


2019

Big Data and Artificial Intelligence: New Challenges for Workplace Equality

Pauline Kim

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

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

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

Repository Citation

Kim, Pauline, "Big Data and Artificial Intelligence: New Challenges for Workplace Equality" (2019).
Scholarship@WashULaw. 432.
https://openscholarship.wustl.edu/law_scholarship/432

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

BIG DATA AND ARTIFICIAL INTELLIGENCE: NEW CHALLENGES FOR WORKPLACE EQUALITY

*Pauline T. Kim**

Big data and artificial intelligence are increasingly being used by employers in their human resources processes in ways that control access to employment opportunities. This Article, based on remarks delivered as the Carl A. Warns, Jr, Keynote Speaker, describes some of those developments and explains how practices like targeted online recruitment strategies and the use of hiring algorithms to screen applicants raise a significant risk of discriminating against protected groups such as women and racial minorities. It then considers some of the challenges these technologies pose for existing anti-discrimination law and suggests ways that the law should be interpreted to address these new threats to workplace equality.

I. INTRODUCTION

Big data and artificial intelligence seem to be everywhere these days. This data revolution promises that as we gather more and more data, and use computers to analyze that information, we can solve all kinds of problems—from optimizing medical care to assembling a winning baseball team.

The workplace is no different. The vast amounts of data and increased computing power now available are transforming the workplace. We are seeing rapid changes in how work gets done, who performs work, and what skills are needed. These developments raise significant challenges, but I am going to put aside the questions that have received the most attention recently—such as whether Uber and Lyft drivers should be classified as employees or independent contractors, or whether robots will take all of our jobs.¹ Instead, I am going to assume that despite the remarkable changes going on around us, the economy will still require workers in the future, and

* Daniel Noyes Kirby Professor of Law, Washington University School of Law. These remarks were delivered as the Carl A. Warns, Jr. Keynote Speaker on June 29, 2018 at the University of Louisville Brandeis School of Law. Many thanks to Adam Hall for research assistance.

¹ See *Razak v. Uber Techs. Inc.*, No. 16-573, 2018 WL 1744467 (E.D. Penn. Apr. 11, 2018) (3d Cir. argued Jan. 15, 2019) (district judge ruled that Uber drivers were independent contractors); David Z. Morris, *Uber Drivers Are Employees, New York Unemployment Insurance Board Rules*, FORTUNE (July 21, 2018), <http://fortune.com/2018/07/21/uber-drivers-employees-new-york-unemployment/>; see also Alex Williams, *Will Robots Take Our Children's Jobs?*, N.Y. TIMES (Dec. 11, 2017), <https://www.nytimes.com/2017/12/11/style/robots-jobs-children.html>.

that many of those workers will be tied to particular firms in employment relationships.

The question I want to focus on today is this: who will have access to the employment opportunities that remain? Or to put it another way, how will jobs be distributed in a data-driven world? Because not only is technology changing how work is performed, it is also transforming companies' personnel practices. Employers are relying on tools built using big data and artificial intelligence to try to solve the most vexing problems facing human resources (HR) departments—such as how to successfully recruit and retain productive employees.²

To explore these issues, I will begin with a story about a fictional tech company, Tech Co.³ It is experiencing rapid growth and needs to hire many computer programmers quickly. It wants to hire the best talent as efficiently as possible. To accomplish this, it decides to pursue an aggressive social media campaign to target potential applicants. Using tools provided by social media platforms like Facebook, it pushes job advertisements to a narrowly tailored audience—those predicted to be the best candidates. These targeted users see the advertisement on their Facebook news feed. The ad contains a link that takes interested viewers to the website of a third-party vendor hired by the employer to collect and screen applications. On the website, applicants provide basic personal information, upload a résumé, and take an online test or personality inventory. The vendor aggregates this information with other data available from third-party data brokers and enters it into its proprietary algorithm. The algorithm sorts and ranks applicants and the results are used to recommend the best candidates to Tech Co.

Very few of the programmers at Tech Co. are women, and they tend to leave at higher rates than their male counterparts. While the company is concerned by these numbers, it believes that data-driven strategies will be more efficient and accurate than traditional recruitment strategies. Moreover, because gender is not a factor in the advertising or hiring algorithms, it assumes that the process is fair. After pursuing this data-driven strategy for some months, the company notices a pattern. First, the percentage of female applicants is much lower than the percentage of female Facebook users.

² See Forbes Coaches Council, *10 Ways Artificial Intelligence Will Change Recruitment Practices*, FORBES (Aug. 10, 2018, 7:00 AM), <https://www.forbes.com/sites/forbescoachescouncil/2018/08/10/10-ways-artificial-intelligence-will-change-recruitment-practices/#75c9d9de3a2c>.

³ What follows is a fictional story, although it is similar to the experience of a real company—Amazon—that tried to develop a hiring algorithm. See Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, REUTERS (Oct. 9, 2018, 10:12 PM), <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.

Second, the females who do apply are rarely recommended for hire. Over time, more women leave the firm and their total numbers fall steadily. The company says that it would like to hire more women, but it believes the algorithm is neutral, so it must be that women lack the objective criteria that predict success on the job.

This story is intended to illustrate the risk that big data and algorithms may disadvantage groups that are protected by anti-discrimination laws—such as racial and ethnic minorities, women, older workers, and individuals with disabilities. As I will explain, algorithms sometimes operate in ways that are systematically biased against certain groups. When that happens, the effects will look very similar to traditional forms of discrimination.

Before discussing concerns about discrimination, let me make a brief aside about privacy. Many big data and artificial intelligence tools rely on extensive data gathering about applicants and employees, which raises significant concerns about privacy. Those concerns warrant separate attention, so I will only mention here that while there are some legal limits on employers' ability to collect personal information, existing privacy laws are quite limited and are unlikely to slow collection of the types of information used to build workplace algorithms. As a result, employers' reliance on big data and algorithms to make personnel decisions is likely to grow. While I do not wish to minimize the very real concerns about privacy, this Article will focus on the risks of unfair discrimination posed by these technologies.

First, I will discuss some of the ways big data and artificial intelligence may control access to employment opportunities. Next, I will explain the risks that these employer practices may produce discriminatory outcomes and why those risks should concern us. Finally, I will consider the challenges these developments pose for the law.

II. BIG DATA, ARTIFICIAL INTELLIGENCE, AND ACCESS TO EMPLOYMENT

To illustrate how big data and artificial intelligence are being used to control access to employment opportunities, I will focus on the practices followed by the fictional Tech Co.

First, the company pursued a social media recruiting strategy. Social media platforms like Facebook offer tools that allow advertisers to narrowly target a particular audience.⁴ Using those tools, an advertiser can select its

⁴ For an explanation in greater detail of how targeted online advertising works and the risks of discrimination it poses, see Pauline T. Kim & Sharion Scott, *Discrimination in Online Employment Recruiting* (Wash. Univ. in St. Louis Legal Studies Research Paper No. 18-07-02),

audience based on demographic characteristics—such as location, age, and gender.

Of course, as we have learned recently, Facebook knows a great deal more about its users than basic demographic information.⁵ It collects data about pretty much everything we do on Facebook, and a great deal about our offline behavior as well.⁶ It knows what we like, who our friends are, what links we click, where we travel, and where we spend our money.⁷ Using this data, Facebook infers “attributes”—tens of thousands of characteristics, preferences, and interests about each of us.⁸ Attributes can identify groups like people who are interested in vegetarianism or who lived in Bangladesh. Advertisers can target specific audiences by including or excluding users who have a certain attribute or combination of attributes.⁹

Facebook also offers a tool called “lookalike audience.”¹⁰ To use this feature, an advertiser provides Facebook with information about an existing group, the “source audience,” that represents the type of people it wants to reach.¹¹ For retailers, this could be a list of its existing customers; in the case of an employer, it might be its current employees. Facebook takes the source audience, analyzes data about them, identifies other users who have similar profiles, and targets ads to this “lookalike” group.¹² The idea behind targeted advertising is that Facebook’s data helps advertisers predict which users will respond to their ads.¹³ In the recruitment context, this means Facebook’s tools help employers predict which users are most likely to apply for their jobs.

The second strategy of the hypothetical Tech Co. was to rely on an applicant screening tool. Automated screening tools can be useful for employers that receive thousands of applications and seek an efficient way

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3214898##; Till Speicher, Muhammad Ali, Girdhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau & Alan Mislove, *Potential for Discrimination in Online Targeted Advertising*, 81 PROC. MACHINE LEARNING RES. 1 (2018).

⁵ See Dylan Curran, *Are You Ready? Here is All the Data Facebook and Google Have on You*, GUARDIAN (Mar. 30, 2018), <https://www.theguardian.com/commentisfree/2018/mar/28/all-the-data-facebook-google-has-on-you-privacy>; Julia Angwin, Surya Mattu & Terry Parris, Jr., *Facebook Doesn't Tell Users Everything It Really Knows About Them*, PROPUBLICA (Dec. 27, 2016), <https://www.propublica.org/article/facebook-doesnt-tell-users-everything-it-really-knows-about-them>; Speicher et al., *supra* note 4, at 7.

⁶ See Curran, *supra* note 5; Angwin et al., *supra* note 5.

⁷ See Curran, *supra* note 5; Angwin et al., *supra* note 5.

⁸ See Angwin et al., *supra* note 5.

⁹ *Id.*

¹⁰ *About Lookalike Audiences*, FACEBOOK BUS., <https://www.facebook.com/business/help/164749007013531> (last visited May 15, 2018).

¹¹ *Id.*

¹² See *id.*

¹³ See *id.*

of sorting through them.¹⁴ Rather than having a human read every application, automated software—often provided by a third-party vendor—offers to screen applicants by analyzing all the data available about each one and deciding which candidates deserve a second look by the company.¹⁵ This is kind of like Tinder for the HR department, except that the computer swipes left and right instead of a human. In screening or scoring applicants, the algorithm is making predictions about which applicants will perform best on the job.¹⁶

Notice that the concept of prediction is at the core of both of the tools used by Tech Co. in its recruiting process, just as it underlies many artificial intelligence systems. Automated decision-making programs take lots and lots of data and analyze it to find statistical relationships between variables. The relationships that are uncovered are used to build models to predict future cases. An algorithm is simply the set of instructions derived from that analysis. It takes data about past cases and uses it to predict outcomes in new cases—such as who is likely to apply and who will make a good employee.

However, the word “predict” can mean different things. When meteorologists predict the weather, they are forecasting an unknown future event (tomorrow’s weather) based on factors they can observe now, like temperature and barometric pressure. These factors are understood to *causally* influence tomorrow’s weather, and therefore they provide a basis for prediction. The forecast may turn out to be right or wrong, but the better meteorologists understand the causal relationship between today’s atmospheric conditions and tomorrow’s weather, the more accurate their predictions will become.

The word “predict” is used in another sense as well. Data scientists often say that certain data can “predict” other characteristics. What they mean is that because data analysis can identify patterns, once a computer knows certain information, it can infer other characteristics. For example, a recent study found that what a person “likes” on Facebook can “predict” intelligence or personality traits.¹⁷ One particular finding in that study was that “liking”

¹⁴ See Josh Bersin, *Big Data in Human Resources: Talent Analytics (People Analytics) Comes of Age*, FORBES (Feb. 17, 2013, 8:00 PM), <http://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/>; Steve Lohr, *Big Data, Trying to Build Better Workers*, N.Y. TIMES (Apr. 20, 2013), <http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html>; Claire Cain Miller, *Can an Algorithm Hire Better Than a Human?*, N.Y. TIMES: THE UPSHOT (June 25, 2015), <http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html>.

¹⁵ Miller, *supra* note 14.

¹⁶ *Id.*

¹⁷ See Michael Kosinski, David Stillwell & Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROC. NAT’L ACAD. SCI. U.S. 5802, 5805

curly fries on Facebook predicts higher intelligence.¹⁸ I hope what I am going to say next is not controversial, but I do not think this is a causal relationship. I would not advise my students to spend less time studying and more time eating curly fries. Even though there is no causal relationship, liking curly fries on Facebook “predicted” intelligence in the sense that these two things are closely correlated, at least in the data studied by those researchers.

In another well-known example, Target Stores discovered that women often start taking certain types of supplements early in their pregnancies.¹⁹ By observing who purchased those supplements, Target could “predict” that a customer was pregnant.²⁰ Once again, this is not a causal relationship. At least so far as I know, taking supplements does not cause pregnancy. Yet the close correlation between these two facts allowed Target to direct advertising for things like diapers and baby clothes to expectant mothers at the moment they needed to purchase these items.²¹

Recruiting and hiring algorithms use predictions of the second type—observed correlations—to make predictions of the first type—forecasts about how humans will behave in the future. However, because these predictions often are not based on causal factors, they can result in significant errors or biases. I will return to this distinction between correlation and causation again shortly.

III. THE RISKS OF DISCRIMINATION

So, then, what are the risks of discrimination if employers rely on big data tools for recruiting and hiring workers? Consider first targeted online advertising. One risk is that an employer might expressly rely on protected characteristics to select who will be included or excluded from seeing a job posting.²² Using the demographic targeting variables, an employer can select to show an advertisement to only women, or only men, or only to individuals in certain age groups.²³

In a lawsuit recently filed in California, the plaintiffs alleged that companies like T-Mobile and Amazon do exactly that—namely, target their job postings on Facebook in a way that excludes older workers from

(2013).

¹⁸ *Id.*

¹⁹ See Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>.

²⁰ *Id.*

²¹ *Id.*

²² Kim & Scott, *supra* note 4, at 5.

²³ *Id.*

receiving them.²⁴ In the case of T-Mobile, the company placed a job posting on Facebook featuring a young man wearing a cap backwards, and inviting users to “Launch a Customer Care career.”²⁵ Facebook allows users to click on a “Why am I seeing this ad?” feature, and doing so revealed that T-Mobile selected an age range focused on younger workers—eighteen- to thirty-eight-year-olds—to receive the ad.²⁶

An employer’s targeted advertising strategy might discriminate even without relying on an explicit age range. Instead, the employer could choose to target an audience with an attribute that correlates with age—such as “Young Professionals” or “Young Lifestyle.”²⁷ They say age is just a number, but folks over forty are not likely to be included in groups like “Young & Hip” or “Millennials,” which means they will not receive job ads targeted at these groups.²⁸

Race is a bit more complicated. In theory, an employer cannot select a Facebook audience based on race, but some interest categories are closely identified with race—such as the attribute of “African American affinity group.”²⁹ Selecting or screening out an audience using this attribute would probably come pretty close to selecting on the basis of race. Even without using ethnic affinity, employment ads might target or exclude along racial or ethnic lines by using proxies, such as “interested in BlackNews.com” or “interested in Nuestro Diario.”³⁰

Finally, an employer might use the “lookalike audience” feature in a way that selects a target audience along protected class lines.³¹ For example, if it provides Facebook with an existing user group that does not include women or racial minorities, the “lookalike audience” will, not surprisingly, look a lot like the original biased population. Note that in each example above, employers might use these tools to deliberately exclude or target particular groups, but it is also possible for employers to rely on Facebook’s targeting features without being aware of their discriminatory impact.

²⁴ Second Amended Class and Collective Action Complaint & Demand for Jury Trial at 2, *Bradley v. T-Mobile US, Inc.*, No. 5:17-cv-07232 (N.D. Cal. Aug. 20, 2018).

²⁵ *Id.*

²⁶ *Id.* at 36–39.

²⁷ See Speicher et al., *supra* note 4.

²⁸ See *id.*

²⁹ See Julia Angwin, *Facebook Says It Will Stop Allowing Some Advertisers to Exclude Users by Race*, PROPUBLICA (Nov. 11, 2016), <https://www.propublica.org/article/facebook-to-stop-allowing-some-advertisers-to-exclude-users-by-race>; Julia Angwin et al., *Facebook (Still) Letting Advertisers Exclude Users by Race*, PROPUBLICA (Nov. 21, 2017), <https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin>.

³⁰ See Speicher et al., *supra* note 4.

³¹ See *id.* at 11.

What about applicant screening and scoring algorithms? These tools can also produce discriminatory results, even when they appear to be neutral. One reason is that, as with targeted advertising, a selection algorithm might rely on a proxy attribute that causes it to sort implicitly on the basis of a protected characteristic.³² This might occur intentionally, if the person creating or using the algorithm knows that a certain trait is correlated with a protected characteristic and uses it to screen out a disfavored group. For example, place of residence is closely correlated with race in many cities.³³ By relying on an algorithm that sorts candidates based on zip code, an employer could screen out racial minority applicants. The effect could be unintentional as well. For example, certain types of personal data might be correlated with health conditions in a way that causes the algorithm to implicitly discriminate against individuals with disabilities, even if the employer neither knows nor intends to screen on that basis.³⁴

Another way an algorithm can produce biased results is if it is built using biased data.³⁵ An algorithm makes predictions about the future by analyzing data about the past. If that data incorporates biased judgments, then the model's predictions will be biased as well.³⁶ Take, for example, a hiring algorithm currently in use that selects candidates by comparing them with an employer's current employees. If the employer's prior hiring practices excluded certain groups—for example, the hypothetical Tech Co. which hired very few women as computer programmers in the past—the algorithm will simply reproduce the previously existing biases. A different type of algorithm could offer to eliminate applicants who will not be a good “cultural fit” with the employer. Depending upon what factors are used to define cultural fit, the selection tool might operate to exclude racial or ethnic minorities.

There are other, more technical reasons an algorithm might be biased that I will not discuss in detail here. To summarize them briefly: if there are problems with the data used to build a model—for example, if it relies on incomplete, error-ridden, or unrepresentative data—the algorithm may produce biased predictions.³⁷ Models are built by analyzing existing datasets, therefore they are only as good as the input data. The old adage applies here: garbage in, garbage out.

³² Kim & Scott, *supra* note 4, at 22.

³³ Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 889 (2018).

³⁴ *See id.* at 884.

³⁵ *See* Kim, *supra* note 33.

³⁶ *Id.*

³⁷ *Id.* at 861; *see also* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677–93 (2016).

In response, big data enthusiasts argue that data can help reduce discrimination³⁸—and sometimes that may be true. Decades of social science research has revealed many ways that human decision making is biased.³⁹ Aside from outright prejudice, human judgments can be distorted by stereotyped beliefs or implicit biases. Algorithms that ignore arbitrary and irrelevant demographic information can help employers avoid this sort of human bias. So, to be clear, I am not opposed to all uses of data and algorithms in the workplace. Used properly, they can be important tools for employers and can help make workplaces fairer. However, because big data and algorithms also raise the risk of introducing new forms of bias, there are strong reasons to be very cautious about how these new tools are deployed.

One reason for caution is that human behavior is notoriously difficult to predict. Unlike the laws of physics, human behavior is not fully determinate and predictions involving humans can go wrong.⁴⁰ Consider two predictions made on October 30, 2016.⁴¹ It was the middle of the World Series and the Chicago Cubs were down three games to one.⁴² A headline read “The Cubs Have a Smaller Chance of Winning than Trump Does.”⁴³ The article predicted the Cubs had only a 15% chance of winning the World Series and that Donald Trump’s chances of winning the presidency were only slightly higher at 21%.⁴⁴ These were not amateur-hour predictions—they were made on FiveThirtyEight, the website run by Nate Silver, who is one of the biggest data geeks around.⁴⁵ Yet, we all know the outcome of both the 2016 World Series and the presidential election.

So, predictions about human behavior inherently involve varying degrees of uncertainty. Depending upon the purpose of the prediction, that uncertainty may be a minor or major concern. If an algorithm erroneously

³⁸ See Kim, *supra* note 33, at 873 (“Data analytics thus hold the potential to reduce biases and increase opportunities in the workplace for traditionally disadvantaged groups.”).

³⁹ There is a vast literature documenting human bias, a small fraction of which is summarized in Jerry Kang & Kristine Lane, *Seeing Through Colorblindness: Implicit Bias and the Law*, 58 UCLA L. REV. 465, 468–89 (2010); Linda Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1186–87 (1995); Charles R. Lawrence III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317, 322 (1987).

⁴⁰ See sources cited *supra* note 39.

⁴¹ Rob Arthur, *The Cubs Have a Smaller Chance of Winning than Trump Does*, FIVETHIRTYEIGHT (Oct. 30, 2016, 7:53 AM), <https://fivethirtyeight.com/features/the-cubs-have-a-smaller-chance-of-winning-than-trump-does/>.

⁴² *Id.*

⁴³ *Id.*

⁴⁴ *Id.*

⁴⁵ *Id.*; see also Benjamin Mullin, *Nate Silver’s FiveThirtyEight is Moving to ABC News*, WALL ST. J. (Apr. 17, 2018, 3:46 PM), <https://www.wsj.com/articles/nate-silvers-five-thirty-eight-is-moving-to-abc-news-1523986333>.

predicts that I am pregnant and sends me coupons for diapers, I can simply ignore them. If, however, it predicts—erroneously—that I will not be a good employee and I am denied a job as a result, it has created a much more significant problem for me. And if an algorithm not only makes an erroneous prediction about an individual worker, but makes predictions across cases or populations in a way that is *systematically* wrong or biased—that raises much broader social concerns. We should be especially concerned if those biases fall along lines of race, sex, or other protected characteristics in ways that reinforce historical patterns of disadvantage.

Another reason for concern is the difference between correlation and causation, discussed earlier. Recall the predictions that “liking curly fries on Facebook predicts intelligence” and taking supplements predicts pregnancy.⁴⁶ Because the relationships used to make these predictions are not causal, the connections are not likely to be stable over time. Curly fries may no longer be popular among the intelligentsia. Medical advice for expectant mothers may shift away from taking certain supplements. Predictions based on the initial correlations will turn out to be wrong. In the employment context, if non-causal correlations are used to decide who should be hired, some workers will lose out on jobs for reasons that turn out to be completely arbitrary.

Big data enthusiasts are unconcerned by the lack of causal explanations.⁴⁷ They believe we do not need to understand the reasons for the statistical relationships that are observed. If there are errors in their predictions, they argue, the machine will learn and update the algorithm to become more accurate over time.⁴⁸ However, an algorithm can only learn based on the data available to it. In the employment context, machine learning is limited because this required feedback and correction will not always occur.⁴⁹ If an algorithm is labeling candidates qualified and unqualified, the employer will only hire applicants identified as qualified. If some of these “qualified” applicants are, in fact, not so qualified, the employer will find out and can provide updated data to correct the model. However, if some of those labeled “unqualified” are in fact qualified, there is no way to know. This is because they will not be offered the job and will have no opportunity to show they are in fact qualified. These errors will go undetected and the model will not learn from these mistakes.

⁴⁶ Kosinski et al., *supra* note 17, at 5804; Duhigg, *supra* note 19.

⁴⁷ Kim, *supra* note 33, at 875.

⁴⁸ *Id.* at 881–82.

⁴⁹ *Id.*

If those errors are not randomly distributed, but are systematically biased against certain groups, they can themselves produce feedback effects. Workers in disfavored groups may become discouraged and no longer apply to certain firms or for certain types of jobs. Even more harmful from a societal perspective, those discouraged workers may no longer invest in developing their human capital.⁵⁰ So, the risk is that biased data models may create feedback loops that further entrench inequality.⁵¹

Relying on big data and algorithms raises the risk of a form of discrimination I called classification bias in an earlier article.⁵² Classification bias occurs when decision makers rely on classification schemes that worsen inequality or disadvantage along the lines of race, sex, national origin, disability, or other protected status.⁵³ When that happens, it raises the same policy concerns that motivate our laws against discrimination in employment.

IV. CHALLENGES FOR THE LAW

Are existing laws up to the challenges posed by discriminatory algorithms? Let's start with targeted recruitment practices. One of the aims of Title VII of the Civil Rights Act of 1964⁵⁴ was to end job segregation. In the first half of the twentieth century, help wanted ads were typically sex-segregated and often expressed racial or ethnic preferences.⁵⁵ In response, Title VII, in addition to prohibiting discrimination, also made it unlawful for employers to publish advertisements that "indicate a preference, limitation, specification or discrimination" based on a forbidden characteristic.⁵⁶ A similar provision in the Age Discrimination in Employment Act (ADEA) prohibits ads indicating a preference based on age.⁵⁷

Following the passage of Title VII, ads expressing racial preferences quickly disappeared.⁵⁸ It took several more years before newspapers stopped printing separate job listings for men and women.⁵⁹ After some initial

⁵⁰ Samuel R. Bagenstos, *Subordination, Stigma, and "Disability,"* 86 VA. L. REV. 397, 464 n.254 (2000) (citing economics literature that discrimination can be self-perpetuating if it discourages members of groups facing discrimination from investing in their human capital).

⁵¹ Kim, *supra* note 33, at 882.

⁵² *Id.* at 890–92.

⁵³ *Id.* at 890–91.

⁵⁴ Civil Rights Act of 1964 §§ 703–716, 42 U.S.C. § 2000e to 2000e–15 (2012).

⁵⁵ See Nicholas Pedriana & Amanda Abraham, *Now You See Them, Now You Don't: The Legal Field and Newspaper Desegregation of Sex-Segregated Help Wanted Ads 1965–75*, 31 L. & SOC. INQUIRY 905, 906 (2006).

⁵⁶ 42 U.S.C. § 2000e–3(b) (2012).

⁵⁷ 29 U.S.C. § 623(e) (2012).

⁵⁸ Pedriana & Abraham, *supra* note 55, at 906.

⁵⁹ *Id.*

ambivalence, the Equal Employment Opportunity Commission (EEOC) took a clear stance against them in 1968 and by the mid-1970s, sex-segregated help wanted ads had largely disappeared.⁶⁰ The provisions in Title VII and the ADEA prohibiting discriminatory ads have rarely been litigated since. However, given the recent legal challenge to targeted online recruiting,⁶¹ the provisions will likely get renewed attention.

One crucial question is whether or when targeted online advertising violates the statutory prohibition on ads that indicate a discriminatory preference. Does the text of the ad itself have to express a preference to fall within the prohibition? The T-Mobile ad described earlier says nothing about preferring or discriminating against any group.⁶² But the ad was targeted to a specific audience in a way that explicitly mentioned age.⁶³ Does selecting an audience using demographic characteristics indicate an unlawful preference in violation of the statute? No court has yet decided this question, but the cases on sex-segregated newspaper ads suggest that it does.⁶⁴

However, there may be practical difficulties in enforcing this provision. Targeting decisions are often opaque. When someone receives an advertisement and clicks on “[w]hy am I seeing this ad?” the explanation is often incomplete.⁶⁵ As for those in the *excluded* group—they will not see the ad at all and cannot ask “why am I *not* seeing this ad?” Although they do not know of the existence of the ad, they arguably have suffered a tangible harm by being denied information about job opportunities.

An employer’s preference may be obvious when it relies on demographic characteristics, but what if it relies on attributes or behaviors that are facially neutral or makes use of Facebook’s “lookalike” tool? If the effect is to exclude women or certain ethnic groups, has the employer violated Title VII? Did the employer know that using a particular attribute would lead to the exclusion of older workers? Did it understand that using the “lookalike audience” tool might result in an audience skewed along gender or ethnic lines? Because of the differences between targeted online ads and the old help wanted ads printed in newspapers, it is uncertain how the prohibitions on discriminatory ads will apply to today’s social media recruitment strategies.

⁶⁰ *Id.* at 914.

⁶¹ See Second Amended Class and Collective Action Complaint & Demand for Jury Trial, *supra* note 24.

⁶² See *id.*

⁶³ See *id.*

⁶⁴ See Kim & Scott, *supra* note 4.

⁶⁵ See Hanna Kozłowska, “Why Am I Seeing This Ad” Explanations on Facebook are Incomplete and Misleading, a Study Says, QUARTZ (Apr. 6, 2018), <https://qz.com/1245941/why-am-i-seeing-this-ad-explanations-on-facebook-are-incomplete-and-misleading-a-study-says/>.

What about hiring algorithms that sort or score job applicants? Does an algorithm that systematically disadvantages members of a protected group violate the law? The basic structure of the law is familiar. There are two well-recognized theories of discrimination: disparate treatment and disparate impact. Disparate treatment theory forbids adverse decisions taken “because of” race, sex, or any other protected class.⁶⁶ Disparate impact cases involve facially neutral employment practices that have discriminatory effects.⁶⁷ If an employer is using a biased algorithm because it wants to produce discriminatory results, that is clearly a form of intentional discrimination prohibited under disparate treatment theory. Proving the employer’s intent may be difficult, but that type of discrimination fits quite well conceptually with traditional doctrine.

Some commentators have proposed that the risk of discrimination can be avoided by purging a dataset or algorithm of variables recording information about characteristics like race or sex.⁶⁸ As we saw earlier, however, other variables may correlate closely enough that they can act as proxies for protected characteristics.⁶⁹ Eliminating express reliance on variables like race and sex simply cannot guarantee non-discriminatory outcomes.

What about unintended bias that results from the use of workforce analytics? Perhaps an employer purchases software to help screen its applicant pool without realizing that the algorithm is systematically biased against a protected group. In such a situation, disparate impact theory would seem to apply. Under current doctrine, disparate impact cases proceed through several steps.⁷⁰ First, the plaintiff must identify an employer practice that has a disparate impact on a protected group.⁷¹ Then, the employer can defend the practice by showing that it is “job related” and “consistent with business necessity.”⁷² If the employer succeeds in this defense, the plaintiff can still prevail by showing that a less discriminatory alternative exists that the employer failed to adopt.⁷³ While the concept of disparate impact captures concerns about unintended discrimination, the specific ways the doctrine has been applied by courts does not quite fit the challenges posed by biased algorithms.⁷⁴ The doctrine was initially elaborated by courts in the 1970s,

⁶⁶ See *Int’l Bhd. of Teamsters v. United States*, 431 U.S. 324, 335–36 n.15 (1977); see also 42 U.S.C. § 2000e–2(a) (2012).

⁶⁷ See *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971); see also 42 U.S.C. § 2000e–2(k).

⁶⁸ Kim, *supra* note 33, at 918.

⁶⁹ Kim & Scott, *supra* note 4, at 22.

⁷⁰ 42 U.S.C. § 2000e–2(k)(1)(A) (2012).

⁷¹ *Id.* § 2000e–2(k)(1)(A).

⁷² *Id.* § 2000e–2(k)(1)(A)(i).

⁷³ *Id.* § 2000e–2(k)(1)(A)(ii).

⁷⁴ Kim, *supra* note 33, at 866–67, 905–09.

when employers commonly used written ability tests.⁷⁵ Civil rights lawyers successfully challenged the use of these tests because they tended to screen out black workers, but often were not related to skills needed for the job.⁷⁶

Workplace algorithms differ in important ways from traditional selection procedures which focused on testing workers for particular abilities or job skills.⁷⁷ These differences mean that mechanically applying existing disparate impact doctrine will be insufficient to address the risks of discrimination. One challenge is that, as described above, big data algorithms often rely on unexplained correlations with observations about an individual's behavior, rather than measuring skills or abilities that are directly relevant to job performance.⁷⁸ Another challenge is that the factors driving the results may be unknown. An algorithm can be so complex that its decision process is completely opaque—even to the programmers who created it.⁷⁹ As a result, an employer that relies on an algorithm may not be able to clearly articulate or explain the reasons for its decision.

Given these characteristics, biased algorithms raise many questions and challenges for disparate impact doctrine. If an algorithm produces a racially discriminatory effect, can the employer meet its burden of showing that it is “job related” by demonstrating that it rests on a robust statistical correlation? Even if the correlation is unexplained? What about in situations where the relationship between the observed behavior (e.g., liking curly fries on Facebook) and the target characteristic (intelligence) is clearly *not* causal? What if the algorithm is so complex that neither the employer nor the creator of the software can explain what factors caused it to have a discriminatory effect? How should the “job related/business necessity” test be applied to a “black box”? Existing disparate impact doctrine is not equipped to deal with issues like these, and so to address classification bias, the law needs to recognize that predictive algorithms differ from traditional ability tests and to adapt accordingly.

Although this is by no means a complete list, I describe here some of the ways that anti-discrimination law can better address data-driven discrimination.⁸⁰ First, although it may seem counterintuitive, data about sensitive characteristics like race and sex should be preserved. Purging these

⁷⁵ See Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 755–60 (2006).

⁷⁶ See, e.g., *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971).

⁷⁷ See Kim, *supra* note 33, at 908.

⁷⁸ *Id.*

⁷⁹ *Id.* at 881.

⁸⁰ For a more detailed discussion of how anti-discrimination law should be interpreted to respond to the risk of classification bias, see Kim, *supra* note 33, at 916–25.

data will not prevent biased outcomes. And only where this data is available is it possible to audit complex algorithms to identify whether they are having discriminatory effects. Second, the mere existence of a statistical correlation should not be sufficient to justify a model with discriminatory effects. In other words, an unexplained correlation should not satisfy the requirement that an employer show that a practice is “job related.”

Third, when an algorithm systematically disadvantages protected groups, the employer should bear the burden of demonstrating that the model is statistically valid and substantively meaningful, as opposed to merely “job related.” The employer, or the vendor who created the algorithm, should have to show it avoids common sources of statistical bias—for example, by showing that it was built using data that is accurate, unbiased, and representative. In addition, the employer should have to provide some explanation of the decision process and explain its relevance to the job—something more than a mere statistical relationship. Only then can we bring to bear societal values and judgments to determine whether its use is justified despite its effects.

Finally, the law should be structured in a way that encourages employers to audit the algorithms they use and to mitigate any discriminatory effects they uncover by refining, revising, or discarding biased models.⁸¹ And when an algorithm is only one part of a much larger selection process, employers should be given the incentive to ensure that the overall process operates in an unbiased manner, recognizing that it may sometimes be difficult to tease apart precisely which factors in an algorithm are causing a disparate impact. This may mean giving employers some protection if they can show that they are seeking in good faith to identify when bias exists in their automated HR processes in order to eliminate or correct for it. Anti-discrimination law will not achieve its purposes if it gives employers an incentive to bury their heads in the sand and avoid finding out the effects of their practices.

V. CONCLUSION

I want to return to my fictional company, Tech Co., and imagine a different story. The company examines data about its current workforce and recognizes that women programmers are underrepresented in its ranks, even compared with other firms in the same industry. Close analysis of its workforce data reveals that teams that include women are more productive, yet women tend to receive lower ratings from supervisors and peers than their male colleagues involved in the same projects. After looking more closely at

⁸¹ Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189 (2017).

other indicia of productivity, the company learns that, consistent with cognitive science literature, the evaluations of female programmers are affected by implicit biases. Examining data on turnover suggests that certain divisions experience higher attrition of female employees. The company follows up by interviewing employees and finds that women in certain divisions are subject to sexual harassment, while in others they are being excluded from key training or mentoring opportunities. It turns its efforts towards improving the workplace culture and correcting the structural biases uncovered by its analysis of the data, and over time, the number of successful women in the company begins to grow steadily. In this version of the story, data and analytic tools are used to promote, rather than undermine, workplace equality.

The explosion of data and the growth of artificial intelligence is transforming how employers recruit, select, and manage their employees. The question is whether those forces will reinforce past patterns of bias and exclusion or will reverse them and make the workplace fairer. The answer will turn to a large extent on the choices companies make about how to use these technologies. But the law has a role to play as well. The law should be interpreted in ways that create incentives for employers to audit their data-driven HR processes for discriminatory impact, and to ask hard questions of vendors who create these algorithms, rather than uncritically accepting their claims of neutrality and legality. At the same time, the law should encourage employers to use data in creative ways to diagnose and solve problems of implicit and structural bias that hinder opportunities for women and racial minorities in the workplace. Perhaps this will require defining best practices; perhaps it will mean creating some kind of safe harbor for employers who use these tools responsibly and for equality-promoting purposes.

I do not have all the answers, but the big data revolution is not going away. We need to think hard—now—about how to avoid biased uses of data, while harnessing its potential for creating a fairer, more inclusive workplace.