Algorithms, Platforms, and Ethnic Bias: An Integrative Essay

Forthcoming in *Phylon: The Clark Atlanta University Review of Race and Culture*

August 21, 2018

Selena Silva
Research Assistant

and

Martin Kenney*
Distinguished Professor
Community and Regional Development Program
University of California, Davis
Davis
&
Co-Director
Berkeley Roundtable on the International Economy
&
Affiliated Professor
Scuola Superiore Sant'Anna

* Corresponding Author

The authors wish to thank Obie Clayton for his encouragement and John Zysman for incisive and valuable comments on an earlier draft.

Keywords: Digital bias, digital discrimination, algorithms, platform economy, racism

Abstract

Racially biased outcomes have increasingly been recognized as a problem that can infect software algorithms and datasets of all types. Digital platforms, in particular, are organizing ever greater portions of social, political, and economic life. This essay examines and organizes current academic and popular press discussions on how digital tools, despite appearing to be objective and unbiased, may, in fact, only reproduce or, perhaps, even reinforce current racial inequities. However, digital tools may also be powerful instruments of objectivity and standardization. Based on a review of the literature, we have modified and extended a “value chain–like” model introduced by Danks and London (2017), depicting the potential location of ethnic bias in algorithmic decision-making. The model has five phases: input, algorithmic operations, output, users, and feedback. With this model, we identified nine unique types of bias that might occur within these five phases in an algorithmic model: (1) training data bias, (2) algorithmic focus bias, (3) algorithmic processing bias, (4) transfer context bias, (5) misinterpretation bias, (6) automation bias, (7) non-transparency bias, (8) consumer bias, and (9) feedback loop bias. In our discussion, we note some potential benefits from the movement of decisions online, as they are then traceable and amenable to analysis. New social challenges arise as algorithms, and digital platforms that depend on them organize increasingly large portions of social, political, and economic life. Formal regulations, public awareness, and additional academic research are crucial, as algorithms will make or frame decisions, often without awareness by either the creators of the algorithms or those affected by them of biases that might affect those decisions.

Racially biased outcomes have increasingly been recognized as a problem that can infect software algorithms and datasets of all types, and these can be expressed by the decisions that the digital platforms are organizing ever greater portions of social, political, and economic life (Gillespie 2014; O’Neil, 2016). We contribute to the rapidly growing literature on software-based ethnic bias by extending the model created by Danks and London (2017) created to identify where it can be introduced in decision-making processes and which types of bias have received the greatest popular and research attention. This field is particularly important, as digital platforms and software algorithms are creating ever larger assemblages of data to inform decision-making (Gillespie 2014; O’Neil, 2016). Further, these digital technologies have become more sophisticated even as they have become progressively more intertwined in social and economic decision-making, either directly or by providing output that shapes human decisions (Mayer-Schönberger & Cukier, 2014). This is particularly important because software and digital platforms structure social activity (Barley, 2015; Scott & Orlikowsi, 2012). Understanding how and, in particular, where in the software operation bias might be introduced is vital for ensuring that society does not reproduce biases from the social world directly in the decision-making machinery of the future. As many scholars have shown, technologies, by their nature, often embody, consciously or unconsciously, the often unarticulated beliefs and goals of their creators (Winner 1980; Noble 1984).

Software algorithms and platforms are already widely spread across society (Orlikowski, 2016) and are constantly insinuating themselves into ever more social, political, and economic activities and concomitantly reorganizing them. Already, software programs sift through ever-increasing volumes of data to provide credit ratings, decide which advertisements should be delivered to whom, match individuals on dating sites, flag unusual transactions on credit cards, determine who qualifies for a mortgage, predict the locations of future crimes, parse résumés and rank job candidates, generate lists of which candidates for bail or probation are likely to reoffend, and perform a wide variety of other tasks (O'Neil, 2016). Facebook's algorithms recommend articles for our newsfeed, and Google auctions advertising to place next to information that we receive. As a result, even if legally proscribed categories such as race and gender are not used directly, indirect identifiers for such categories are certain to proliferate as algorithms render decisions based on variables that are highly correlated with race, existing biased databases are mined, and individuals express their biases, consciously or unconsciously, in their platform-mediated behavior.

Although concerns about algorithms in decision-making have a long history, their salience in public discourse increased after a 2014 report to the administration of President Barack Obama meant to focus on privacy also addressed the issue of digital discrimination (U.S. Executive Office of the President, 2014). The Obama administration then commissioned a second report titled "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights" that addressed these issues directly (U.S. Executive Office of the President, 2016). With this and an increasing number of newspaper articles and reports from various organizations, a discussion that had largely been confined to a few data scientists and some civil society organizations burst into public consciousness.

Our goal is to identify in the existing literature the ways in which bias can be introduced into online social and economic decisions. For that purpose, in this paper, we focus on ethnic bias. Our review categorizes and summarizes the existing literature on the ways in which algorithmic decision systems can reproduce, accentuate, or even create ethnically biased outcomes. Importantly, the model by Danks and London (2017) that we have extended could be generalizable to other biases, though it is likely that each set of biases manifest themselves differently or concentrate in particular forms. Our goal is motivate others build upon the general conclusions here and apply them to other specific instances.

Algorithms and Digital Platforms

For convenience, let us define the term “algorithm” as it is used here. At its most elementary level, an algorithm is a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer. In computation, algorithms are rendered into software that can then process input data according to a set of rules and produce output. Algorithms are used to make decisions or advise on decision-making in nearly every part of social life.

For us, a digital platform is composed of software-executing algorithms that allow various populations to interact or work together. Thus, Microsoft’s operating system (OS) was a platform that allowed computer makers to build machines that could, through the OS platform, interface with software built by developers. The application software allows users to perform useful functions. The recent growth of online platforms, such as Airbnb, Amazon, Facebook, Google, LinkedIn, Lyft, and Uber, enables two or more sides of a market to interact. In performing these linking functions, platforms can structure human activity. Yet platforms are not neutral; the decisions are made by the platform owners for their own benefit. They are built to frame and drive activities (see, e.g., Gillespie, 2018). As Seaver (2012) and Piskorski (2014) point out, specific theories about the correspondences between persons and things are built into the infrastructure of algorithms and online platforms. Used as filters, these theories shape our perception of the phenomena they describe. For example, the algorithmic choices of a dating site in serving certain matches shape the user’s choices, not only by limiting them but also by validating them as “scientifically” chosen. These issues will become even more difficult to address when outcomes are based on advanced machine-learning techniques. Machine learning is problematic because the outcomes of the algorithmic black box cannot be explained (Knight, 2017; Lepri 2017). If the reason for a decision cannot be explained, then it is possible that a decision is biased, and yet difficult to disprove allegations of bias.

Whereas machine learning and tools for using machine learning are quickly spreading, today most algorithmic decision-making is based not on artificial intelligence (AI) or machine learning but, rather, regression analysis. Regression analysis is statistical prediction that predicts only the odds that something will happen in a specific case. For example, the odds of committing a crime can be predicted, but what is beyond a reasonable doubt—10%, 5%, 1%, or some other percentage? Going further, Barabas et al. (2017, p. 8) suggest that in criminal contexts, such as regression results, such broad-brush statistical results are ill-suited for “effective diagnosis and intervention.” Of course, as in any statistical exercise, results can only be as good as the data entered into the algorithms.

What about Bias?

Because bias is profoundly and deeply integrated into the fabric of society, it consequently can be expected to appear in data used for algorithmic decision-making. For example, in a linguistic study of the correlation among internet word associations, Caliskan et al. (2017, p. 1) find that machine learning can acquire biases from textual data on the internet and concluded that “if we build an intelligent system that learns enough about the properties of language to be able to understand and produce it, in the process, it will also acquire historic cultural associations, some of which can be objectionable.” Until recently, bias research has largely focused on interpersonal dynamics. Yet, in online rating systems, even though obviously individuals are making the biased decision, because aggregated results are reported, the results may be "a function of societal, i.e. macro-level bias." (Greenwood et al., 2017, p. 23). The result is that preexisting bias will become embedded in the software that makes or influences real-world decisions. For example, in a critique of "evidence-based" sentencing, Starr (2014, p. 806) argues against "the use of demographic, socioeconomic, family, and neighborhood variables to determine whether and for how long a defendant is incarcerated," concluding that the practice is neither "progressive" nor particularly scientific and likely unconstitutional. Barocas, Bradley, Honavar, and Provost (2017) attribute this infringing practice to the fact that software developers are not well versed in issues such as civil rights and fairness. However, the issue may be far deeper than many assume, as the bias may not be immediately visible in the big datasets or the ways in which the algorithms structure decisions or platforms force interaction into specific patterns.

Bias issues may be subtle or quite obvious. To illustrate the obvious, Facebook had ethnic "affinity groups" that allowed advertisers to use them for targeting. After complaints from various community groups, Facebook "agreed to not allow ads related to housing, employment, and financial services be targeted using the attribute and renamed it 'multicultural affinity'" (Speicher et al., 2018, p. 1). More subtly, Sweeney (2013) finds that searches on African-American–sounding names were accompanied by a significantly higher percentage of advertisements that had "arrest" in their text. Such explicit use of sensitive attributes is not the sole way that an algorithm can result in discriminatory behavior or output (Barocas et al., 2017). Consider the implications of a data-mining exercise by an insurance firm that secures access to an individual's shopping habits (cigarettes, processed food, sodas, alcohol), ZIP codes, and other physical characteristics, such as body mass, and used this to price life insurance. Although life insurance pricing using such characteristics may be understandable, these consumption patterns might correlate with car accidents or mortgage delinquency. It is here that the issue becomes more complex: what if these characteristics are also more prevalent among certain ethnic groups or socioeconomic strata? These characteristics then also could proxy for categories of people who are legally protected or, depending upon the situation, illegal.

Software is increasingly being developed to recognize images. Image recognition is performed by a computer, which examines visible patterns and generates a statistical prediction based on what it "learned" from its training data. Facial recognition software accuracy differs by race. There is a recent example, in China, Apple's facial recognition software could not distinguish between two Chinese faces. Apple claims it was simply an expected but statistically highly improbable event (Zhao, 2017).   Recently, a press report suggests that better and more training data sets using non-white faces can reduce the error rate (Dormehl 2018). In any system based upon statistical predictions that misidentification is inevitable, and when it appears to be less successful with one race than another, what appear to be discriminatory outcomes can result. Whether lack of recognition is a manifestation of racism, technical issues, lack of interest in a particular market segment (not a sufficient number of training pictures) is uncertain. Other high-visibility instances of failures in facial recognition software have occurred.

Although much of the discussion has centered on exclusion, algorithms and machine learning can also be used to target certain groups. For example, algorithms might identify certain ethnic groups as susceptible to abusive mortgage offers pitched in a certain way. At the same time, algorithms could be used to monitor mortgage offers to identify racial targeting, thereby alerting authorities to potential legal violations. It is this multivalence that makes algorithms so interesting. They can be used not only to discriminate but also to detect discrimination.

Methodology

This paper concerns solely race and ethnic bias, and thus we omit the voluminous literature on gender unless the contribution also addressed race and ethnicity in a substantive way. However, we believe that the value-chain model can be extended to gender and other such issues. Because the literature on algorithms is so broad, we did a comprehensive search on “algorithmic bias” and “algorithmic discrimination.” We also searched sectorally for articles that addressed software and algorithms in fields that included words such as "sentencing” or "education.” Additionally, we searched for “algorithm regulations” and “algorithm legislation” for articles pertaining to actions being taken to regulate the use of algorithms. When articles were identified as candidates for inclusion, their references were examined for additional articles. We did not include articles that discussed general problems with software and algorithms in which discrimination was mentioned in passing. Further, papers that were purely mathematical methodologies discussing how to address “discrimination” with purely technical means generally were not included. We had intended to limit the search to peer-reviewed journal articles, however, we also included papers from conferences that had been published in venues such as arXiv, if they were directly relevant.

Many articles on digital ethnic bias have appeared in the popular press, in particular about specific instances, so we also assembled a database of those articles. Selection of popular press articles for inclusion was difficult, as reports are often reprinted, or derivative articles are published. In cases in which the press article referred to academic research, we obtained the source article. Reports from various organizations were included if they dealt directly with ethnic bias if they were substantive, though that is admittedly subjective. We endeavored to be comprehensive and inclusive yet still focused.

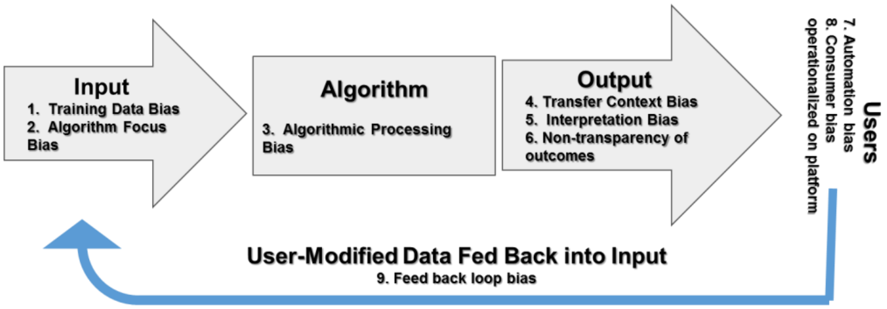
Potential Sources of Bias in Algorithmic Decision-Making Processes

To organize our paper, we built on the model developed by Danks and London (2017), which can be thought of as an algorithmic value chain, as a way to identify where bias can enter an algorithmic production process. This model suggests that the value chain has five phases: input, algorithmic operations, output, users, and feedback. The final phase, interestingly, involves the users themselves, as they can affect the outcome through their own actions and, of course, affect the feedback. Our review of the literature inductively produced nine types of potential bias, as shown in Figure 1: (1) training data bias, (2) algorithmic focus bias, (3) algorithmic processing bias, (4) transfer context bias, (5) misinterpretation bias, (6) automation bias, (7) non-transparency bias, (8) consumer bias, and (9) feedback loop bias. We also included articles with general discussions of bias[1] that can be found in different parts of the algorithmic value chain.[2] These types of bias are not mutually exclusive, as it is possible for a particular algorithmic process or digital platform to have multiple sources of biases, and they may interact in ways that may be difficult to understand. We also identified the activity (i.e., criminal justice, education), technology (facial recognition, search, etc.), or, in certain cases, the firm that was being studied (Uber, Airbnb, etc.). This analysis provides insight into the sectors where algorithms or platforms operate that have attracted the greatest attention. The next section discusses the nine phases and types of bias that we identified.

[1] In slide 30 of her PowerPoint presentation, Mitchell (2017) identifies far more sources of bias, but we have reduced the number to nine.

[2] The only other reference to an "algorithmic value chain" that we found was from an anonymous expert reported in Rainee and Anderson (2017).

Figure 1: Types of Potential Bias and Where They Can Be Introduced in the Algorithmic Value Chain



Source: Expanded from Danks, D., & London, A. J. (2017).

Types of Bias

1. Training Data Bias

The dataset on which an algorithm is trained serves as the foundation for the eventual computations, outputs, and decisions (Eaglin, 2017). Bias introduced into the software by its training data has received the most research attention. Training data bias is one of the most common types of algorithmic bias because all predictive algorithms must be traced. Effectively, the tracing data determines the future operation of the algorithm (Barocas et al., 2017; Danks & London, 2017). Although, in principle, training data bias should be relatively easy to detect, doing so is nearly impossible in reality because data sources are rarely released to the public. Barocas et al. (2017) suggest that this form of bias is more pernicious than using protected categories because it is easier for the developers and those using the algorithm to overlook or completely miss bias that may exist in their source data especially if they did not collect it themselves.

Biased training data is likely widespread. For example, software is being adopted by criminal justice agencies in the hope of preventing crime, rather than waiting for it to happen

(Barrett, 2017; Lum & Issac, 2016), as detailed below. Many of these programs have been developed by private companies, such as PredPol, Palantir, HunchLabs, and IBM (Lum & Issac, 2016). Each company's software is the product of different data sources and measurements, thus leading to discrepancies in measurement validity and different degrees of bias (Christin, Rosenblatt, & Boyd, 2015; Eaglin, 2017; Lum & Issac, 2016). For example, PredPol develops predictive algorithms trained on measurements derived from historical crime data (Barret, 2017). This means that an individual is likely to receive different risk scores, and therefore a different sentence, depending on which algorithm the jurisdiction employs.

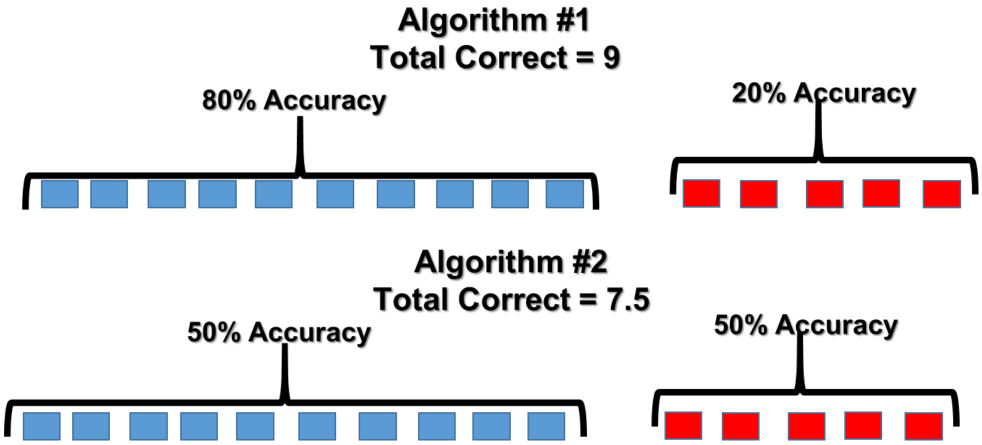
The criminal justice system is an ideal setting for understanding training data bias. It is widely accepted that the justice system already suffers from bias, and an algorithm based on its data will reflect those biases, as crime data are obtained from police records. The result is that the bias is "hardwired" into the results produced by the software. The dataset's prediction software is trained on, are not a measurement of the incidence of crime but, instead, interaction among the community–police relations, policing strategy, and criminality (Lum & Issac, 2016). Therefore, the adoption of predictive policing results in disproportionately high levels of over-policing in low-income and minority communities. If this analysis is accepted, then, paradoxically, predictive policing effectively predicts future policing, not future crime (Lum & Issac, 2016, p. 16).

The criminal justice system's incorporation of prediction software extends to the courtroom as well. Risk assessment software is used as a tool to help judges with setting bail and determining criminal sentences. Researchers at ProPublica conducted an experiment comparing risk assessments derived from Northpoint, a for-profit company, to actual recidivism rates among released prisoners who had been convicted of a crime (Angwin, Larson, Mattu, & Kirchner, 2016). They found that racial disparities existed in the software that resulted in falsely flagging black defendants high risk twice as often as white defendants. Meanwhile, white defendants were falsely labeled low risk more often than black defendants. This could imply that the data used were weighted toward higher expected recidivism rates among black defendants. As in the crime prediction software, risk assessments from software can vary by training data, which varies by company (Christin et al., 2015). Some assessments include criminal history and age, while others include psychological assessments and subjective measurements, such as "personal attitude," often performed by general prison staff with little to none psychological training. members. Also, the number of cases used to create the training dataset is crucial for an accurate reflection of the population, yet some software is trained on as little as several thousand cases. Large variances in training data can result in a dramatically different outcome, depending on the developer. This leads us to question the validity, accuracy, and fairness of the measurements.

Training data bias can still occur with large numbers of cases if the original data points are not obtained through random sampling, thus resulting in either underrepresentation or overrepresentation of groups (Goodman, 2016; Barocas et al., 2017). This is critical, as the algorithms make assumptions about individuals according to statistics derived from sampled groups. For instance, facial recognition software often does not have a diverse set of training data, which results in poor performance and low accuracy for the underrepresented sample, which often comprises minorities (Klare, Burge, Klontz, Bruegge, & Jain, 2012). Database photos are taken from driver's license and ID photos, allowing law enforcement to compare them against the entire city or county population. However, this software has been shown not to be as accurate for individuals with a darker complexion, possibly because of insufficient training. The result could be an increase in false positives and negatives (Garvie, 2016, p. 54). Today, facial recognition software has been used in the criminal justice system to identify criminals by creating "virtual line-ups" developed by algorithms, instead of witnesses.

When building algorithms, firms make decisions regarding where to concentrate their investment. Algorithm producers might choose to optimize their algorithm for concrete reasons. For example, a firm might optimize a facial recognition algorithm to recognize a social group more accurately. As a thought experiment, compare Algorithm 1 in Figure 2, which is more accurate at prediction for the much larger Group 1 than for Group 2, against Algorithm 2, which is equally effective for both populations. From a technical and likely business perspective, Algorithm 1 is superior. Now consider a case in which Group 1 members individually have a significantly higher business value than those in Group 2. In this case, it would be even more valuable for the firm to continue to improve the algorithm for identifying Group 1 and Group 2 would languish. If Group 2 is a "protected" group, then the firm could be labeled biased, though, in fact, it would be a business decision regarding the direction to orient R&D.

Figure 2: Graphical Representation of the Conundrum Posed by the Difference between Two Algorithms in Terms of Overall Identification Accuracy for Groups 1 and 2



However, the implications of such decisions can be a powerful generator of inequality. For example, consider whether facial recognition software is less effective for certain social groups. In policing, if the recognition software is less effective for a certain group, that group's members would be more likely to be stopped unnecessarily and not stopped when they should be. It has been shown that facial recognition software is less effective with African Americans, but whether this is technical problem or due to insufficient training data is not clear (Pulliam-Moore, 2015). Lower accuracy of facial recognition of minorities by software is seen in other countries as well. East Asian algorithms were found to perform better on East Asian faces and less well on Caucasian faces (Phillips, Jiang, Narvekar, Ayyad, & O'Toole, 2011). The training data samples are nearly always larger for the majority population, for which they have greater accuracy.

Skewed or biased training data can affect the hiring process and employment opportunities for minorities as well. Algorithms can be used to rank candidates' applications to determine the best choice for the job. However, the way in which the data are acquired could skew the results toward specific groups of people. To illustrate, consider a case in which the algorithm uses the existing database of employees for training, thereby the algorithm developer will select individuals who are similar to those already employed (Ajunwa, Freidler, Scheidegger, & Venkatasubramanian, 2016). As a result, applications of individuals from groups who are not already part of the firm's employment profile will be excluded.

The nature of training and the data on which training take place can introduce bias. Algorithms can be trained on neutral data, such as crime statistics, but they are often imbued with historical prejudice because of differential enforcement or for a variety of other reasons. For example, historical stereotypes and biases have been found to be embedded in word choice. When an algorithm is trained on written language, it will develop the same human biases unintentionally

(Caliskan, Bryson, & Narayanan, 2017). Credit scores are also susceptible to historically biased training data. If a firm uses historical loan payment based on demographic characteristics, such as location or race, fewer samples will consist of disadvantaged groups that were denied access to credit. As a result, those already denied access to credit will continue to be denied access, thus resulting in continued marginalization (Goodman, 2016). The problem, of course, is that nearly all training is done with proprietary data, so, in the absence of litigation compelling discovery and a reanalysis of the data, there is no way to know whether the training data were biased.

2. Algorithmic Focus Bias

Algorithmic focus bias occurs within the dataset itself. As a society, we have established certain categories, such as race and gender, that cannot be incorporated into algorithms. Biases can occur from both the inclusion and exclusion of available information. This means developers must carefully consider the effects of the variable(s) in question and whether the benefit outweighs the potential harm from using sensitive information (Danks & London, 2017). For instance, exclusion of gender or race in a health diagnostic algorithm can lead to an inaccurate and harmful report. However, the inclusion of gender or race in a sentencing algorithm can lead to discrimination against protected groups. Yet Barocas et al. (2017) point out that, in certain cases, these variables must intentionally be used to weigh groups differently in order to reach a “fair” outcome.

Targeted online advertising is prone to algorithmic focus bias, which has recent received great attention. As mentioned earlier, Angwin and Parris (2016) discovered that Facebook allowed advertisers to filter groups classified by “ethnic affinity.” They were able to purchase advertising targeted at house hunters and exclude individuals identified as having “African-American,” “Asian-American,” and “Hispanic” affinity. Facebook maintained that its “ethnic affinity” identifier is not the same as race but, instead, is assigned based on the user's "likes," though the category is listed in the "demographics" section. Facebook's response to the backlash was to rename "ethnic affinity" as "multicultural affinity" and place the modifier under the "behaviors" classification instead of "demographics" (Angwin & Tobin, 2017). Although Facebook advertising does not include race or ethnicity as a choice, "ethnic affinity" or "multicultural affinity" appears to be a close proxy. More recently, Facebook introduced a new advertising feature, "custom audiences," allowing them to select "personally identifying information" (PII) (Speicher et al., 2018). Advertisers cannot identify any targeted profiles, however, the identifiers used are very personal "including phone numbers, email addresses, and combinations of name with other attributes (such as date of birth or ZIP code)" (Speicher et al., 2018, p. 4). They can use this system to target certain groups by making a PII list of desired attributes. Advertisers can gather sensitive data from various public records and then use that data to target users with specific characteristics. Meanwhile, those who lack these characteristics would be unaware that they were omitted.

Algorithmic focus bias potentially influences the criminal justice system as well. Angwin et al. (2016) obtained generated risk scores for 7,000 people arrested in Broward County, Florida and checked who committed crimes after two years. Through their analysis, they discovered that race may have been used as an attribute in the risk assessment's crime prediction calculations: or, more likely, a highly correlated proxy variable such as income or address.

> [They] ran a statistical test that isolated the effect of race from criminal history and recidivism, as well as from defendant's age and gender. Black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind. (Angwin et al., 2016)

However, this evidence is not proof of algorithmic focus bias and may actually be training data bias (see Bias 1) or both types interacting to produce an arguably discriminatory result. Berry-Jester, Casselman, and Goldstein (2015) explain that, even if these measures are accurate at predicting rates of recidivism, it is unfair to measure variables or characteristics that are outside the defendants' control (race, gender, age, etc.). Although it is not difficult for developers to consciously avoid sensitive variables, proxy variables that are highly correlated with protected categories may result in discriminatory measures. To illustrate, high school attendance, street address, or even first names can be a proxy for race and provide results that are functionally equivalent to those that use race as a variable (Ajunwa et al., 2016).

3. Algorithmic Processing Bias

Bias can also be embedded in the algorithms themselves, hence the term "processing" bias. One common source of such bias is created when variables are weighted (Danks & London, 2017). Another arises when the algorithms do not account for differences in cases, resulting in unfair or inaccurate outputs. For example, grading algorithms are being introduced in classroom instruction. Guskey and Jung (2016) found that when an algorithm was given student grades over the course of a lesson, it scored students based on the average of their assignments. However, when teachers were given the same data, they adjusted the students' score according to their progress and understanding of the material by the end of the lesson. The grades by the teachers appear to be fairer than that of the algorithm. Of course, the algorithm could be reprogrammed and to take student progress into account. This algorithm was unable to process the information "fairly" despite being given a neutral dataset.

Other cases are more complex. Consider a widely reported case in which one company's chief scientist found that "one solid predictor of strong coding is an affinity for a particular

Japanese manga site" (Peck, 2013, para. 50). The discovery of this data point could then be given a particularly high weighting in the algorithm and used for selection. This might introduce unknown bias against or for certain groups. Effectively, the weighting in the algorithm tends to exclude non-readers of that site from consideration, thereby introducing bias (King & Mrkonich, 2016, p. 576). Bias can also be introduced in a subtler way. Design choices may end up promoting certain values or advantaging certain social subgroups. As Surden (2017) observes, these preferences or biases may be a byproduct of design decisions made for technical, efficiency, or functionality reasons.

Algorithmic processing bias is one of the most difficult forms of bias to reveal, as developers have incentives not to disclose their source code to the public (Citron & Pasquale, 2014). In some instances, consumers have discovered this type of bias after repeated usage, such as by users on Booking.com, whose algorithm that did not allow users to rate any hotel below 2 out of 10. Users described unsatisfactory experiences in comments, but the algorithm's design inflated hotel scores, resulting in significantly higher scores on the hotel reservation platform (Eslami, Vaccaro, Karahalios, & Hamilton, 2017).

4. Transfer Context Bias

Bias can occur after the algorithm provides the user with an output. One type of bias occurs when the output is put into an inappropriate or unintended context, leading to discriminatory or biased decisions. This is called "transfer context bias" (Dank & London, 2017). One common example is the use of credit scores as a variable in employment. It has been estimated that roughly one-third of employers request credit checks on job candidates (Gallagher, 2005). Effectively, bad credit is being equated with bad job performance. Yet little evidence indicates that credit is related to work performance. Because having bad credit is not a protected category, the employer is free to use it as a screening device. This can relate to bias because certain groups do not have the same access to credit, so this variable would exclude them from employment. With the advent of big data and machine learning, using it may increase the potential for this sort of transfer bias.

5. Interpretation Bias

Interpretation bias is another way that users can cause a “neutral” output to reflect bias. The bias arises when users interpret an ambiguous output according to their own internalized biases. The problem of interpretation bias occurs because algorithmic outcomes are stochastic—each prediction is in fact only the risk that something will occur. For example, a recidivism prediction may itself not be biased, and it is ultimately up to the judge to interpret that score and decide on the punishment or bail amount for the defendant (Eaglin, 2016). The difficulty is that individual judges may be susceptible to bias. One judge may interpret a risk score of 6 as high, while another judge may see this as low.

Problems arise in a variety of areas. For example, facial recognition software is used to compare evidence obtained at the scene of a crime, such as video surveillance, to a database of photo identification to determine possible suspects (Garvie, 2016). The software cannot unequivocally determine an exact match. Instead, it indicates the likeliness of a match on a scale of, say, 1 to 100. An officer must determine whether the threshold for a sufficient match has been reached and then decide whether the evidence is sufficient for an arrest (Garvie, 2016). Interpretation bias arises because the user interprets the data provided by the computer to, for example, decide whether the evidence is sufficiently compelling either to act or not to act. Effectively, the interpretation can be affected by the interpreter’s biases.

6. Non-Transparency of Outcomes Bias

Because of the increasing use of machine learning, enormous databases containing large numbers of variables, and algorithms that are constantly being breached, a situation has been creating in which the reasons for the outcomes are becoming increasing opaque (Knight, 2017). Interestingly, the reasons for the outcome may even be inexplicable to the creator of the algorithm or the software owner. For example, when a loan application is denied, even the bank may not know the exact reasons for the denial.

Such outcomes are particularly problematic when a few powerful actors, such as corporations and governments, possess enormous datasets and sophisticated tools that make important decisions that are not transparent (Lepri et al., 2017, p. 9). The absence of transparency makes it difficult for the subjects of these decisions to identify discriminatory outcomes or even the reasons for the outcome. Within the realm of predictive policing, transparency is an issue for both those using the algorithms and those who are being judged based on their outputs (Barrett 2017). This is exacerbated, as many of the firms providing systems attempt to suppress even the knowledge that their software is used in decision-making. Thus, secrecy regarding the software's usage can further contribute to the problem of non-transparency.

Firms such as Google, Yahoo, and Microsoft have created tools to allow the user to control and select which type of advertisements will be displayed on Google services and partner's websites (Datta, Tschantz, & Datta, 2015). However, Datta et al. (2018) find that the user's selection of ad settings preferences had only a small effect on the ad outcomes. Not surprisingly, an individual's browsing history and personal demographic characteristics were much larger determinants of the ads reviewed. To illustrate: when users browsed websites related to disabilities, they received ads depicting mobility lifters and standing wheelchairs. A control group that did not browse disability-related websites did not receive any of these advertisements. Interestingly, this attribute was not displayed in the agent’s ad settings. They conclude from this study that the “ad settings” function did not provide full information regarding what Google uses to make decisions.

AI and machine learning further complicate the provision of transparency. Some of these systems use “neural networks,” modeled after the human brain based upon millions of computation clusters. The software constantly updates and changes its nodes, in response to new data, resulting in extremely complex code and computations. Researchers and developers of machine learning systems are often unable to explain how the machine derives its output and which variables are being used (Gershgorn, 2017; Kuang, 2017). As a result, it is virtually impossible to provide clear explanations for the outcome of any decision that relies on machine learning.

7. Automation Bias

Automation bias results when the user views the algorithms’ output as objective or factual. The user believes that the machine provides “neutral” statistical computation. It results from the belief that the output, which is stochastic, is objectively true, rather than being a prediction with a confidence level (Goodman, 2016). For instance, Zarsky (2016) finds that automation bias can affect credit decisions because the scores are fully automated, relying on group statistics and personal credit history to produce a result. Such a process effectively identifies some people as having lower credit scores and then limits their access to credit. If they lack access to credit, their scores cannot improve. Effectively, the algorithm traps them.

Automation bias also can be seen in the criminal justice system. In these cases, the algorithm generates risk assessments to inform decisions regarding sentencing and bail (Jackson, Banks, Woods, & Dawson, 2017). Judges and others may give more credence to computer-based recommendations, in contrast to comparable human-based assessments, because of the belief that a computer-generated, analytics-based decision may be more objective. Importantly, Dressel and Farld (2018) showed that commonly used recidivism prediction software was no more accurate than the average citizen provided with minimal information on the defendant. This human tendency to unduly ascribe value neutrality to technology and to defer to mathematical analysis can lead to erroneous decisions (Surden, 2017).

Algorithms lack intuition and empathy, and thus using them may result in discriminatory outcomes. There is a counterargument, assuming that the training data and algorithms were not biased: that using outputs as objective values can eliminate discrimination that could arise with human intervention. The danger in automation bias is that algorithm users are unaware of the assumptions and hidden bias in the algorithms' operation and uncritically accept the decisions generated.

8. Consumer Bias

Consumer bias is bias that people can express on digital platforms. Effectively, it transfers their biases from the analog world to the online environment. In this respect, consumer bias is not significantly different from that in the offline world, but digital platforms can exacerbate or give expression to latent bias. Moreover, discrimination forbidden in the physical world may be expressed in platform-mediated context. Bias has been shown to occur in online purchasing behavior on various online retail sites. For example, in an experiment with online rating systems, Greenwood et al. (2017) that manipulated service provider histories and photos they showed that when good service was provided the ratings showed no gender or racial bias. However, when service was altered to be perceived as bad, there were more severe rating penalties for women and minorities – effectively there was more severe punishment. Because of the importance of rating systems, users have significant power because their evaluations are fed back into the platform as data.

One perverse example of the power of users was the 2016 introduction by Microsoft of its Tay chatbot on Twitter. Within 24 hours, Tay had “learned” from Twitter users to tweet out racist, sexist, and offensive responses and had to be taken offline (Vincent, 2016). On many other platforms, such as Facebook, YouTube, and Google Search, results and all kinds of other sites must be monitored and scrubbed of racist (and other offensive content). In some cases, such material is merely offensive and not illegal or close to the boundaries of legality.

Digital platforms, such as Uber/Lyft and Airbnb, compete with taxis and hotels, respectively, which are regulated by the government and have a statutory responsibility to serve all customers. If a platform is properly designed, using it should decrease bias.[3] Edelman and Luca (2014) find that Airbnb and similar services, sometimes known as peer to peer (P2P) e-commerce, actually facilitate discrimination (see also Kakar, Franco, Voelz, & Wu, 2016). This is because P2P e-commerce demands personalization, which results in extensive profiles that include identifying and personal information, including name, gender, a profile picture, and possibly sexual orientation, all of which provide information that can trigger bias (Kakar et al., 2016). Effectively, access to this information provides individuals with an opportunity to pick and choose with whom to engage in a transaction or what type of evaluation to provide, thus allowing

[3] Of course, ample evidence shows that taxi drivers, in particular, discriminate against certain classes of customers directly by, for example, not picking up particular customers or not serving certain neighborhoods (Ambinder, 1995), even though it is illegal for taxi drivers to discriminate (Brown, 2018). Some evidence indicates that Uber and Lyft drivers may be less likely to discriminate based on race than taxi drivers.

unrestrained biased responses. In fact, in digital environments, racist or otherwise unsavory comments may be even more common.

On multisided platforms, the biases can be expressed on either side. For example, on platforms such as Uber/Lyft and Airbnb, the discrimination can be from either service providers or customers. For instance, controlling for possible confounding variables for Airbnb listings' cost of rental, Edelman and Luca (2014) find that black hosts earn roughly 12% less "for a similar apartment with similar ratings and photos relative to [non-black] hosts" (Edelman & Luca, 2014, p. 10). In another study, Kakar et al. (2016) find that Hispanic hosts' listings have prices that are 9.6% lower than those of equivalent non-Hispanic hosts, while Asian hosts' listings had prices that were 9.3% lower. They suggest two possible explanations for this difference: taste-based discrimination and statistical discrimination. Taste-based discrimination occurs when the customer favors or disfavors a renter because of user preference, in this case, with respect to race. Statistical discrimination arises when customers infer a difference between the hosts' properties and make a decision to minimize their risk (Edelman & Luca, 2014; Kakar et al., 2016). In this case, the customer infers qualities of the property using the race of the host as a guide. Of course, both are cases of discriminatory behavior by the customer, but the reasons may vary by individual.

9. Feedback Loop Bias

As Zuboff (1988) pointed out, one of the most powerful features of computation-based systems is that all activities on them create more data. Consider the Google Search algorithm, which responds to a query, which it records, so the query becomes input for succeeding searches. This dynamic results in better future search outcomes. The conundrum is that the algorithm is learning from user behavior. The training data and the platform's algorithm were tested and without bias, but the algorithm learns from user behavior.

What if consumers or providers systematically rate a certain class of individuals differently (Bias 8)? These ratings are input for further analysis, which leads a class of individuals to be considered suspect. Yet another form of feedback loop bias occurs in predictive policing, when software relies exclusively on historical crime data (Barrett, 2017). Crudely described, on average, when police officers go into a neighborhood to find crime, they find crime. This then raises the crime rate of the community, leading to the dispatch of more police, who make more arrests, initiating a self-reinforcing feedback loop. Of course, fixes for this exist. For example, Barrett (2017) suggests the use of a variety of data sources to avoid these dynamics.

Benefits of Platforms and Algorithms

The academic and popular press articles referenced in this paper indicate that algorithms are reproducing social biases that already exist and possibly reinforcing them. For example, Caliskan et al. (2018) find that these biases are inherent in the materials on the internet, suggesting that the problems may be either far greater than imagined or even close to unsolvable.[4] To address bias of all types, pornography, and other forms of objectionable content, firms using algorithms for decision-making or serving data to the public through the internet will have to increase both automated and human content moderation, and, as the popular press has reported, they are doing so (Glaser, 2018; Ho & Salna, 2017; Solon, 2017). As Chander (2017) observes, given the potential liability, it is unlikely that any legitimate software provider or platform owner would deliberately write discriminatory software or specifically include socially objectionable material.

[4] If this is inherent in the language or the entire corpus of material on websites, then the composition of the technology workforce may not be the source of bias.

What is less often researched is whether algorithmic decision-making and platforms may ameliorate social biases, which, of course, were the claims of the utopian advocates of digitization. With regard to the issue of race, our literature review discovered little academic research that found that amelioration of social bias. In part, this was because nearly all the research has focused on finding bias and generally confirms that the algorithms merely reproduced existing results.

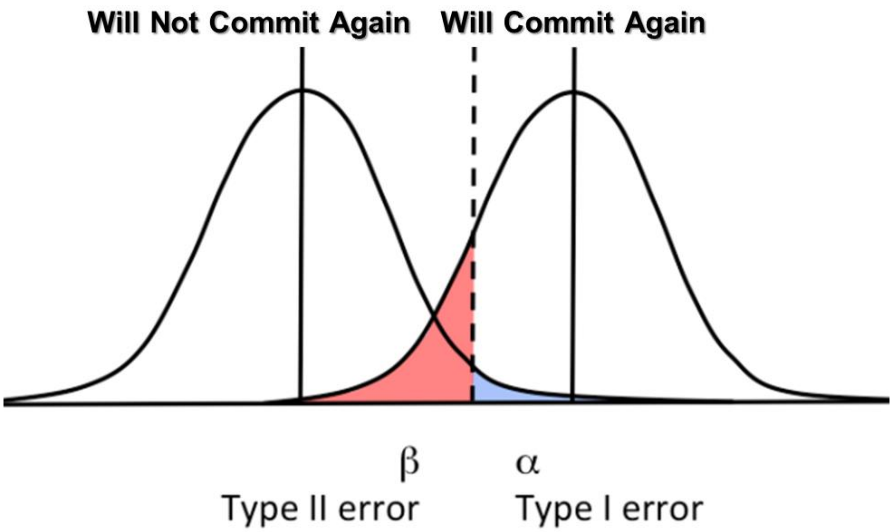
Some studies have found results that are anomalous. For example, Ortega and Hergovich (2017) find a statistically significant relationship between the rapid increase in the use of online dating sites and an increase in interracial marriages—an outcome they attribute to matching of people with similar interests by the online dating platforms who would never have met because of the racial separation endemic in US society.[5] This observation may be extendable to other areas, such as online sales platforms, particularly when buyer and seller are anonymous. They allow vendors from less advantaged groups to sell their wares without experiencing bias. In such cases, the platforms connect buyers and sellers in an economically "perfect," i.e., bias-free, marketplace. In such marketplaces, all reviews would be unbiased, thus eliminating evaluation bias, which numerous studies have shown is endemic. In fact, to address evaluation bias, many studies have suggested increasing anonymity (see, e.g., Doleac & Stein, 2013; Edelman et al., 2017).[6]

[5] This result was present for all racial groups. The study controls for the fact that interracial marriages also increased prior to the introduction of online dating. This paper does not suggest that racial discrimination does not affect platform-based dating sites. In fact, it is likely widespread on these sites, as in society.

[6] Of course, it may not be possible to remove all indicators that trigger bias. For example, Airbnb could eliminate all photos of the host(s), and yet show zip codes and addresses, which may serve as cues to race (Edelman et al. 2017). Further, particularly suspicious guests could use other platform services, such as Streetview, to obtain information that could lead to making a biased choice.

The studies of algorithmic decision-making in activity such as criminal sentencing, deciding probation, and even deciding whom police should stop have repeatedly shown that it leads to biased outcomes. Because all these decisions are based on predictions, human factors are intimately involved and weigh in the decisions. As our discussion of training data bias (Bias 1) indicates, this is not surprising, and we should be mindful that decisions in law enforcement are also based on preconceived notions and learned behavior. In principle, an algorithm should not be biased. For example, it is well known that members of the law enforcement community make decisions that are affected by the defendant's “demeanor,” dress, and a variety of other characteristics, such as race and gender. In the US, judicial decisions are predicated upon the standard of “beyond a reasonable doubt,” which is inherently subjective. The use of statistics raises concern over the commission of Type I and Type II errors (false positives and negatives), i.e., will or will not the individual commit a crime again (see Figure 3). This trade-off is fundamental in criminal justice decision-making. Effectively, bias occurs when characteristics that should not affect the judicial decision have a positive or negative impact. For example, individuals might be granted release because they are friends of the judge, or, alternatively, denied release because the judge does not like an individual’s “looks.”

Figure 3: Type I and Type II Errors Applied to Recidivism Assessment



In a recent paper, Kleinberg et al. (2017) test their machine-learning algorithm on existing data and find that it can do a better job than judges in making bail decisions. The first benefit of the algorithm is that it creates uniformity, whereas judges have significant individual discretion. Here, it is simply a matter of balancing out judges who are too lenient with those who are too strict. The algorithm can be constrained to provide outcomes that decrease the overall rate of reoffending, releasing more individuals while preventing an increase in racial disparities among those denied bail. Another constraint could change the algorithm in such a way as to reduce racial disparities among those who are incarcerated while also reducing the rate of reoffending by those who are released after serving their sentences. This is possible because individuals from some racial groups are incarcerated disproportionately highly, while individuals from other racial groups may have a higher propensity to reoffend. Many criticisms do not acknowledge that it may also be possible to use these tools to decrease both inequality and crime rates.

The dangers of these software machines have been explored throughout the paper. However, if it is correct to see them as machines, then in principle they should be adaptable for a variety of purposes. Understanding how they operate to generate output or results is vital for the development of better machines. In some cases, such as sales websites, reorienting the site to, for example, provide greater anonymity could decrease the possibility of bias, though biased evaluations are easier to identify and thus control. In cases such as evaluations of Uber drivers/customers or Airbnb hosts/renters, it may be more difficult to remove the causes of bias. As Kleinberg et al. (2017) indicate, in some cases, merely creating greater visibility, efficiency, and uniformity in human decision-making processes could lead to decreases in bias. Further, algorithms are machines that can be "quality controlled" to provide socially desirable output. Finally, and most important, because all digital actions leave a trace, they can be identified and addressed.

Another neglected opportunity for addressing bias is possible because all actions on a platform leave a data trail. Thus, for digital platforms it is a relatively trivial task to analyze user behavior. For example, Uber or Airbnb or a government agency could study the patterns of either provider or customer behavior to identify harmful bias. Similarly, Goel et al. (2017) argue that administrative data regarding police decisions to stop and frisk civilians could be mined to establish whether, stochastically speaking, this activity tended to target certain groups. If such a tendency exists, a correction could be introduced. However, they further argue that it is possible to statistically determine which types of stops are more likely to be productive. Thus, it is possible to use data to decrease false positives (searching with no success), while not affecting overall success rates. The important observation is that, although platforms and algorithms can often be used in anti-social ways, they also have the potential to assist in addressing inequality and bias. While social values are embodied in tools, it also possible for other parties to appropriate and reframe tools for different purposes. In this respect, human decisions remain vital.

Algorithms, AI, and machine learning may offer a vehicle for addressing the pernicious bias that affects our criminal justice system. To quote Barabas et al. (2017, p. 1):

> If machine learning is operationalized merely in the service of predicting individual future crime, then it becomes difficult to break cycles of criminalization that are driven by the iatrogenic effects of the criminal justice system itself. We posit that machine learning should not be used for prediction, but rather to surface covariates that are fed into a causal model for understanding the social, structural and psychological drivers of crime.

Effectively, this suggests that the algorithmic tools can be used to identify some of the deeper causes and types of bias and expose them for consideration. As tools, algorithms and platforms contain and express the notions and desires of their producers, which frames but does not entirely determine the actions of their users, which suggests that the sources of bias can emerge from the actions of either the tool makers or the tool users. In the case of multisided platforms, bias can emerge or be expressed on any side of the platform, even particular interactions between the sides. Yet because all the actions are digital and therefore leave a trace, they are visible and can be addressed.

Discussion and Conclusion

Given that algorithms and digital platforms have become central in socioeconomic life, it is not surprising to find increasing scrutiny of their operations and outcomes. As our review indicates, the issue of bias in algorithms and, by extension, platforms deserves significant attention, but only recently has it attracted such attention. During the Obama administration, a series of reports on the digital discrimination were commissioned (U.S. Executive Office of the

President, 2016, p. 14). Despite resulting in little action, they raised awareness of the issues. In 2018, the European Union's new directive on data protection stipulated that users had a right to explanations of the reasons that relevant algorithms generated particular results, though its response was, in certain respects, problematic (Kaminski, 2018). More recently, not only the popular press but also the social sciences are considering the implications of digital bias (Martin, 2018).

Our review has demonstrated the variety of vectors through which bias can be expressed in algorithmic processes. Some of them may be easily remedied, for example, ensuring that the data input are not congenitally biased or that the algorithms are not unwittingly developing bias. Other biases, such as user bias, may, in certain cases, be virtually unstoppable, though as Edelman and Luca (2014) suggest, digital platforms could take concrete actions to minimize the opportunities for bias. It has been argued that organizations have an interest in preventing algorithmic bias, and thus it would not be induced deliberately (Chander, 2016). That may be true, but it is often the case that the producers of algorithms may not even consider the large number of ways in which bias can be manifested.

Research on how to mitigate algorithmic bias has increased, but "correcting" the data to increase fairness is also hampered by determining what is "fair" (Courtland, 2018, p. 359). Some have suggested that transparency would provide protection against bias and other socially undesirable outcomes (Annany & Crawford, 2016). However, firms will resist because it is often in their best interest to protect their information and algorithmic decision-making processes, both legitimately to prevent gaming their processes and to prevent scrutiny (Citron & Pasquale 2016; Lepri et al., 2017). To prevent such gaming, some aspects of the algorithmic process must be kept hidden, including key inputs and outputs. Also, given the complicated nature of many of these algorithms and now with machine learning, in which computer scientists are not entirely certain how the machine derives outcomes, transparency may be a “protection” that cannot in practice provide accountability (Kuang, 2017; Gershgorn, 2017).

Bias can also be generated indirectly, making identification more difficult, especially when it simply mirrors current beliefs in the larger society. In such cases, the elimination of biases requires active countermeasures. Further, a single software program may contain more than one source of bias, and they may interact, creating greater difficulty in unraveling their sources. Diagnosis of bias often requires that ethnicity be entered into the model directly—something that is prohibited in employment or housing-related decision-making. Paradoxically, the use of a protected category in the algorithm to overcome possible discrimination due to indirect factors could expose the responsible organization to the charge of using protected data categories for decisions and thus legal violations (d'Alessandro, O'Neil, & LaGatta, 2017; Žliobaitė & Custers, 2016, p. 3). It might also invite litigation from what might be called “unprotected” parties, arguing that remedies actually introduce other biases.

Big data derived from sources such as social networks offers ever more opportunity for distinguishing and treating users more individually. In China, a new smartphone payment system generates an individual’s credit score not only on the basis of the user’s payment history but also the credit scores of those in the individual’s network. Because those with higher credit scores obtain discounts and need not make security deposits, individuals are reticent about adding individuals with lower credit scores to their network and, in fact, are likely to drop individuals from the network if their scores decline (Hvistendahl, 2017). The point of this illustration is that big data permits the shaping of social networks that could create surprising new opportunities for bias to be expressed.

Given the increased use of algorithms, big data, machine learning, and, especially, platforms in nearly all aspects of human activity, considering their impact on the difficult social issue of bias seems particularly important. Fortunately, this increasing reliance and integration of data into our daily lives also creates a rich living dataset in countless industries. This allows researchers to collect data and evaluate bias more easily than ever before.

The emerging frontier of bias research is in the realm of machine learning, in which the computer analyzes the data and builds its own predictors of outcomes. In an increasing number of cases, even the algorithm's creators may not be able to explain the computational outcomes, and thus mandating "transparency" is unlikely to prevent untoward decisions (De Laat, 2017). The difficulties seem even greater, as many of the most important algorithms and the data upon which they make decisions are evolving constantly. For example, Google's search algorithm is updated nearly daily, with larger updates monthly or quarterly (e.g., SearchEngineLand, 2017). In other words, many of these algorithmic systems and platforms are constantly evolving, making the auditing of their results perpetually contingent.

Platforms, algorithms, big data, and machine learning have become more important in our lives, shaping choices, alternatives, and outcomes. It is more important than ever to understand where and how social ills such as ethnic bias can be expressed and reinforced through these digital technologies. However, the critics who suggest that these technologies necessarily exacerbate bias may be too pessimistic. Although machine learning may reproduce bias in previously unforeseen ways, it is also true that all the inputs and activities on digital platforms create a digital record that can be examined with software tools. Whereas ethnic or other types of discrimination in the analog world were difficult and expensive to reveal and identify, in the digital world, they are both permanent and easier to analyze. For a society or researchers, this new world offers new dangers for reinforcing old biases with new tools but also tools that can facilitate identifying and addressing the continuing social problems related to ethnic and other types of bias.

REFERENCES

Ajunwa, I., Friedler, S., Scheidegger, C. E., & Venkatasubramanian, S. (2016). Hiring by algorithm: predicting and preventing disparate impact. Available at SSRN, https://friedler.net/papers/SSRN-id2746078.pdf.

Ambinder, L. P. (1995). Dispelling the Myth of Rationality: Racial Discrimination in Taxicab Service and the Efficacy of Litigation under 42 USC 1981. *George Washington Law Review*, 64, 342-378.

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 1, 17.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. ProPublica (May 23), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Angwin, J., & Parris, T. (2016). Facebook lets advertisers exclude users by race. *ProPublica* (October 28), https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race.

Angwin, J., & Tobin, A. (2017). Facebook (still) letting housing advertisers exclude users by race. ProPublica (November 21), https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

Barabas, C., Dinakar, K., Virza, J. I., & Zittrain, J. (2017). Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. arXiv preprint arXiv:1712.08238, https://arxiv.org/abs/1712.08238

Barley, S. R. (2015). Why the internet makes buying a car less loathsome: How technologies change role relations. *Academy of Management Discoveries*, 1(1), 5-35.

Barocas, S., Bradley, E., Honavar, V., & Provost, F. (2017). Big data, data science, and civil rights. arXiv preprint arXiv:1706.03102, https://arxiv.org/abs/1706.03102.

Barrett, L. (2017). Reasonably suspicious algorithms: Predictive policing at the United States border. *NYU Review of Law & Social Change*, 41, 327-365.

Barry-Jester, A. M., Casselman, B., & Goldstein, D. (2015). The new science of sentencing. The Marshall Project (August 8), https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing/.

Brown, A. E. (2018). Ridehail revolution: Ridehail travel and equity in Los Angeles. PhD dissertation, UCLA.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

Chander, A. (2016). The racist algorithm. *Michigan Law Review*, 115, 1023-1046.

Christin, A., Rosenblatt, A., & Boyd, D. (2015). Courts and predictive algorithms. *Data & CivilRight.* https://www.law.nyu.edu/sites/default/files/upload_documents/Angele%20Christin.pdf.

Citron, D. K., & Pasquale, F. A. (2014). The scored society: Due process for automated predictions. *Washington Law Review* (89)1, 1-33.

Courtland, C. (2018). The bias detectives. *Nature* 558 (June 21), 357-361.

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2), 120-134.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4691-4697). Melbourne, 2017. AAAI Press.

Datta, A., Datta, A., Makagon, J., Mulligan, D. K., & Tschantz, M. C. (2018, January). Discrimination in Online Advertising: A Multidisciplinary Inquiry. In *Conference on Fairness, Accountability and Transparency* (pp. 20-34). New York City.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1), 92-112.

De Laat, P. B. (2017). Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology*, 1-17. https://doi.org/10.1007/s13347-017-0293-z.

Doleac, J. L., & Stein, L. C. (2013). The visible hand: Race and online market outcomes. *Economic Journal*, 123(572), F469-F492.

Dormehl, L. (2018). Facial recognition has a race problem — here’s how Gyfcat is fixing that. Digitaltrends (January 25) https://www.digitaltrends.com/cool-tech/facial-recognition-gyfcat-race/

Dressel, J. & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*. 4(1) http://advances.sciencemag.org/content/advances/4/1/eaao5580.full.pdf

Eaglin, J. M. (2017). Constructing recidivism risk. *Emory Law Journal*, 67(59), 60-122.

Edelman, B. G., & Luca, M. (2014). Digital discrimination: The case of Airbnb.com. Harvard Business School Working Paper 14-054, January 10.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2), 1-22.

Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). "Be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In International AAAI Conference on Web and Social Media (pp. 62-71). Montreal,

Gallagher, K. (2005). Rethinking the Fair Credit Reporting Act: When requesting credit reports for employment purposes goes too far. *Iowa Law Review*, 91, 1593-1620.

Garvie, C. (2016). The perpetual line-up: Unregulated police face recognition in America. Georgetown Law, Center on Privacy & Technology.

Gershgorn, D. (2017). AI is now so complex its creators can't trust why it makes decisions. Quartz Media. https://qz.com/1146753/ai-is-now-so-complex-its-creators-cant-trust-why-it-makes-decisions/.

Gillespie, T. (2014). The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, ed. Gillespie, T., Boczkowski, P. and Foot, K. Cambridge, MA: MIT Press.

Gillespie, T. (2018). Platforms are not intermediaries. *Georgetown Law and Technology Review*, 2, 198-218.

Glaser, A. (2018). Want a terrible job? Facebook and Google may be hiring. Slate, January 18, https://slate.com/technology/2018/01/facebook-and-google-are-building-an-army-of-content-moderators-for-2018.html.

Goel, S., Perelman, M., Shroff, R., & Sklansky, D. A. (2017). Combatting police discrimination in the age of big data. *New Criminal Law Review: An International and Interdisciplinary Journal*, 20(2), 181-232.

Goodman, B. W. (2016, June). Economic models of (algorithmic) discrimination. *In 29th Conference on Neural Information Processing Systems* (Vol. 6). Barcelona.

Greenwood, B. N., Adjerid, I., & Angst, C. M. (2017). Race and gender bias in online ratings: An origins story. papers.ssrn.com,
https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1339&context=icis2017

Guskey, T. R., & Jung, L. A. (2016). Grading: Why you should trust your judgement. *ASCD: Educational Leadership*, 73(7), 50-54.

Ho, Y., and Salna, K. (2017). Indonesia says Google Agrees to monitor negative YouTube content. Bloomberg (August 4, 2017), https://www.bloomberg.com/news/articles/2017-08-04/indonesia-says-google-agrees-to-monitor-negative-youtube-content/.

Hogan, K. (2017). How Apple says it prevented Face ID from being racist. Gizmodo (October 16, 2017), https://gizmodo.com/how-apple-says-it-prevented-face-id-from-being-racist-1819557448/.

Hvistendahl, M. (2017). Inside China's vast new experiment in social ranking. *Wired* (December 14), https://www.wired.com/story/age-of-social-credit/.

Jackson, B. A., Banks, D., Woods, D., & Dawson, J. C. (2017). Future-proofing justice. Rand Corporation, Santa Monica.

Kakar, V., Franco, J., Voelz, J., & Wu, J. (2016). Effects of host race information on Airbnb listing prices in San Francisco. San Francisco State University, Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/69974/1/MPRA_paper_69974.pdf

Kaminski, M. E. (2018). The right to explanation, explained (June 15). Available at SSRN: https://ssrn.com/abstract=3196985/.

Kenney, M., & Zysman, J. (2016). The rise of the platform economy. *Issues in Science and Technology, 32*(3), 61-69.

King, A. G., & Mrkonich, M. J. (2015). Big data and the risk of employment discrimination. *Oklahoma Law Review,* 68, 555-584.

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6), 1789-1801.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. National Bureau of Economic Research. Working paper w23180.

Knight, W. (2017). Intelligent machines: The dark secret at the heart of AI. *Technology Review* (April 11). https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Kuang, C. (2017). Can A.I. be taught to explain itself? *New York Times*. (Nov. 2017).

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2017). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 1-17, https://doi.org/10.1007/s13347-017-0279-x

Lessig, L. (2009). *Code: And Other Laws of Cyberspace*. New York: Basic Books.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14-19.

Martin, K. (2018). Ethical implications and accountability of algorithms. *Journal of Business Ethics* 1-16, https://doi.org/10.1007/s10551-018-3921-3

Mayer-Schonberger, V., & Kenneth, C. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt.

Ma, A. (2018). China ranks citizens with a social credit system. *Business Insider* (April 8).

Mitchell, M. (2017). The seen and unseen factors influencing knowledge in AI systems. Powerpoint presentation at the 2017 Conference on Fairness, Accountability, and Transparency in Machine Learning, Halifax, Canada (August 14).

Noble, D. (2017). *Forces of Production: A Social History of Industrial Automation*. New York: Knopf.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Broadway Books.

Orlikowski, W. J. (2016). Digital work: A research agenda. In *A Research Agenda for Management and Studies,* ed. B. Czarniawska. Northampton, MA: Edward Elgar, 88-96.

Ortega, J., & Hergovich, P. (2017). The strength of absent ties: Social integration via online dating. arXiv preprint arXiv:1709.10478. https://arxiv.org/pdf/1709.10478.pdf.

Peck, D. (2013). Watching you at work. *Atlantic* (December), https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, *8*(2), 1-11.

Piskorski, M. J. (2014). *A Social Strategy: How We Profit from Social Media*. Princeton: Princeton University Press.

Pulliam-Moore, C. (2015). Google Photos identified black people as "gorillas," but racist software isn't new. July 1, 2015. Retrieved from https://splinternews.com/google-photos-identified-black-people-as-gorillas-but-1793848829/.

Rainie, L., & Anderson, J. (2017). Theme 7: The need grows for algorithmic literacy, transparency and oversight. Pew Research Foundation (February 8). Retrieved from http://www.pewinternet.org/2017/02/08/theme-7-the-need-grows-for-algorithmic-literacy-transparency-and-oversight/.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *Knowledge Engineering Review*, 29(5), 582-638.

Scott, S. V., & Orlikowski, W. J. (2012). Reconfiguring relations of accountability: Materialization of social media in the travel sector. *Accounting, Organizations and Society*, 37(1), 26-40.

SearchEngineLand. (2017). 8 major Google algorithm updates, explained. (September 19), https://searchengineland.com/8-major-google-algorithm-updates-explained-282627/.

Seaver, N. (2012). Algorithmic recommendations and synaptic functions. *Limn*, 1(2).

Solon, O. (2017). Facebook is hiring moderators. But is the job too gruesome to handle? Guardian (May 4, 2017). Retrieved from https://www.theguardian.com/technology/2017/may/04/facebook-content-moderators-ptsd-psychological-dangers/.

Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., ... & Mislove, A. (2018, January). Potential for discrimination in online targeted advertising. In *Conference on Fairness, Accountability and Transparency* (pp. 5-19).

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, 66, 803-872.

Surden H. 2017 Values embedded in legal artificial intelligence. University of Colorado Law Legal Studies Research Paper No. 17-17.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3), 1-19.

U.S. Executive Office of the President. (2014). Big Data: Seizing Opportunities, Preserving Values. (May). Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

U.S. Executive Office of the President. (2016). Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. (May). Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Vincent, J. (2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. The Verge (March 24, 3016). Retrieved from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist/.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121-136.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1), 118-132.

Zhao, C. (2017). Is the iPhone racist? Apple refunds device that can't tell Chinese people apart. *Newsweek*, December 18, 2017. Retrieved from http://www.newsweek.com/iphone-x-racist-apple-refunds-device-cant-tell-chinese-people-apart-woman-751263/.

Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2), 183-201.

Zuboff, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*. New York: Basic Books.