an offer.[112] The algorithm proves superior at identifying "less traditional candidates"—individuals without elite educations or references—suggesting that human screeners are biased toward picking candidates from a specific subpopulation, but still pick randomly enough that the machine learning algorithm can learn from a sufficient number of less traditional candidates in the training dataset.[113] Illustrating further evidence of the uncertainties surrounding the decisions made by the résumé screeners, the research found that human screeners change their minds if they are aware of the machine's decision.[114] This shows that the even if human screeners hold biases, they are not rigid enough to empower the screener to overrule the judgements made by the algorithms.

It is evident that interviews also result in noisy decision-making. Dana et al. conduct an experiment in which students predict their classmates' GPAs based on past GPAs, test scores, and interview questions.[115] The researchers found that when interviewers were aware that their interviewees were giving random answers, they still asserted that the interview responses aided prediction and preferred having a random interview to no interview at all. This illustrates the powerful effect of "sense-making"—the construction of a confident impression of a candidate, regardless of the inputs.[116] In addition, it supports the conjecture that the conclusions reached by the interviewers are

[112] Cowgill, Bo. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening."

[113] *Ibid.*

[114] *Ibid.*

[115] Dana, Jason, Robyn Dawes, and Nathaniel Peterson. "Belief in the Unstructured Interview: The Persistence of an Illusion." *Judgement and Decision Making*8, no. 5 (2013): 512-20.

[116] *Ibid.*

also quite random, given that they inconsistently weigh the attributes of applicants in order to fabricate some level of coherence about them.

B. *Algorithms can update their beliefs more rapidly than human screeners*

One of the greatest issues in hiring is that the résumé screeners and interviewers rarely see if the people they select fare well on the job, namely, they receive little outcome feedback about their decisions. Consequently, the evaluators tend to rely on beliefs that are updated slowly, and rarely validated. Psychological studies show that when people perform tasks repeatedly without finding out anything about the results of their efforts, they are unlikely to improve performance solely through repetition.[117] Rather, it is through feedback about the efficacy of their decisions that humans better their judgement. This is echoed in the employee selection process where Barr and Bojilov find that more experienced interviewers are not necessarily more competent at screening candidates than their less experienced colleagues.[118] While interviewers gain the ability to identify candidates who will accept offers, they do not get better at selecting candidates who will remain at the company, precisely because the interviewers receive timely feedback about who accepts their offer, but not about who lasts longest on the job.[119] In contrast, the machine learning framework of frequently providing feedback to the algorithm encourages the model to constantly update its beliefs about the world.

C. *Algorithms can identify complex underlying relationships when evaluating candidates*

Increasingly, algorithms are harnessed to quantify a large number of micro data points captured in digitized candidate screening processes. Some companies are

[117] Harvey, N., and Fischer, I. Development of experience-based judgment and decision making: The role of outcome feedback.
[118] Barr, Tavis and Bojilov, Raicho. Working Paper on the Effectiveness of Human Screeners at a Call Center Company. 2018.
[119] *Ibid*.

adopting online video games to screen and select applicants based on scenarios that measure skills needed for the job.[120] Video games enables employers to virtually simulate the demands of the job and measure how candidates react to certain scenarios, all while collecting an extensive set of data points for each applicant.[121] The nuances and richness of this dataset crave analysis using machine learning techniques, which have in turn proved quite successful. One company wishing to identify "idea generators" in its applicant pool found that the video game it used to screen applicants could correctly identify the top 10% of all idea generators.[122] Another area in which machine learning algorithms are gaining traction is in video interviewing. Large companies including Goldman Sachs and Unilever have implemented image and voice recognition technology to track thousands of barely perceptible changes in posture, facial expression, vocal tone, and word choice during video interviews.[123] The data are then translated into predictive scores which compare the data points of new candidates to high-performing existing employees.[124] Though the effectiveness of the scores are somewhat unclear, the company marketing the technology asserts that its customers see lower turnover and better performance amongst the individuals hired by the algorithms.[125]

D. Clearly defining a prediction outcome

Hiring algorithms with a narrowly defined outcome variable have the potential to elicit useful outcomes. The success of the machine learning algorithm in Cowgill's research may be explained by the fact that the algorithm is used for a very limited task:

[120] Savage, David, and Richard Bales. "Video Games in Job Interviews: Using Algorithms to Minimize Discrimination and Unconscious Bias." *ABA Journal of Labor and Employment Law* 32, no. 2 (Winter 2017): 211-28.
[121] *Ibid.*
[122] *Ibid.*
[123] Buranyi, Stephen. "'Dehumanising, Impenetrable, Frustrating': The Grim Reality of Job Hunting in the Age of AI." *The Guardian*, March 4, 2018. Accessed April 29, 2018.
[124] *Ibid.*
[125] *Ibid.*

predicting whether or not candidates are likely to pass the interview and accept a job.[126] The outcome variable is reasonably unambiguous and objective; it is clear whether or not candidates in the historical data passed the interview and accepted the offer. The algorithm's prediction task is also narrow in scope because the interview occurs only a short period after the résumé has been screened by the algorithm. The model does not attempt to forecast far off outcomes, such as expected performance or tenure, which are impacted by factors which do not occur until the candidate is on-the-job. Therefore, it is unlikely that time-related factors unknown to the algorithm will influence whether the candidate passes the interview or not. Perhaps most crucially, the algorithm is not doing away with human decision-making altogether. The résumé screening algorithm fills one step in the employee selection process, leaving humans with the autonomy to make the most crucial decision of all, ultimately choosing who to hire or not.

E. Considerable gains to slight improvements in prediction

Even if machine learning algorithms do not generate perfect predictions, the financial gains can be substantial if they prove marginally superior to humans. For example, a human capital analytics firm demonstrates that the financial impact of a small percentage reduction in employee turnover can have sizable impact on the bottom line.[127] Selecting just slightly better candidates from the applicant pool can help firms reach this goal. Especially in situations where screeners demonstrate predictive capabilities no better than random chance, implementing algorithms simply to dampen the noise in the predictions can still be useful.

At the very least, machine learning algorithms encourage a standardization of hiring protocols, as they promote a consistent decision-making process. For illustration, imagine a company with a large number of stores where in each store, the manager

[126] Cowgill, Bo. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening."
[127] "TalenTeck." TalenTeck. Accessed April 30, 2018. https://www.talenteck.com/.

reviews applications, interviews candidates, and ultimately makes hiring decisions. Under a traditional selection process, outcomes would appear quite different across stores, as managers each have their own way of evaluating candidates and maintain varying perceptions about what constitutes a good employee. The quality of the workforce in each store is thus largely dependent on the effectiveness of the manager's screening decisions, so it is likely that individual store performance will vary accordingly. Now imagine that the company executives decide to implement a hiring algorithm, necessitating a standardized application and structured interview. The hiring process is formalized across store locations to optimize the quality of the workforce in *all* stores. As a result, the company can ensure hiring of the best-match candidates at the aggregate level without relying on the inconsistent, often misled intuition of individual managers within stores.

6. Conclusion

Part II looks at machine learning algorithms used during the employee selection process. It examines the characteristics of the data sources used to build the algorithms, showing that favorable outcomes are achieved via a data-driven process. Utilizing rich, relevant data sources free of systematic biases and defining clear outcomes have the potential to generate useful algorithmic predictions. Considering the demonstrated flaws in human judgement during the employee selection process, hiring is a decision-making realm desperately in need of assistance. Under the right conditions, machine learning algorithms have the potential to be a promising solution.

PART III: SOLUTIONS

1. Introduction

The goal of Part III is to identify practical ways to address data bias and present policy recommendations, particularly within the employee selection process. It will first outline technical solutions and then address some of the legal and ethical aspects surrounding machine learning algorithms.

2. Technical Solutions

Motivated by heightening concerns over "algorithmic bias" from the public sphere, the machine learning community has formulated a number of tools to identify and correct biases in data and mitigate discrimination.[128] These methods are broadly known as "fairness-aware algorithms." The literature focuses on two aspects of fairness: group fairness and individual fairness.[129] The former refers to statistical parity, a basic rule that if a protected group is denied a benefit at a higher rate than an unprotected group, then discrimination is said to have occurred.[130] The latter requires that similar individuals are treated similarly and consistently.[131] This section will examine some of these techniques, providing commentary on their practicality and potential effectiveness in hiring applications.

A. Discovering group discrimination: how to identify biases in the data

Simply quantifying the presence of biases in the training data is an informative first step. Identifying potentially discriminatory characteristics (PDCs)—protected classes such as gender, age, and ethnicity—and then examining how outcomes vary for those groups can provide a rough indication of whether historical biases exist.[132] In

[128] Friedler, Sorelle, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan Hamilton, and Derek Roth. "A Comparative Study of Fairness-enhancing Interventions in Machine Learning." *arXiv preprint*, February 13, 2018.

[129] Ibid.

[130] Grigić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. "On Fairness, Diversity and Randomness in Algorithmic Decision Making." *arXiv preprint*, June 30, 2017.

[131] Ibid.

[132] Friedler, Sorelle, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan Hamilton, and Derek Roth. "A Comparative Study of Fairness-enhancing Interventions in Machine Learning."

hiring, this may involve looking at rejection rates across applicants of protected classes. Other ways of unearthing bias utilize unsupervised data mining techniques. For example, Haijan et al. demonstrate that classification rule mining is useful in discrimination discovery.[133] This process identifies the frequency that members of a certain group suffer a harm, such as how often black women are rejected from a job. It can also unearth channels for indirect discrimination, arising when a PDC occur frequently with a non PDC and, in turn, that non PDC occurs frequently with a harm (in econometrics, this is the definition of a proxy variable).[134] For example, if zip codes are correlated with race and zip codes are correlated with whether candidates are rejected, then using zip codes in a prediction model is a proxy for race, and therefore a source of indirect discrimination.

Though the techniques of discrimination discovery are quite simple, actually applying these methods to data used in the employee selection process is less clear. First of all, discrimination is a difficult concept to prove. If a certain group is hired at a lower rate than other groups, does it actually mean that employers are discriminating against the group or is the group actually less qualified and competent for the job? Furthermore, when outcome variables of interest are employment performance variables, such as tenure or productivity, and certain groups have different outcomes, it does not necessarily imply that disparate impact is caused by discrimination in the workplace. A technical issue of applying these methods to the employee selection process is the high dimensionality of training datasets. It is computationally expensive to search for all significant classification rules linking PDCs to outcomes, in addition to finding relationships between non PDCs, outcomes, and PDCs in order to unearth avenues for indirect discrimination.

[133] Hajian, S. and J. Domingo-Ferrer. "A methodology for direct and indirect discrimination prevention in data mining." *TKDE*, 25(7):1445-1459, 2013.

[134] *Ibid*.

B. Correcting for group biases in the data

If bias is uncovered in the training data, there are a several approaches for correcting them, either by preprocessing the dataset before modelling, incorporating fairness constraints into the model, or post processing the outputs.[135] Preprocessing involves "sanitizing" the training dataset at the outset. Techniques include altering the outcomes, assigning weights to certain individuals, and duplicating or deleting observations completely.[136] All of these methods involve some strategy of reshaping the data based on a notion of what the outcomes "should have been." When modifying the data, Kamiran and Calders show an inherent trade-off between removing discrimination and optimizing predictive power, suggesting that modifying records near the decision boundary is the most effective way to preserve predictive power.[137] In the hiring algorithm context, this might involve modifying the outcomes of applicants at the threshold of hire/don't hire in the training dataset.

Building machine learning algorithms that incorporate fairness constraints is another approach to mitigating bias. A number of techniques have been developed by researchers in the data science community to address different algorithmic frameworks. Kamiran et al. constructs a discrimination aware decision tree by forcing it to simultaneously maximize accuracy and avoid discrimination on the basis of a protected class attribute.[138] This is achieved by modifying the splitting criterion for tree nodes, requiring that a split occurs only if it contributes to accuracy and does not increase discrimination in the resulting tree. Finally, certain leaves of the grown tree are relabeled to mitigate any lingering discrimination. Zafar et al. propose a method for

[135] Hajian, Sara, Francesco Bonci and Carlos Castillo. "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining." Lectured delivered at KDD '16. August 13-17, 2016. San Francisco, CA, USA. Accessed at http://francescobonchi.com/algorithmic_bias_tutorial.html.

[136] Kamiran, F. and T. Calders. "Data preprocessing techniques for classification without discrimination. *KAIS*, 33(1):1-33, 2012.

[137] *Ibid*.

[138] Kamiran, F, T. Calders and M. Pechenizkiy. "Discrimination Aware Decision Tree Learning." *Computer Science Reports* 1013, 2010. Eindhoven: Technische Universiteit Eindhoven.

SVM and logistic regression where fairness constraints are imposed when a decision boundary hyperplane is fitted to a training dataset to provide "an optimal tradeoff between fairness and accuracy."[139]

The tradeoff between fairness and accuracy is a central dilemma for algorithms utilized in the employee selection process. Better predictions usually translate into more profit, so employers will not necessarily be incentivized to correct for biases if it means significant loss of predictive power. However, compliance with discrimination laws and a growing demand for algorithmic accountability suggest that these techniques are still highly applicable to the workplace setting. Furthermore, correcting discriminatory outcomes may also result in selection of a more productive workforce if discrimination barred productive workers from gaining employment in the past.

C. *Auditing the model: Identifying disparate outcomes*

Beyond correcting for bias in the raw data and constraining the model framework, it is also advantageous to audit the algorithm after it is constructed. This involves examining predictions to see whether decision procedures operate differently for protected groups.[140] For a clear illustration, imagine a scenario in which there are two groups, A and B, and two possible outcomes, Good and Bad. The algorithm assigns a risk score to each person representing the likelihood they are Bad. The first concept is whether or not the algorithm is *well-calibrated*.[141] This means that for all individuals with a certain risk score, the proportion that are actually Bad is the same for both groups. The second concept is called *balance for the positive class*.[142] This refers to achieving equal false positive rates across groups. For example, if Good people from group A are more often

[139] Zafar, M.B, I. Valera, M.G. Rodriguez, and K.P. Gummadi. "Fairness Constraints: A Mechanism for Fair Classification". *ArXiv preprint*, July 19, 2015.

[140] Kleinberg, Jon, and Sendhil Mullainathan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *ArXiv Preprint*, November 17, 2016.

[141] *Ibid*.

[142] *Ibid*.

misclassified as Bad than people from group B, then balance for the positive class fails. An analogous notion is *balance for the negative class*, which refers to equality in false negative rates across groups.[143]

Significantly, Kleinberg et al. show that calibration and balance for the positive and negative classes cannot be simultaneously satisfied if the distribution of outcomes differs between groups.[144] For example, suppose that a hiring algorithm is used to predict good/bad performers and that the mean performance level of women is inherently higher than that of men. Therefore, at least one of the following must be true: the algorithm systematically skews the scores upward or downward for at least one gender, the algorithm assigns a higher average score to good performers in one gender, or the algorithm assigns a higher average score to bad performers in one gender. In hiring, the inability to satisfy all fairness requirements is not necessarily problematic in itself. If certain groups are truly better performers than others, then it is not in the firm's best interest to hire equally from both groups. However, if the algorithm is specifically designed to comply with diversity guidelines or anti-discrimination laws, the results from Kleinberg et al. show that fairness cannot be achieved in all dimensions. More importantly through, this deliberate quest for fairness can produce unfairness as an unintended consequence.

D. *Identifying and alleviating individual discrimination*

Individual discrimination concerns the consistency in how observations are handled by the machine learning algorithm.[145] The goal of remedying individual discrimination is to treat similar individuals similarly, disregarding potentially discriminatory attribute variables. K-nearest neighbors is a diagnostic method that

[143] *Ibid.*

[144] *Ibid.*

[145] Luong, B. T, S. Ruggieri and F. Turini. "k-nn as an implementation of situation testing for discrimination discovery and prevention." *KDD*, 2011.

compares the outcomes for similar observations from a protected group to the outcomes of similar observations from any group.[146] In effect, the procedure quantifies whether protected class membership has an impact on an individual's outcome. Zemel et al. formalize an algorithmic procedure for achieving both group and individual fairness.[147] The procedure first maps each observation to a new representation space in which all protected class identifiers are removed before implementing the algorithm. This establishes a "two-step system" consisting of "an impartial party attempting to enforce fairness, and a vendor attempting to classify individuals."[148]

E. The importance of randomness

When a machine learning algorithm replaces human decision-making, it is likely to eliminate their considerable noise and inconsistencies. While this can improve predictive power, it does not automatically translate to favorable outcomes. To illustrate this abstractly, consider the question: what has enabled life to exist over billions of years? Evolution, of course, a process wholly driven by random mistakes during the DNA replication process. While many of these mistakes are complete failures, some have enabled crucial resiliencies to environmental changes and dominance over those without the slightly randomized genetic material. Now think about the machine learning algorithm again, a process similar to that of replicating DNA, but without the life-enabling capability for making mistakes. In effect, machine learning algorithms can only replicate decisions made in the past, generating outcomes which lack the diversity to withstand unpredictable change over time. In hiring, if a machine learning algorithm selects candidates based on the employees who succeeded in the past, then it will build a progressively homogenous workforce optimized on a single outcome variable. But

[146] Ibid.

[147] Zemel, Richard, Yu Wu, Kevin Swersky, and Toniann Pitassi. "Learning Fair Representations." *Proceedings of the 30 th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

[148] *Ibid.*

what if there is a radical disruption in the skillset required for the company to remain viable? Arguably if the workforce has been built under noisier selection processes, it would have maintained a greater degree of heterogeneity and, more crucially, greater resilience.

One strategy for fostering randomness is by increasing the number of algorithms in use. Grgić-Hlača et al. lament the "implicit loss in decision process diversity" which occurs when a large number of inconsistent human decision-makers are replaced by a single machine learning algorithm.[149] The researchers propose that instead of employing one algorithm for a decision-making process, a collection of algorithms could be trained on different subsets of historical data and then randomly assigned for making predictions on new observations.[150] In effect, the algorithms would each represent "a school of thought" based on the selection of data they were trained on.[151] This scheme is largely inspired by the use of machine learning algorithms for making bond decisions in the criminal justice system. Implementing a single algorithm across many courtrooms represents a great loss in the randomness produced through the arbitrary assignment of judges to cases and the noisiness of decisions made by the decision-makers themselves. In large enough companies, the employee selection process represents a similar environment. A traditional hiring process involves a number of hiring managers with heterogeneous beliefs, who are likely to generate noisy decisions about applicants. If a machine learning algorithm is implemented company-wide, a singular set of beliefs for all candidates is instated, removing noise and variance from decision-making. In contrast, building a collection of algorithms to embody several different types of hiring managers and assigning candidates randomly to those algorithms would preserve some of the heterogeneity, thus increasing variation in candidate selection.

[149] Grigić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. "On Fairness, Diversity and Randomness in Algorithmic Decision Making."

[150] *Ibid.*

[151] *Ibid.*

Another opportunity to add randomness into the decisions made by machine learning algorithms is in designing ways to gain insight into type II errors. If the data an algorithm is trained on are suspected of bias, then forcing the algorithm to reverse some of the decisions it makes on new data is a potential remedying method. In other words, if a hiring algorithm selects a subset of candidates to reject, the algorithm should be forced to randomly hire a small number of those candidates anyways in order to gain insight into its type II errors. Candidates from this group represent a valuable learning ground for the algorithm: if the candidates end up being good matches, then the algorithm can appropriately adjust its weights to prevent the same mistakes in the future. On the other hand, if the candidates are bad matches like the algorithm originally predicted, the model achieves greater confidence in its beliefs about the relationships between the characteristics of those applicants and the outcome.

Incorporating randomness into a machine learning algorithm as a way to alleviate bias is largely unexplored territory with limited empirical verification.[152] However, evidence that models perform best when the training dataset they learn from are noisy supports this claim. When data is biased, it is a natural reaction to distrust it. The process of incorporating randomness into the algorithm programs skepticism into it, effectively requiring the model to second-guess its decisions and undermine the notion that the data it learns from represents absolute truth.

F. Importance of a theoretical framework

Perhaps most crucial to developing effective machine learning algorithms is the importance of grounding the models in theoretical frameworks. It is tempting to utilize any data available to improve the predictive power of the model, but through this blind generation of predictions, a greater vulnerability to unintended consequences arises. As

[152] *Ibid.*

King and Mrkonich explain, "The problem with Big Data is that there is no reason the algorithm that best fits the data on Monday will do so on Tuesday. Because there is no understanding of why the correlation exists, there is no basis for surmising how long it will persist."[153] Machine learning prediction is therefore most effective when it is complemented by theory and experimentation, which provide insight into causal relationships. In the employee selection process, algorithms should utilize concepts of labor, psychological, and organizational theory in order to identify the most appropriate predictor variables, understand workforce dynamics, incentives, and more. Pairing the predictive power of machine learning with the explanatory power of theory presents a more robust approach to problem solving.

3. Policy solutions

Machine learning algorithms require more than technical solutions to ensure that biases and discrimination are not fostered by the technology. Outdated laws and regulations must be updated to reflect new concerns over privacy, transparency, and accountability. Furthermore, the limitations of algorithmic solutions need be recognized so that problems can be solved from more fundamental levels.

A. Updating laws to reflect the Age of the Algorithm

If discrimination is found in machine learning algorithms, it is unlikely that employers or the builders of the algorithm will be held accountable under the law. Existing laws in the United States cannot accommodate for the complexities of algorithmic discrimination.[154] Title VII covers employment discrimination, prohibiting unfair treatment based on race, color, religion, sex and national origin.[155] To prove

[153] King, Allan G. and Marko Mrkonich, ""Big Data" and the Risk of Employment Discrimination."
[154] Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact."
[155] *Ibid*.

discrimination, it must be shown that discrimination was intentional (disparate treatment), or neutral policies impact protected classes differently (disparate impact).[156] Furthermore, policies engendering disparate impact must have no business justification and less-discriminatory alternatives must exist.[157]

Disparate impact concerns unintentional discrimination—the more likely scenario an algorithm used in the employee selection process would find itself accused of. Substantiating this type of discrimination is difficult, as the law lacks experience with the nuances of big data algorithms.[158] A plaintiff must first prove that disparate impact occurred, based on the four-fifths rule.[159] Second, the defendant must fail to show that there is a business justification for the disparate impact.[160] The left-hand-side variable may be a good indication of whether or not the algorithm's predictions are relevant to the business. For example, using tenure as an outcome variable may be problematic, because someone's expected duration is not fundamentally necessary to perform the job. The validity of the models also reflects business relevance.[161] A plaintiff should question whether the scores reflect the performance of existing employees and if they accurately predict future performance of scored applicants. If business relevance is established, the third aspect to proving disparate impact requires the court to show that a less-discriminatory alternative exists.[162] If the models are discriminatory, then clearly the less discriminatory alternative is to fix them. However, this is difficult in practice, as there is no well-defined process of determining which features will make models less discriminatory. Additionally, removing dependent variables generally decreases the accuracy of the model, so alternative, equally-well performing models may be impossible to achieve under the requisite that they contain no information that is

[156] *Ibid.*

[157] *Ibid.*

[158] King, Allan G. and Marko Mrkonich, ""Big Data" and the Risk of Employment Discrimination."

[159] *Ibid.*

[160] *Ibid.*

[161] Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact."

[162] *Ibid.*

indicative of a protected class. Improving the breadth and granularity of usable data may be costly, time-consuming, and practically unfeasible, illustrating that less-discriminatory alternatives are not achieved through simple solutions.

Legal scholars suggest that policy should be developed to guide how much disparate impact created by the algorithms is tolerable in employment.[163] Banning the technology completely, however, would be "a perverse outcome, given how much even imperfect data mining can do to help reduce the very high rates of discrimination in employment decisions."[164] The European Union has crafted policies to address the increasing presence of algorithmic decision-making, including a "right to explanation" which enables individuals to question why an automated decision was made about them.[165] Beginning next month, the E.U.'s General Data Protection Regulation bans algorithmic profiling on the basis of "particularly sensitive" personal data, meaning algorithms cannot use protected categories in models.[166] However these policies are not designed with a deep understanding of the framework behind machine learning algorithms in mind. For instance, how can an algorithm satisfy the "right to explanation" if a distinguishing characteristic of the technology is that no one really understands how predictions are formulated? Even after protected categories are removed from the models, how does the law address the effect of proxy variables?

Despite major gaps in the legislation, inadequate policy is better than no policy, as in the case of the United States. In 2014, the Obama Administration released the first ever report to address Big Data concerns about privacy and fairness.[167] The report called for an expansion of technical expertise to "identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes," though no

[163] *Ibid*.

[164] *Ibid*.

[165] Garcia, Megan. "Racist in the Machine: The Disturbing Implications of Algorithmic Bias."

[166] Wing, Jeannette. "Data for Good". Lecture delivered for Columbia ADI.

[167] United States. Executive Office of the President. *Big Data: Seizing Opportunities, Preserving Values*. By John Podesta. 2014.

real policy developments transpired afterward. Marko J. Mrkonich, a panelist participating in a 2016 Equal Employment Opportunity Commission meeting on the issue addressed the gaps in current policy, remarking, "Employers continue to operate in a legal environment based on rules and regulations developed in an analog world with few guideposts that translate seamlessly into the world of Big Data."[168]

B. Increasing Transparency and User Autonomy

The opaqueness of machine learning is such that very few people know what factors enter an algorithm, and even fewer, or no one at all, know how the factors are weighted to produce a prediction. Increasing transparency of input data is one way in which algorithms can be better tested and held accountable for their biases. Several researchers have proposed a society-wide algorithmic auditing scheme, similar to Consumer Reports for buying used cars.[169] The unbiased, third-party would "combine both expert and everyday evaluations, test algorithms on the public's behalf and investigate and report situations where algorithms have gone wrong."[170] Expanding understanding of how machine learning algorithms apply weights to input variables is also a crucial area for future research.

Greater transparency enhances user autonomy over the personal data fed into the algorithms making decisions about them. For example, if job applicants are aware that the selection algorithm will incorporate their web browsing activity, it may alter how they spend time online. Similar to the idea that job applicants can curate their résumés, LinkedIn profiles, and references, revealing the criteria used in employment selection algorithms would lead to greater curation of the algorithm's inputs as well. A

[168] Equal Employment Opportunity Commission. "Use of Big Data Has Implications for Equal Employment Opportunity, Panel Tells EEOC." News release, October 13, 2016. Eeoc.gov. Accessed April 29, 2018.

[169] Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. "Can an Algorithm Be Unethical?" *Paper Presented to the 65th Annual Meeting of the International Communication Association.*, May 2015.

[170] *Ibid.*

caveat is that, like résumés and references, the algorithms would likely become less predictive of actual employment outcomes since applicants would figure out what factors generate a good score and then alter their inputs accordingly. Grgić-Glača et al. suggest that utilizing many algorithms is a possible solution to this problem, as each algorithm would optimize on input variables in unique ways so that there would be no clear way to take advantage of the system.[171]

C. Solving Underlying Problems

When used with the mindset that the technology represents a catch-all solution, machine learning algorithms provide the answers to problems without ever addressing the causes of the problems themselves. However, these root problems are often of greater social, political, and economic importance. For instance, algorithms take the workplace environment as a given, enabling companies to focus all of their energies on identifying better employees instead of trying to fix company problems preventing certain employees from succeeding. Barocas and Selbst argue that employers could focus on "a more family-friendly workplace, greater on-the-job training, or a workplace culture more welcoming to historically underrepresented groups" to boost employee tenure and performance, rather than identifying the "right" candidates from the start.[172] Algorithms authorize complacency in addressing tough problems: policing algorithms identify potential crime hotspots so policymakers can avoid tackling the root causes of crime such as wealth and educational inequalities;[173] school teacher rankings sort teachers based on students standardized test scores, neglecting socioeconomic factors which may play a greater role in student performance;[174] hiring algorithms emphasize

[171] Grigić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. "On Fairness, Diversity and Randomness in Algorithmic Decision Making."
[172] Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact."
[173] O'Neil, Cathy. *Weapons of Math Destruction.*
[174] *Ibid.*

an applicant’s propensity for success rather than a company’s ability to propel a diverse group of employees to flourish. Algorithms provide only temporary relief to more fundamentally broken systems.

FINAL REMARKS

This paper shows that remedying biases in machine learning algorithms is far from a simple task, regardless of the situation in which the models are used. Utilizing algorithms in employee selection—a complex, nuanced, and ever-changing decision-making process—requires careful curation of the data fed into the algorithm and serious contemplation over its design and context. The machine learning algorithm is a delicate mechanism. Under very precise conditions, it has the potential to minimize biases inherent in human intuition, increase accountability, and accurately make predictions. It enables us to reproduce our best decisions consistently. But with the slightest alteration of the data it learns from, the algorithm can come to mimic and propagate our worst judgements and the biases inherent in society. Because of the tremendous influence algorithms command over decisions affecting real humans, it is pertinent that we ensure the algorithm is operating within the former conditions. While we hope for it to achieve blindness, we should never use the algorithm blindly. In order to achieve optimal outcomes, we must fundamentally recognize that when using algorithmic technology, past outcomes in the training data will completely shape future ones. Machine learning algorithms do not create bias; they simply provide a mechanism to mimic our biases. Therefore, developing ways to generate rich, complete, unbiased data should be prioritized over formulating fancier mathematical models. This will enable us to harness the power of machine learning technology with minimal unintended consequences.

Works Cited

"5 Companies Using Big Data to Transform Human Resources." Everwise. September 29, 2016. Accessed April 30, 2018. https://www.geteverwise.com/human-resources/5-companies-using-big-data-to-transform-human-resources/.

Arrow, Kenneth J. "The Theory Of Discrimination." *Discrimination in Labor Markets*, 1973, 1-33. doi:10.1515/9781400867066-003.

Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *California Law Review*104 (2016).

Barr, Tavis and Bojilov, Raicho. Working Paper on the Effectiveness of Human Screeners at a Call Center Company. 2018.

Becker, Gary S. "Investment in Human Capital: A Theoretical Analysis." *Journal of Political Economy* 70, no. 5 (1962): 9-49.

Bersin, Josh. "People Analytics: Here With A Vengeance." Josh Bersin. December 19, 2017. Accessed April 30, 2018. https://joshbersin.com/2017/12/people-analytics-here-with-a-vengeance/.

Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*94, no. 4 (September 2004): 991-1013. doi:10.3386/w9873.

Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16, no. 3 (2001): 199-215.

Brehmer, B. “In one word: Not from experience.” *Acta Psychologica*, 45, 223-241. 1980.

Buranyi, Stephen. "'Dehumanising, Impenetrable, Frustrating': The Grim Reality of Job Hunting in the Age of AI." *The Guardian*, March 4, 2018. Accessed April 29, 2018.

Burn-Murdoch, John. "The Problem with Algorithms: Magnifying Misbehaviour." The Guardian. August 14, 2013. Accessed April 30, 2018. https://www.theguardian.com/news/datablog/2013/aug/14/problem-with-algorithms-magnifying-misbehaviour.

*Cambridge Dictionary.*Accessed April 29, 2018. https://dictionary.cambridge.org/us/dictionary/english/discrimination.

Chaney, Allison, Brandon Stewart, and Barbara Engelhardt. "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility." *Working Paper*, October 30, 2017. Accessed April 29, 2018.

Chouldechova, Alexandra, Emily Putnam-Hornstein, and Diana Benavides-Prado. "A Case Study of Algorithm-assisted Decision Making in Child Maltreatment Hotline Screening Decisions." *Proceedings of Machine Learning Research*81 (2018): 1-15. Accessed April 29, 2018.

"Community Facts." American FactFinder. October 05, 2010. Accessed April 30, 2018. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

Cowgill, Bo. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening." *Working Paper*, March 16, 2018.

Dana, Jason, Robyn Dawes, and Nathaniel Peterson. "Belief in the Unstructured Interview: The Persistence of an Illusion." *Judgement and Decision Making*8, no. 5 (2013): 512-20.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings." *Proceedings on Privacy Enhancing Technologies*2015, no. 1 (April 18, 2015). Accessed April 29, 2018.

Dawes, Robyn M., David Faust, and Paul E. Meehl. "Clinical versus Actuarial Judgment." *Heuristics and Biases*, 1989, 716-29. Accessed April 29, 2018.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology*, 2014. Accessed April 29, 2018.

"Discrimination." *Cambridge Dictionary.* Accessed April 29, 2018.

"Discrimination by Type." Types of Discrimination. Accessed April 30, 2018. https://www.eeoc.gov/laws/types/.

Dressel, Julia, and Hany Farid. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances*4, no. 1 (January 17, 2018). Accessed April 29, 2018. doi:10.1126/sciadv.aao5580.

Esteva, Andre, Brett Kuprel, Roberto Novoa, and Justin Ko. "Dermatologist-level Classification of Skin Cancer with Deep Neural Networks." *Nature*542 (January 25, 2017): 115-18. Accessed April 29, 2018.

Eubanks, Virginia. *Automating Inequality How High-tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martins Press, 2018.

Fleck, Carole. "Using Algorithms to Build a Better Workforce." SHRM. June 1, 2016. Accessed April 30, 2018. https://www.shrm.org/hr-today/news/hr-magazine/0616/pages/using-algorithms-to-build-a-better-workforce.aspx.

Friedler, Sorelle, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan Hamilton, and Derek Roth. "A Comparative Study of Fairness-enhancing Interventions in Machine Learning." *Working Paper*, February 13, 2018.

Garcia, Megan. "Racist in the Machine: The Disturbing Implications of Algorithmic Bias." *World Policy Journal*33, no. 4 (Winter 2016). Accessed April 29, 2018.

Gautier, Kathryn, and Lalith Munasinghe. "Is Human Capital Analytics a Worthwhile Investment for Your Organization? Linking Data-driven HR Policies to the Bottom-line." Talenteck.com. March 2018. Accessed April 29, 2018.

Gautier, Kathryn, and Lalith Munasinghe. "Quantifying employee learning curves and their contributions to store revenue at a large U.S. retailer." *White Paper, TalenTeck LLC*. March 2018. Accessed April 29, 2018.

Goldin, Claudia, and Cecilia Rouse. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." *American Economic Review*90, no. 4 (September 200): 715-41. doi:10.3386/w5903.

Grigić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. "On Fairness, Diversity and Randomness in Algorithmic Decision Making." *Working Paper*, June 30, 2017.

Hajian, Sara, Francesco Bonci and Carlos Castillo. "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining." Lectured delivered at KDD '16. August 13-17, 2016. San Francisco, CA, USA. Accessed at http://francescobonchi.com/algorithmic_bias_tutorial.html.

Hajian, Sara and J. Domingo-Ferrer. "A methodology for direct and indirect discrimination prevention in data mining." *TKDE*, 25(7):1445-1459, 2013.

Hamilton, Jon. "Think You're Multitasking? Think Again." *NPR*, October 2, 2008. Accessed April 29, 2018.

Harvey, N., and Fischer, I. Development of experience-based judgment and decision making: The role of outcome feedback. In T. Betsch & S. Haberstroh (Eds.), *The routines of decision making* (pp. 119-137). 2005.

Highhouse, Scott. "Stubborn Reliance on Intuition and Subjectivity in Employee Selection." *Industrial and Organizational Psychology*1, no. 03 (2008): 333-42. doi:10.1111/j.1754-9434.2008.00058.x.

Hoffman, Mitchell, Lisa Kahn, and Danielle Li. "Discretion in Hiring." *The Quarterly Journal of Economics*133, no. 2 (May 2018): 765-800. doi:10.3386/w21709.

Hurley, Dan. "Can an Algorithm Tell When Kids Are in Danger?" *New York Times Magazine*, January 2, 2018. Accessed April 29, 2018.

Kamiran, F. and T. Calders. "Data preprocessing techniques for classification without discrimination. *KAIS*, 33(1):1-33, 2012.

Kamiran, F, T. Calders and M. Pechenizkiy. "Discrimination Aware Decision Tree Learning." *Computer Science Reports* 1013, 2010. Eindhoven: Technische Universiteit Eindhoven.

Kaneman, Daniel, Andrew Rosenfield, Linnea Gandhi, and Tom Blaser. "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making." *Harvard Business Review*, October 2016.

King, Allan G. and Marko Mrkonich, ""Big Data" and the Risk of Employment Discrimination." 68. Oklahoma Law Review. 555, 2016.

Kleinberg, Jon, and Sendhil Mullainathan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *ArXiv Preprint*, November 17, 2016. Accessed April 29, 2018.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human Decisions and Machine Predictions." *NBER Working Paper*, February 2017. Accessed April 29, 2018.

Kuncel, Nathan, Deniz Ones, and David Klieger. "In Hiring, Algorithms Beat Instinct." *Harvard Business Review*, May 2014.

Lam, Bourree. "For More Workplace Diversity, Should Algorithms Make Hiring Decisions?" *The Atlantic*, June 22, 2015.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." *Pro Publica*, May 23, 2016. Accessed April 29, 2018.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. "The Parable of Google Flu: Traps in Big Data Analysis." *Science*343, no. 6176 (2014): 1203-205. doi:10.1126/science.1248506.

Lewis-Kraus, Gideon. "The Great A.I. Awakening." *New York Times Magazine*, December 14, 2016. Accessed April 29, 2018.

Luong, B. T, S. Ruggieri and F. Turini. “k-nn as an implementation of situation testing for discrimination discovery and prevention.” *KDD*, 2011.

Mann, Gideon, and Cathy O'Neil. "Hiring Algorithms Are Not Neutral." *Harvard Business Review*, December 9, 2016.

Martinez, Joseph P. *Unpunished Criminals: The Social Acceptablity of White Collar Crimes in America*. Master's thesis, Eastern Michigan University, 2014.

Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt, 2013.

Mukherjee, Siddhartha. "A.I. versus M.D." *The New Yorker*, April 3, 2017. Accessed April 29, 2018.

Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*31, no. 2 (Spring 2017): 87-106. Accessed April 29, 2018.

Munro, Paul. "Bias Variance Decomposition." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Web, 100-01. Springer, 2011.

Neumark, David, Roy Bank, and Kyle Van Nort. "Sex Discrimination in Restaurant Hiring: An Audit Study." *The Quarterly Journal of Economics*111, no. 3 (August 1996): 915-41. doi:10.3386/w5024.

Nilsson, Nils J. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge ; New York: Cambridge University Press, 2010.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

O'Neil, Cathy. *Weapons of Math Destruction*. S.l.: PENGUIN BOOKS, 2017.

Pasquale, Frank. *Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press, 2016.

Peck, Don. "They're Watching You at Work." *The Atlantic*, December 2013.

"People Analytics: Recalculating the Route." Deloitte Insights. Accessed April 30, 2018. https://www2.deloitte.com/insights/us/en/focus/human-capital-trends/2017/people-analytics-in-hr.html.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *The American Economic Review* 62, no. 4 (1972): 659-61.

United States. Executive Office of the President,. *Big Data: Seizing Opportunities, Preserving Values*. By John Podesta.

Rivera, Lauren A. "Hiring as Cultural Matching." *American Sociological Review*77, no. 6 (2012): 999-1022. doi:10.1177/0003122412463213.

Rivlin, Gary. "The Long Shadow of Bad Credit in a Job Search." *The New York Times*, May 11, 2013. Accessed April 29, 2018.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. "Can an Algorithm Be Unethical?" *Paper Presented to the 65th Annual Meeting of the International Communication Association.*, May 2015.

Savage, David, and Richard Bales. "Video Games in Job Interviews: Using Algorithms to Minimize Discrimination and Unconscious Bias." *ABA Journal of Labor and Employment Law*32, no. 2 (Winter 2017): 211-28.

Schmidt, Frank L., and John E. Hunter. "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings." *Psychological Bulletin*124, no. 2 (1998): 262-74. doi:10.1037/0033-2909.124.2.262.

"Stop-and-Frisk Data." New York Civil Liberties Union. April 04, 2018. Accessed April 29, 2018. https://www.nyclu.org/en/stop-and-frisk-data.

"TalenTeck." TalenTeck. Accessed April 30, 2018. https://www.talenteck.com/.

"The Digitization of Human Resources." Accenture. Accessed April 30, 2018. https://www.accenture.com/us-en/insight-digitization-human-resources.

"Use of Big Data Has Implications for Equal Employment Opportunity, Panel Tells EEOC." Equal Employment Opportunity Commission. News release, October 13, 2016. Eeoc.gov. Accessed April 29, 2018.

Watzenig, Daniel, and Martin Horn. *Automated Driving: Safer and More Efficient Future Driving*. Cham: Springer International Publishing, 2017.

Wing, Jeannette. "Data for Good." Lecture, Columbia ADI Lecture, Columbia University, New York, April 3, 2018.

Zafar, M.B, I. Valera, M.G. Rodriguez, and K.P. Gummadi. “Fairness Constraints: A Mechanism for Fair Classification”. *ArXiv preprint*, July 19, 2015.

Zemel, Richard, Yu Wu, Kevin Swersky, and Toniann Pitassi. “Learning Fair Representations.” *Proceedings of the 30 th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.