Received 22 June 2018
Revised 22 June 2018
Accepted 22 June 2018

Detecting racial bias in algorithms and machine learning

Nicol Turner Lee
Center for Technology Innovation, Brookings Institution, Washington, District of Columbia, USA

Abstract

Purpose – The online economy has not resolved the issue of racial bias in its applications. While algorithms are procedures that facilitate automated decision-making, or a sequence of unambiguous instructions, bias is a byproduct of these computations, bringing harm to historically disadvantaged populations. This paper argues that algorithmic biases explicitly and implicitly harm racial groups and lead to forms of discrimination. Relying upon sociological and technical research, the paper offers commentary on the need for more workplace diversity within high-tech industries and public policies that can detect or reduce the likelihood of racial bias in algorithmic design and execution.

Design/methodology/approach – The paper shares examples in the US where algorithmic biases have been reported and the strategies for explaining and addressing them.

Findings – The findings of the paper suggest that explicit racial bias in algorithms can be mitigated by existing laws, including those governing housing, employment, and the extension of credit. Implicit, or unconscious, biases are harder to redress without more diverse workplaces and public policies that have an approach to bias detection and mitigation.

Research limitations/implications – The major implication of this research is that further research needs to be done. Increasing the scholarly research in this area will be a major contribution in understanding how emerging technologies are creating disparate and unfair treatment for certain populations.

Practical implications – The practical implications of the work point to areas within industries and the government that can tackle the question of algorithmic bias, fairness and accountability, especially African-Americans.

Social implications – The social implications are that emerging technologies are not devoid of societal influences that constantly define positions of power, values, and norms.

Originality/value – The paper joins a scarcity of existing research, especially in the area that intersects race and algorithmic development.

Keywords Artificial intelligence, Communication, Advertising, Computer ethics, Civil society, Civil race relations, Race and political rights

Paper type Research paper

Introduction

The online economy has not resolved the issue of racial bias in its applications. In 2013, online search results for "black-sounding" names were more likely to link arrest records with profiles, even when false (Lee, 2013). Two years later, Google apologized for an algorithm that automatically tagged and labeled two African–Americans as "gorillas" after an innocuous online word search (Kasperkevic, 2015). The online photo-shopping application, FaceApp, was later found to be lightening the darker skin tones of African–Americans because European faces dominated the training data, thereby defining the standard of beauty for the algorithm (Morse, 2017).

Algorithms are procedures that facilitate automated problem-solving, or a sequence of unambiguous instructions (C.T., 2017). In their controversies, Google explained their biases

Journal of Information, Communication and Ethics in Society
Vol. 16 No. 3, 2018
pp. 252-260
© Emerald Publishing Limited
1477-996X
DOI 10.1108/JICES-06-2018-0056

as problems associated with the algorithm or the inappropriate meta-tagging of images. The developer who was applying the racist filter to lighten skin tones blamed it on the lack of representation of African–Americans in the application's training data. While these biases can be corrected, the intended or unintended consequences either support or degrade both the fairness and the confidence in algorithms.

Algorithmic bias is a byproduct of sophisticated online predictive modeling. While scholars have proven that algorithms are more efficient and expeditious in identifying and solving problems, how and when these procedures lead to bias is worth further exploration, along with the role of public policies to address and potentially legislate that bias. On the latter point, policymakers may find themselves determining which biases are more harmful to protected classes, including people of color, women, people with disabilities, and run counter to the public interest.

In this article, I argue that algorithmic biases explicitly or implicitly harm racial groups and share examples where discrimination has been identified and reported. The paper also examines the role of unconscious bias in the design and execution of certain algorithms, especially those that reinforce inequalities. Relying on sociological and technical research, I offer commentary on the need for increased diversity within high-tech industries to address the skewed assumptions of developers and conclude with current public policies that are being used to detect and reduce the likelihood of racial bias in algorithmic design and execution.

Big data and explicit bias

Online data are collected in real-time from users through a series of interactions with websites, social media communities, e-commerce vehicles, and general online inquiries for information of interest. These small portions of data become compiled, mined and eventually repurposed for commercial or public use. Big Data serves a variety of purposes, from helping to advance breakthroughs in science, health care, energy and transportation to enhancing government efficiencies by aggregating citizen input.

Big Data can also exclude populations. In a report published by the Federal Trade Commission, the agency whose responsibilities include regulatory oversight of high-tech companies, when Big Data analytics are misapplied, online users can be tracked or profiled based on their online activities and behaviors (Ramirez *et al.*, 2016). Consequently, online users can be denied credit based on their Web browsing history, or aggregated predictive analytics can wrongly determine an individual's suitability for future employment, personal credit or an educational opportunity. Online proxies, including one's zip code, can also be used by marketers to extrapolate an individual's socioeconomic status based on her neighborhood, resulting in incorrect assumptions about one's lifestyle or preferences (Noyes, 2015). In these and other examples, Big Data, when misused, can lead to the disparate treatment of protected classes, which are distinguishable by their race, gender, age, ability, religion, and sexual orientation.

Not surprisingly, high-tech companies often find themselves wedged between the values of "permissionless innovation," which seeks to remove barriers to entry for technology experimentation and the social responsibility to protected classes (Daigle, 2014). In the latter case, the algorithm, when applied to these vulnerable populations, may replicate structural or explicit discrimination or generate new forms of bias, usually implicit or unconscious. While the developer's intent for the algorithm may not start out being discriminatory or with prejudicial intent, the resulting algorithm can adapt over time to the explicit and implicit societal biases that exist, fostering stereotypes and unfair profiling. The algorithmic output can also be used to bias unsuspecting online users.

Some independent contractors of companies, such as Airbnb, Uber and Lyft, whose business models are enabled by Big Data, have used the information exposed by algorithms to explicitly discriminate against consumers. In 2017, Airbnb, an online home-sharing company and application, found that some hosts were rejecting renters based on race, age, gender, and other factors (Murphy, 2016). In these cases, the intent was due to their hostility toward the online, public profiles of specific customers. While Airbnb has worked to eradicate bias on their site through community commitment agreements (Airbnb, 2018) that reinforce legal compliance, the harmed renters were subjected to unfair or disparate racial treatment.

Among ride-sharing services, researchers exposed similar occurrences of discrimination. Uber and Lyft drivers were found to be either canceling rides or extending the wait times of African–American customers in Boston and Seattle (Ge *et al.*, 2016). In a sample of 1,500 rides in both cities, the study found that Uber drivers were more likely to cancel on riders with "black-sounding" names, and that African–American men typically waited longer to be picked up (Ge *et al.*, 2016). African–American customers were also "screened out" by Lyft drivers through a review of their names and faces upon the order. Other studies concluded that women were taken on longer routes to extend the cost of the fare (Ge *et al.*, 2016).

In 2017, a report by ProPublica exposed a controversial online function on Facebook, the social media platform, that allowed advertisers to exclude members of "ethnic affinity" groups, primarily people of color, from targeted marketing for certain ads (Angwin *et al.*, 2017). Those ads were specifically focused on housing, employment, and the extension of credit. Facebook immediately disabled that function and forbade advertisers from engaging in discriminatory practices on their site. In these cases, the data collected from the algorithms were being used to target and exclude online users because of their race.

Countering certain forms of explicit discrimination can be remedied in the US under federal laws that govern equal opportunity for protected classes. In 1964, Congress passed Public Law 88-352, which "forbade discrimination on the basis of sex, as well as race in hiring, promoting, and firing" (United States Congress, 1964). The Civil Rights Act of 1968 was later amended to include the Fair Housing Act, which further prohibits discrimination in the sales, rental and financing of dwellings, and in other housing-related transactions to federally-mandated protected classes (Fair Housing Center, 2018). Enacted in 1974, the Equal Credit Opportunity Act prohibits any creditor from discriminating against any applicant from any type of credit transaction based on protected characteristics (Department of Justice, 2017). While these laws may redress unlawful explicit bias and harm emanating from the online economy, they fall short in the mitigation of implicit and unconscious biases that are often embedded in the algorithmic design.

Implicit and unconscious bias

The Kirwan Center for the Study of Race and Ethnicity defines implicit bias as "the attitudes or stereotypes that affect our understanding, actions, and decisions in an unconscious manner" (Kirwin Center, 2015). Citing individuals' common susceptibility to these biases, the Kirwan Center found that it is the nature of homogenous associations and relationships to harbor particular feelings and attitudes about others based on race, ethnicity, age, and appearance (Kirwin Center, 2015). Outside of technology, implicit bias has been seen in colleges and universities when white professors are less likely to respond to students of colors requesting time with them during their office hours (Vedatam, 2014). It can also surface as more privileged classes apply negative stereotypes to racial and ethnic groups without proof of their assumptions.

Compared to explicit bias, implicit bias in the digital age can appear in courts and parole boards that have become increasingly reliant upon predictive analytics to determine future criminal behavior or appropriate bail and sentencing limits. Some researchers have found that many of the predictive algorithms are inaccurate or wrought with societal stereotypes, leading to African–Americans being depicted as more likely to commit violent crimes than whites (Angwin and Larson, 2016). For example, questions have emerged around the race neutrality of predictive sentencing models, such as the popular COMPAS (Yong, 2017) algorithm, which assigns risk scores between 1 and 10 to assess the likelihood of a defendant's future criminal activity. Based on the algorithm, defendants with scores of 7 are more likely to reoffend at twice the rate as those with scores of 3 (Corbette-Davies *et al.*, 2017). High-risk defendants are more likely to be detained while awaiting trial based on their COMPAS score. Unfortunately, when these predictions are not accurate, certain groups suffer irreparable effects, especially blacks who are historically unjustly punished and more harshly penalized than whites.

Implicit bias also presents itself in the complex calculations of machine learning and artificial intelligence (AI). In her research on "word embedding," which is commonly used in language translation apps, Joanna Bryson found that this type of bias creates issues for machines that do not have the moral compass of humans when it comes to identifying stereotypical traits (Caliskan *et al.*, 2017). Researchers discovered that words that included "female" and "women" were more likely to be associated with arts and humanities occupations, while "male" and "man" were often correlated with math and engineering jobs, thereby creating false positives and negatives (Caliskan *et al.*, 2017). The same study also surfaced that European American-sounding names were more likely as associated with pleasant word associations, while "black-sounding" names were often associated with unpleasant words (Caliskan *et al.*, 2017). Consequently, stereotypes about African–Americans remain pervasive.

The negative implicit assumptions associated with words, along with the racial bias emanating from predictive criminal justice models, unmask the fact that algorithms are not necessarily devoid of societal biases, prejudices, stereotypes, and even incorrect assessments about people and their circumstances.

The lack of diversity in high-tech industries

The arguments around how biases surface within algorithms still remain unsettled. Further, the relationship between less diverse workforces and algorithmic bias is provocative, especially given the intentionality of such discrimination. Recent diversity statistics report these companies employ less than 2 per cent of African–Americans in senior executive positions and 3 per cent of Hispanics compared to 83 per cent of whites (Atwell, 2016). Asian–Americans comprise just 11 per cent of executives in high-tech companies (Atwell, 2016). In the occupations of computer programmers, software developers, database administrators and even data scientists, African–Americans and Hispanics collectively are under 6 per cent of the total workforce, while whites make up 68 per cent (EEOC, 2016).

Even when people of color are employed in high-tech industries, the feelings of professional and social isolation also have been shown to marginalize these employees, potentially restricting their active workplace engagement, affecting their participation in the feedback loop and contributing to higher rates of attrition (Scott *et al.*, 2017). At Google, employees have been subjected to anti-diversity memos (Conger, 2017) and women have experienced documented backlash from male employees on hiring. This alienation within high-tech workforces neither encourages nor welcomes diverse input into work products.

For example, the Google applications developer whose algorithm led to the misidentification of African–Americans as "gorillas" pointed out that he did not anticipate the technology's faulty translation of darker-skinned faces (Miller, 2017), which could have been averted with a more diverse work team who would be sensitive to these issues.

Coined by some researchers as "inattentional blindness," technologists are not necessarily trained to identify cues that are outside of their cultural context and can be fenced into work groups that share similar experiences, values and beliefs. For example, when the algorithm for FaceApp lightened the skin tones of black users, it was unconsciously (and perhaps explicitly) signaling mainstream, or European American, standards of beauty and applying them to blacks – a compelling reason for why racial diversity was needed on the design team (Morse, 2017).

These unconscious bias errors strongly support why high-tech companies should be striving for more diverse workforces to identify and quell online discrimination. Companies that are disrupting societal norms through the sharing economy, social media and the internet of things must do better to address the less than remarkable representation of people of color as creators, influencers and decision-makers.

Mitigating racial bias in algorithms

Recent research has already started to dive into the various types of racial bias described in this paper. While most research acknowledges the problem, differences do exist as to where and when bias is conceived in the modeling process. Barocas, Bradley, Honvar *et al.* (2017) argue that the origins can be traced to the training data, especially when non-representative samples of populations lead to algorithmic models that exhibit systematic errors (Corbette-Davies *et al.*, 2017). These deficits often lead AI and machine learning to struggle with the identification and coding of people of color when compared to largely European populations, largely due to insufficient examples of minority behavior in the system (Corbette-Davies *et al.*, 2017). According to these authors, these discrepancies not only affect the accuracy of data used to create the algorithm but also may encode prejudicial and biased assessments. Moreover, when the training data are insufficient, algorithms are more likely to generate either disparate outcomes or unfair treatment toward specific individuals or communities of users when dark skin tones are manipulated to be lighter or blacks are more likely to be seen as violent.

Bucher (2012) argues that AI and machine learning can also contribute to online invisibility, especially given the dominance of certain algorithms to sort, classify and rank information (Bucher, 2012). Analyzing the algorithms that power the news feeds on Facebook, Bucher concludes that algorithms are not simple constructions of "black-boxes," but instead organized regimes for visibility, which prioritize certain content over others (Bucher, 2012).

However, what is still undetermined in the research is if racial bias can be detected in the beginning stages of the algorithm's development and execution. On this point, a discussion of algorithmic fairness seems appropriate. Among data scientists, three different approaches to achieving some level of fairness are being explored. *Statistical parity* ensures that training data sets have an equal proportion of subjects, such as the same of defendants from the same racial group. *Conditional statistical parity* controls for a set of plausible risk factors within an equal proportion of subjects, i.e. defendants with the same number of convictions and of equal racial proportion. Finally, *predictive equality* assumes that "the accuracy of decisions is equal across race groups, as measured by false positive rate (FPR)" (Corbette-Davies *et al.*, 2017). Within these three approaches is the goal of removing or reducing unwanted discrimination in the model-building process (Barocas *et al.*, 2017). However, given the limits on certain training data, technologists often find themselves juxtaposing fairness against accuracy, or vice versa, when constructing models.

Policymakers, however, tend to be stuck on the concept of algorithmic equity. Under former President Obama, a report addressing algorithmic systems and civil rights proposed an equal opportunity design framework to mitigate discrimination along historical, social and technological contexts (Obama White House, 2016). The findings argued that technology should not be designed in a vacuum, but rather account for all of the potential disparities in its platform and execution. Consequently, algorithms that discriminate should be dealt with, fixed or abandoned.

While policymakers view equity as the ultimate goal when mitigating algorithmic bias, it is not an easy task. As discussed, computer and data scientists are constantly weighing the sacrifices of hard data trade-offs against accuracy in their models when and where data are lacking for certain populations or issues. On this first point, more collaboration is needed to engender robust and ethical models that can either predict or eliminate racial bias from the onset, while protecting certain groups from unnecessary harms.

Second, algorithms are not detached from the cumulative social inferences that surface about online users. In her book, *Weapons of Math Destruction*, author Cathy O'Neil writes that "[t]he [algorithmic] models being used today are opaque, unregulated, and uncontestable, even when they are wrong" (O'Neil, 2016). Researcher Latanya Sweeney found online search results that constantly bombarded African–American users with ads asking, "Have you ever been arrested?" (Bray, 2013) This same inferential intelligence prompted users of a Grindr, a location-based social networking tool for gay men, to download a sex offender location-tracking app (Ananny, 2011). What is identified in these cases are algorithms that tap into the casual associations of online users to infer other behaviors, which relate to broader issues of algorithmic transparency.

Incorrect social inferences can be mitigated by giving consumers control over their digital footprint. In this case, the transparency of algorithms and how users can control what is implied about them becomes critically important (Barocas *et al.*). But more often than not, the biased assumption has already occurred and online users are not aware of what data are factored into the judgement.

Given the unsettled opinion on when and how racial discrimination should be mitigated for algorithms, what is the role of policymakers to address and redress known bias? Moreover, how can policymakers collaborate with the community of data scientists and developers on the design and execution of models to promote both transparency and fairness?

How policymakers are addressing algorithmic bias

The role of public policy in both identifying and mitigating racial biases are complicated, given the plausibility of social inferences factored into algorithms and the resulting disparate treatment and outcomes for vulnerable populations. In this last section, I share examples of how policymakers are confronting algorithmic bias to protect the public interest.

Recently, Congress introduced legislation on algorithmic bias called the "Future of AI Act" established a 19-person federal advisory committee within the US Commerce Department to track its growth and recommend best practices (Breland, 2017). While it is unclear if legislation will be the solution to this problem, policymakers desire to become more aware of disparate systems and impacts in machine learning and AI.

New York City has already instituted legislation to establish its own task force to review the algorithms that the city uses, ranging from educational to public safety applications. The task force will facilitate the testing of algorithms, determine how citizens can request input on algorithmic decisions (especially when they do not prefer the outcome), and investigate whether the source code for city agencies should be publicly available. The city will also explore if certain algorithms are biased against certain residents and how they hold companies accountable. However, company-controlled proprietary algorithms are not usually open and available to city governments, thereby limiting the city's ability to institute accountability over code outputs.

While it is not yet determined if New York City's somewhat proscriptive expectations will reduce algorithmic bias, other cities and counties are identifying additional ways to tackle this problem. In Allegheny County, PA, the Department of Human Services has built its own algorithm to better manage children protective services and the more than 10,000 calls received daily to their office. Working with researchers, county human service officials created the Allegheny Family Screening Tool, which improves the staff's decision-making processes by refocusing resources on children with the highest needs (Giammarise, 2017). The county's predictive modeling resource embeds more than 100 variables, ranging from child welfare services to exposure to criminal justice systems to generate a risk score ranging from 1 to 20, with the latter being the predictor of highest risk and the likelihood of a referral or home removal for the child (Giammarise, 2017). The tool primarily assists with screenings related to general protective services, including living conditions, home supervision, substance abuse by parents or guardians, among other non-abuse calls. Compared to New York City, Allegheny County owns the rights to their algorithm and, as a result, can update the code and track when the algorithm generates incorrect predictions.

What is being done in New York City, Allegheny County and by federal lawmakers highlight the tension between policymakers and industries over the openness and transparency of algorithms, especially in applications that are nested in public domains. For example, advocates for algorithmic openness have been critical of New York City's legislation, suggesting that existing closed models, or algorithms, insulate identified biases from public scrutiny (Abraham, 2017). On the other hand, some companies suggest that the release of proprietary algorithms stifles innovation and discourages them from working with cities and other policymakers. However, the Allegheny County model offers insight into how cities should be innovating and collaborating with developers to maintain pace with new technology. The county also employs a data scientist to build and manage in-house algorithms, which helps harmonize the city's goals with the technical architecture.

What is clear in all of government proposals is that policymakers have at least acknowledged that these computations are not divorced from systemic bias that has consequences for vulnerable populations. Moving forward, the extent to which regulatory policy or legislation is needed to ensure equity in algorithms should be further debated.

Conclusion

The presentation of the explicit and implicit examples of algorithmic bias in this paper still leaves open a critical question of whether computers or humans are the culprits in engineering many of these systemic inequalities online. Just like in society, bias – much like racism – is a learned behavior. The algorithm's adaptation or re-training from its original code conforms with power structures, societal expectations, beliefs and values, resulting in the reasons why algorithms oppress (Noble, 2018).

While technologists are working to create technical processes that promote more fairness in predictions and policymakers are embracing their share of responsibility around online bias, threats of discrimination will continue to pervade machine learning and AI. As algorithms ultimately mimic the existing power structures evidenced in society, blind faith or homogenous workforces are no longer applicable frameworks for innovating these fields.

To ensure that more of these biases do not become commonplace and end up further deepening the inequalities already imposed on people of color, it is important that policymakers and technologists agree upon principles and values for what a "bias-free" zone within innovation should adhere to. Companies also need to be more proactive in employing people of color, or those from different backgrounds, who can factor diversity into the design and execution of algorithms. Until then, algorithms and their associated biases will become mirrors of structural discrimination, rather than bridges to opportunity, equality and efficiency.

References

Abraham, R. (2017), "New York city passes bill to study biases in algorithms used by the city", *Motherboard*, available at: https://motherboard.vice.com/en_us/article/xw4xdw/new-york-city-algorithmic-bias-bill-law (accessed 5 March 2018).

Airbnb (2018), "Airbnb's nondiscrimination policy: our commitment to inclusion and respect", Airbnb, available at: https://goo.gl/a9wDAS (accessed 5 March 2018).

Ananny, M. (2011), "The curious connection between apps for gay men and sex offenders", *The Atlantic*, available at: www.theatlantic.com/technology/archive/2011/04/the-curious connection-between-apps-for-gay-men-and-sex-offenders/237340/ (accessed 5 March 2018).

Angwin, J. and Larson, J. (2016), "Bias in criminal risk scores is mathematically inevitable, researchers say", Propublica, available at: https://goo.gl/S3Gwcn (accessed 5 March 2018).

Angwin, J., Tobin, A. and Varner, M. (2017), "Facebook (still) letting housing advertisers exclude users by race", *ProPublica*, available at: https://goo.gl/Vk4jrs (accessed 5 March 2018).

Atwell, J. (2016), "Lack of women and minorities in senior investment roles at venture capital firms", *Deloitte*, available at: https://goo.gl/iah1VZ (accessed 5 March 2018).

Bray, H. (2013), "Racial bias alleged in google ad results", *Boston Globe*, available at: www.bostonglobe.com/business/2013/02/06/harvard-professor-spots-web-searchbias/PtOgSh1ivTZMfyEGj00X4I/story.html (accessed 5 March 2018).

Breland, A. (2017), "Lawmakers introduce bipartisan AI legislation", *The Hill*, available at: http://thehill.com/policy/technology/364482-lawmakers-introduce-bipartisan-ai-legislation (accessed 5 March 2018).

Bucher, T. (2012), "Want to be on top? Algorithmic power and the threat of invisibility on Facebook", *New Media and Society*, Vol. 14 No. 7, pp. 1164-1180.

C.T. (2017), "What are algorithms?", *The Economist*, available at: https://goo.gl/C32K1f (accessed 5 March 2018).

Caliskan, A., Bryson, J. and Narayanan, A. (2017), "Semantics derived automatically from language corpora contain human-like biases", *Science*, available at: https://goo.gl/TxaG6d (accessed 5 March 2018).

Conger, K. (2017), "Here's the 10-page anti-diversity screed circulating internally at google", *Gizmodo*, available at: https://gizmodo.com/exclusive-heres-the-full-10-page-anti-diversityscreed-1797564320 (accessed 5 March 2018).

Corbette-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. (2017), "Algorithmic decision making and the cost of fairness", *Conference on Knowledge, Discovery, and Data Mining*, available at: https://goo.gl/WDaqTX (accessed 5 March 2018).

Daigle, L. (2014), "Permissionless innovation — openness, not anarchy", *Internet Society*, available at: https://goo.gl/2GtAQ8 (accessed 5 March 2018).

Department of Justice (2017), "The equal credit opportunity act", *The US Department of Justice*, available at: https://goo.gl/7FZGCH (accessed 5 March 2018).

EEOC (2016), "Diversity in high-tech", *EEOC*, available at: https://goo.gl/EwKBUJ (accessed 5 March 2018).

Fair Housing Center (2018), "Federal fair housing act", *The Fair Housing Center of Greater Boston*, available at: https://goo.gl/gKsevR (accessed 5 March 2018).

Ge, Y. Knittel, C.R., MacKenzie, D. and Zoepf, S. (2016), "Racial and gender discrimination in transportation network companies", *The National Bureau of Economic Research*, available at: https://goo.gl/eY10GF (accessed 5 March 2018).

Giammarise, K. (2017), "Allegheny county using algorithm to assist in child welfare screening", *Pittsburgh Post Gazette*, available at: www.postgazette.com/local/region/2017/04/09/Allegheny-County-using-algorithm-to-assist-in-childwelfare-screening/stories/201701290002 (accessed 5 March 2018).

Kasperkevic, J. (2015), "Google says sorry for racist auto-tag in photo app", *The Guardian*, available at: https://goo.gl/ZEJYng (accessed 5 March 2018).

Kirwin Center (2015), "Implicit bias review 2015: state of the science", *Kirwan Center for the Study of Race and Ethnicity*, available at: https://goo.gl/RRmSLG (accessed 5 March 2018).

Lee, D. (2013), "Google searches expose racial bias, says study of names", *BBC*, available at: https://goo.gl/P8oodF (accessed 5 March 2018).

Miller, D. (2017), "Design biases in Silicon Valley are making the tech we use toxic, expert says", *Australian Broadcasting Company*, available at: https://goo.gl/6WGaon (accessed 5 March 2018).

Morse, J. (2017), "App creator apologizes for 'racist' filter that lightens skin tones", *Mashable*, available at: https://mashable.com/2017/04/24/faceapp-racism-selfie/#zeUItoQB5iqI (accessed 5 March 2018).

Murphy, L. (2016), "Airbnb's work to fight discrimination and build inclusion", *Airbnb Blogs*, available at: https://goo.gl/RUXc6j (accessed 5 March 2018).

Noble, S. (2018), *Algorithms of Oppression*, New York University Press, New York, NY.

Noyes, K. (2015), "Will big data help end discrimination – or make it worse? fortune", available at: https://goo.gl/VnPM1i (accessed 5 March 2018).

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

Obama White House (2016), "Big data: a report on algorithmic systems, opportunity, and civil rights", *White House Archives*, available at: https://goo.gl/By6wST (accessed 5 March 2018).

Ramirez, E., Brill, J., Ohlhausen, M.K. and McSweeny, T. (2016), "Big data: a tool for inclusion or exclusion", *FTC*, available at: https://goo.gl/wUxwU1 (accessed 5 March 2018).

Scott, A., Kapor Klein, F. and Onovakpuri, U. (2017), "Tech leavers study", *Kapor Center*, available at: www.kaporcenter.org/wpcontent/uploads/2017/08/TechLeavers2017.pdf (accessed 5 March 2018).

United States Congress (1964), *Civil Rights Act of 1964*, available at: https://goo.gl/tWyQIi (accessed 5 March 2018).

Vedatam, S. (2014), "Evidence of racial, gender biases found in faculty mentoring", *NPR*, available at: https://goo.gl/hMKsbd (accessed 5 March 2018).

Yong, E. (2017), "A popular algorithm is no better at predicting crimes than random people", *The Atlantic*, available at: https://goo.gl/VRnD6K (accessed 5 March 2018).

Corresponding author
Nicol Turner Lee can be contacted at: nturnerlee@brookings.edu