BROOKINGS

Report

Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms

Nicol Turner Lee, Paul Resnick, and Genie Barton Wednesday, May 22, 2019

Introduction

The private and public sectors are increasingly turning to artificial intelligence (AI) systems and machine learning algorithms to automate simple and complex decision-making processes.[1] The mass-scale digitization of data and the emerging technologies that use them are disrupting most economic sectors, including transportation, retail, advertising, and energy, and other areas. AI is also having an impact on democracy and governance as computerized systems are being deployed to improve accuracy and drive objectivity in government functions.

The availability of massive data sets has made it easy to derive new insights through computers. As a result, algorithms, which are a set of step-by-step instructions that computers follow to perform a task, have become more sophisticated and pervasive tools for automated decision-making.[2] While algorithms are used in many contexts, we focus on computer models that make inferences from data about people, including their identities, their demographic attributes, their preferences, and their likely future behaviors, as well as the objects related to them.[3]

“Algorithms are harnessing volumes of macro- and micro-data to influence decisions affecting people in a range of tasks, from making movie recommendations to helping banks determine the creditworthiness of individuals.”

In the pre-algorithm world, humans and organizations made decisions in hiring, advertising, criminal sentencing, and lending. These decisions were often governed by federal, state, and local laws that regulated the decision-making processes in terms of fairness, transparency, and equity. Today, some of these decisions are entirely made or influenced by machines whose scale and statistical rigor promise unprecedented efficiencies. Algorithms are harnessing volumes of macro- and micro-data to influence decisions affecting people in a range of tasks, from making movie recommendations to helping banks determine the creditworthiness of individuals.

[4] In machine learning, algorithms rely on multiple data sets, or training data, that specifies what the correct outputs are for some people or objects. From that training data, it then learns a model which can be applied to other people or objects and make predictions about what the correct outputs should be for them.[5]

However, because machines can treat similarly-situated people and objects differently, research is starting to reveal some troubling examples in which the reality of algorithmic decision-making falls short of our expectations. Given this, some algorithms run the risk of replicating and even amplifying human biases, particularly those affecting protected groups.[6] For example, automated risk assessments used by U.S. judges to determine bail and sentencing limits can generate incorrect conclusions, resulting in large cumulative effects on certain groups, like longer prison sentences or higher bails imposed on people of color.

In this example, the decision generates "bias," a term that we define broadly as it relates to outcomes which are systematically less favorable to individuals within a particular group and where there is no relevant difference between groups that justifies such harms.[7] Bias in algorithms can emanate from unrepresentative or incomplete training data or the reliance on flawed information that reflects historical inequalities. If left unchecked, biased algorithms can lead to decisions which can have a collective, disparate impact on certain groups of people even without the programmer's intention to discriminate. The exploration of the intended and unintended consequences of algorithms is both necessary and timely, particularly since current public policies may not be sufficient to identify, mitigate, and remedy consumer impacts.

With algorithms appearing in a variety of applications, we argue that operators and other concerned stakeholders must be diligent in proactively addressing factors which contribute to bias. Surfacing and responding to algorithmic bias upfront can potentially avert harmful impacts to users and heavy liabilities against the operators and creators of algorithms, including computer programmers, government, and industry leaders. These actors comprise the audience for the series of mitigation proposals to be presented in this paper because they either build, license, distribute, or are tasked with regulating or legislating algorithmic decision-making to reduce discriminatory intent or effects.

Our research presents a framework for *algorithmic hygiene*, which identifies some specific causes of biases and employs best practices to identify and mitigate them. We also present a set of public policy recommendations, which promote the fair and ethical deployment of AI and machine learning technologies.

This paper draws upon the insight of 40 thought leaders from across academic disciplines, industry sectors, and civil society organizations who participated in one of two roundtables.[8] Roundtable participants actively debated concepts related to algorithmic design, accountability, and fairness, as well as the technical and social trade-offs associated with various approaches to bias detection and mitigation.

Our goal is to juxtapose the issues that computer programmers and industry leaders face when developing algorithms with the concerns of policymakers and civil society groups who assess their implications. To balance the innovations of AI and machine learning algorithms with the protection of individual rights, we present a set

of public policy recommendations, self-regulatory best practices, and consumer-focused strategies–all of which promote the fair and ethical deployment of these technologies.

Our public policy recommendations include the updating of nondiscrimination and civil rights laws to apply to digital practices, the use of regulatory sandboxes to foster anti-bias experimentation, and safe harbors for using sensitive information to detect and mitigate biases. We also outline a set of self-regulatory best practices, such as the development of a bias impact statement, inclusive design principles, and cross-functional work teams. Finally, we propose additional solutions focused on algorithmic literacy among users and formal feedback mechanisms to civil society groups.

The next section provides five examples of algorithms to explain the causes and sources of their biases. Later in the paper, we discuss the trade-offs between fairness and accuracy in the mitigation of algorithmic bias, followed by a robust offering of self-regulatory best practices, public policy recommendations, and consumer-driven strategies for addressing online biases. We conclude by highlighting the importance of proactively tackling the responsible and ethical use of machine learning and other automated decision-making tools.

Examples of algorithmic biases

Algorithmic bias can manifest in several ways with varying degrees of consequences for the subject group. Consider the following examples, which illustrate both a range of causes and effects that either inadvertently apply different treatment to groups or deliberately generate a disparate impact on them.

Bias in online recruitment tools

Online retailer Amazon, whose global workforce is 60 percent male and where men hold 74 percent of the company's managerial positions, recently discontinued use of a recruiting algorithm after discovering gender bias.[9] The data that engineers used to create the algorithm were derived from the resumes submitted to Amazon over a 10-year period, which were predominantly from white males. The algorithm was taught to recognize word patterns in the resumes, rather than relevant skill sets, and these data were benchmarked against the company's predominantly male engineering department to determine an applicant's fit. As a result, the AI software penalized any resume that contained the word "women's" in the text and downgraded the resumes of women who attended women's colleges, resulting in gender bias.[10]

Amazon discontinued a recruiting algorithm after discovering that it led to gender bias in its hiring. (Credit: Brian Snyder/Reuters)

Bias in word associations

Princeton University researchers used off-the-shelf machine learning AI software to analyze and link 2.2 million words. They found that European names were perceived as more pleasant than those of African-Americans, and that the words "woman" and "girl" were more likely to be associated with the arts instead of science and math, which were most likely connected to males.[11] In analyzing these word-associations in the training data, the machine learning algorithm picked up on existing racial and gender biases shown by humans. If the learned associations of these algorithms were used as part of a search-engine ranking algorithm or to generate word suggestions as part of an auto-complete tool, it could have a cumulative effect of reinforcing racial and gender biases.

Bias in online ads

Latanya Sweeney, Harvard researcher and former chief technology officer at the Federal Trade Commission (FTC), found that online search queries for African-American names were more likely to return ads to that person from a service that renders arrest records, as compared to the ad results for white names.[12] Her research also found that the same differential treatment occurred in the micro-targeting of higher-interest credit cards and other financial products when the computer inferred that the subjects were African-Americans, despite having similar backgrounds to whites.[13] During a public presentation at a FTC hearing on big data,

Sweeney demonstrated how a web site, which marketed the centennial celebration of an all-black fraternity, received continuous ad suggestions for purchasing "arrest records" or accepting high-interest credit card offerings.[14]

Bias in facial recognition technology

MIT researcher Joy Buolamwini found that the algorithms powering three commercially available facial recognition software systems were failing to recognize darker-skinned complexions.[15] Generally, most facial recognition training data sets are estimated to be more than 75 percent male and more than 80 percent white. When the person in the photo was a white man, the software was accurate 99 percent of the time at identifying the person as male. According to Buolamwini's research, the product error rates for the three products were less than one percent overall, but increased to more than 20 percent in one product and 34 percent in the other two in the identification of darker-skinned women as female.[16] In response to Buolamwini's facial-analysis findings, both IBM and Microsoft committed to improving the accuracy of their recognition software for darker-skinned faces.

Bias in criminal justice algorithms

Acknowledging the possibility and causes of bias is the first step in any mitigation approach.

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, which is used by judges to predict whether defendants should be detained or released on bail pending trial, was found to be biased against African-Americans, according to a report from ProPublica.[17] The algorithm assigns a risk score to a defendant's likelihood to commit a future offense, relying on the voluminous data available on arrest records, defendant demographics, and other variables. Compared to whites who were equally likely to re-offend, African-Americans were more likely to be assigned a higher-risk score, resulting in longer periods of detention while awaiting trial.[18] Northpointe, the firm that sells the algorithm's outputs, offers evidence to refute such claims and argues that wrong metrics are being used to assess fairness in the product, a topic that we return to later in the paper.

While these examples of bias are not exhaustive, they suggest that these problems are empirical realities and not just theoretical concerns. They also illustrate how these outcomes emerge, and in some cases, without malicious intent by the creators or operators of the algorithm. Acknowledging the possibility and causes of bias is the first step in any mitigation approach. On this point, roundtable participant Ricardo Baeza-Yates from NTENT stated that "[companies] will continue to have a problem discussing algorithmic bias if they don't refer to the actual bias itself."

Causes of bias

Barocas and Selbst point out that bias can creep in during all phases of a project, "…whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors."[19] Roundtable participants focused especially on bias stemming from flaws in the data used to train the algorithms. "Flawed data is a big problem," stated roundtable participant Lucy Vasserman from Google, "…especially for the groups that businesses are working hard to protect." While there are many causes, we focus on two of them: *historical human biases* and *incomplete or unrepresentative data*.

Historical human biases

Historical human biases are shaped by pervasive and often deeply embedded prejudices against certain groups, which can lead to their reproduction and amplification in computer models. In the COMPAS algorithm, if African-Americans are more likely to be arrested and incarcerated in the U.S. due to historical racism, disparities in policing practices, or other inequalities within the criminal justice system, these realities will be reflected in the training data and used to make suggestions about whether a defendant should be detained. If historical biases are factored into the model, it will make the same kinds of wrong judgments that people do.

The Amazon recruitment algorithm revealed a similar trajectory when men were the benchmark for professional "fit," resulting in female applicants and their attributes being downgraded. These historical realities often find their way into the algorithm's development and execution, and they are exacerbated by the lack of diversity which exists within the computer and data science fields.[20]

Further, human biases can be reinforced and perpetuated without the user's knowledge. For example, African-Americans who are primarily the target for high-interest credit card options might find themselves clicking on this type of ad without realizing that they will continue to receive such predatory online suggestions. In this and other cases, the algorithm may never accumulate counter-factual ad suggestions (e.g., lower-interest credit options) that the consumer could be eligible for and prefer. Thus, it is important for algorithm designers and operators to watch for such potential negative feedback loops that cause an algorithm to become increasingly biased over time.

Incomplete or unrepresentative training data

Insufficient training data is another cause of algorithmic bias. If the data used to train the algorithm are more representative of some groups of people than others, the predictions from the model may also be systematically worse for unrepresented or under-representative groups. For example, in Buolamwini's facial-analysis experiments, the poor recognition of darker-skinned faces was largely due to their statistical under-representation in the training data. That is, the algorithm presumably picked up on certain facial features, such as the distance between the eyes, the shape of the eyebrows and variations in facial skin shades, as ways to

detect male and female faces. However, the facial features that were more representative in the training data were not as diverse and, therefore, less reliable to distinguish between complexions, even leading to a misidentification of darker-skinned females as males.

Turner Lee has argued that it is often the lack of diversity among the programmers designing the training sample which can lead to the under-representation of a particular group or specific physical attributes.[21] Buolamwini's findings were due to her rigor in testing, executing, and assessing a variety of proprietary facial-analysis software in different settings, correcting for the lack of diversity in their samples.

Conversely, algorithms with too much data, or an over-representation, can skew the decision toward a particular result. Researchers at Georgetown Law School found that an estimated 117 million American adults are in facial recognition networks used by law enforcement, and that African-Americans were more likely to be singled out primarily because of their *over-representation* in mug-shot databases.[22] Consequently, African-American faces had more opportunities to be falsely matched, which produced a biased effect.

Bias detection strategies

Understanding the various causes of biases is the first step in the adoption of effective algorithmic hygiene. But, how can operators of algorithms assess whether their results are, indeed, biased? Even when flaws in the training data are corrected, the results may still be problematic because *context* matters during the bias detection phase.

"Even when flaws in the training data are corrected, the results may still be problematic because context matters during the bias detection phase."

First, all detection approaches should begin with careful handling of the sensitive information of users, including data that identify a person's membership in a federally protected group (e.g., race, gender). In some cases, operators of algorithms may also worry about a person's membership in some other group if they are also susceptible to unfair outcomes. An examples of this could be college admission officers worrying about the algorithm's exclusion of applicants from lower-income or rural areas; these are individuals who may be not federally protected but do have susceptibility to certain harms (e.g., financial hardships).

In the former case, systemic bias against protected classes can lead to collective, *disparate impacts*, which may have a basis for legally cognizable harms, such as the denial of credit, online racial profiling, or massive surveillance.[23] In the latter case, the outputs of the algorithm may produce *unequal outcomes* or unequal error rates for different groups, but they may not violate legal prohibitions if there was no intent to discriminate.

These problematic outcomes should lead to further discussion and awareness of how algorithms work in the handling of sensitive information, and the trade-offs around fairness and accuracy in the models.

Algorithms and sensitive information

While it is intuitively appealing to think that an algorithm can be blind to sensitive attributes, this is not always the case.[24] Critics have pointed out that an algorithm may classify information based on online proxies for the sensitive attributes, yielding a bias against a group even without making decisions directly based on one's membership in that group. Barocas and Selbst define online proxies as "factors used in the scoring process of an algorithm which are mere stand-ins for protected groups, such as zip code as proxies for race, or height and weight as proxies for gender."[25] They argue that proxies often linked to algorithms can produce both errors and discriminatory outcomes, such as instances where a zip code is used to determine digital lending decisions or one's race triggers a disparate outcome.[26] Facebook's advertising platform contained proxies that allowed housing marketers to micro-target preferred renters and buyers by clicking off data points, including zip code preferences.[27] Thus, it is possible that an algorithm which is completely blind to a sensitive attribute could actually produce the same outcome as one that uses the attribute in a discriminatory manner.

"While it is intuitively appealing to think that an algorithm can be blind to sensitive attributes, this is not always the case."

For example, Amazon made a corporate decision to exclude certain neighborhoods from its same-day Prime delivery system. Their decision relied upon the following factors: whether a particular zip code had a sufficient number of Prime members, was near a warehouse, and had sufficient people willing to deliver to that zip code.[28] While these factors corresponded with the company's profitability model, they resulted in the exclusion of poor, predominantly African-American neighborhoods, transforming these data points into proxies for racial classification. The results, even when unintended, discriminated against racial and ethnic minorities who were not included.

Similarly, a job-matching algorithm may not receive the gender field as an input, but it may produce different match scores for two resumes that differ only in the substitution of the name "Mary" for "Mark" because the algorithm is trained to make these distinctions over time.

There are also arguments that blinding the algorithm to sensitive attributes can *cause* algorithmic bias in some situations. Corbett-Davies and Goel point out in their research on the COMPAS algorithm that even after controlling for "legitimate" risk factors, empirically women have been found to re-offend less often than men in many jurisdictions.[29] If an algorithm is forbidden from reporting a different risk assessment score for two

criminal defendants who differ only in their gender, judges may be less likely to release female defendants than male defendants with equal actual risks of committing another crime before trial. Thus, blinding the algorithm from any type of sensitive attribute may not solve bias.

While roundtable participants were not in agreement on the use of online proxies in modeling, they largely agreed that operators of algorithms must be more transparent in their handling of sensitive information, especially if the potential proxy could itself be a legal classificatory harm.[30] There was also discussion that the use of sensitive attributes as part of an algorithm could be a strategy for detecting and possibly curing intended and unintentional biases. Because currently doing so may be constrained by privacy regulations, such as the European Union's General Data Protection Rules (GDPR) or proposed U.S. federal privacy legislation, the argument could be made for the use of regulatory sandboxes and safe harbors to allow the use of sensitive information when detecting and mitigating biases, both of which will be introduced as part of our policy recommendations.

Detecting bias

When detecting bias, computer programmers normally examine the set of outputs that the algorithm produces to check for anomalous results. Comparing outcomes for different groups can be a useful first step. This could even be done through simulations. Roundtable participant Rich Caruana from Microsoft suggested that companies consider the simulation of predictions (both true and false) before applying them to real-life scenarios. "We almost need a secondary data collection process because sometimes the model will [emit] something quite different," he shared. For example, if a job-matching algorithm's average score for male applicants is higher than that for women, further investigation and simulations could be warranted.

However, the downside of these approaches is that not all unequal outcomes are unfair. Roundtable participant Solon Barocas from Cornell University summed this up when he stated, "Maybe we find out that we have a very accurate model, but it still produces disparate outcomes. This may be unfortunate, but is it fair?" An alternative to accounting for unequal outcomes may be to look at the equality of error rates, and whether there are more mistakes for one group of people than another. On this point, Isabel Kloumann of Facebook shared that "society has expectations. One of which is not incarcerating one minority group disproportionately [as a result of an algorithm]."

As shown in the debates around the COMPAS algorithm, even error rates are not a simple litmus test for biased algorithms. Northpointe, the company that developed the COMPAS algorithm, refutes claims of racial discrimination. They argue that among defendants assigned the same high risk score, African-American and white defendants have almost equal recidivism rates, so by that measure, there is no error in the algorithm's decision.[31] In their view, judges can consider their algorithm without any reference to race in bail and release decisions.

It is not possible, in general, to have equal error rates between groups for all the different error rates.[32] ProPublica focused on one error rate, while Northpointe honed in on another. Thus, some principles need to be established for which error rates should be equalized in which situations in order to be fair.

The COMPAS algorithm, which is used by judges to predict whether defendants should be detained or released on bail pending trial, has drawn scrutiny over claims of potential racial discrimination. (Credit: Stephen Lam/Reuters)

However, distinguishing between how the algorithm works with sensitive information and potential errors can be problematic for operators of algorithms, policymakers, and civil society groups.[33] "Companies would be losing a lot if we don't draw a distinction between the two," said Julie Brill from Microsoft. At the very least, there was agreement among roundtable participants that algorithms should not perpetuate historical inequities, and that more work needs to be done to address online discrimination.[34]

Fairness and accuracy trade-offs

Next, a discussion of trade-offs and ethics is needed. Here, the focus should be on evaluating both societal notions of "fairness" and possible social costs. In their research of the COMPAS algorithm, Corbett-Davies, Goel, Pierson, Feller, and Huq see "an inherent tension between minimizing violent crime and satisfying common notions of fairness."[35] They conclude that optimizing for public safety yields decisions that penalize defendants of color, while satisfying legal and societal fairness definitions, and may lead to more releases of

high-risk defendants, which would adversely affect public safety.[36] Moreover, the negative impacts on public safety might also disproportionately affect African-American and white neighborhoods, thus creating a fairness cost as well.

If the goal is to avoid reinforcing inequalities, what, then, should developers and operators of algorithms do to mitigate potential biases? We argue that developers of algorithms should first look for ways to reduce disparities between groups without sacrificing the overall performance of the model, especially whenever there appears to be a trade-off.

A handful of roundtable participants argued that opportunities exist for improving both fairness and accuracy in algorithms. For programmers, the investigation of apparent bugs in the software may reveal why the model was not maximizing for overall accuracy. The resolution of these bugs can then improve overall accuracy. Data sets, which may be under-representative of certain groups, may need additional training data to improve accuracy in the decision-making and reduce unfair results. Buolamwini's facial detection experiments are good examples of this type of approach to fairness and accuracy.

Roundtable participant Sarah Holland from Google pointed out the risk tolerance associated with these types of trade-offs when she shared that "[r]aising risk also involves raising equity issues." Thus, companies and other operators of algorithms should determine if the social costs of the trade-offs are justified, the stakeholders involved are amenable to a solution through algorithms, or if human decision-makers are needed to frame the solution.

Ethical frameworks matter

What is fundamentally behind these fairness and accuracy trade-offs should be discussions around ethical frameworks and potential guardrails for machine learning tasks and systems. There are several ongoing and recent international and U.S.-based efforts to develop ethical governance standards for the use of AI.[37] The 35-member Organization for Economic Cooperation and Development (OECD) is expected shortly to release its own guidelines for ethical AI.[38] The European Union recently released "Ethics Guidelines for Trustworthy AI," which delineates seven governance principles: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, nondiscrimination and fairness, (6) environmental and societal well-being, and (7) accountability.[39] The EU's ethical framework reflects a clear consensus that it is unethical to "unfairly discriminate." Within these guidelines, member states link diversity and nondiscrimination with principles of fairness, enabling inclusion and diversity throughout the entire AI system's lifecycle. Their principles interpret fairness through the lenses of equal access, inclusive design processes, and equal treatment.

Yet, even with these governmental efforts, it is still surprisingly difficult to define and measure fairness.[40] While it will not always be possible to satisfy all notions of fairness at the same time, companies and other operators of algorithms must be aware that there is no simple metric to measure fairness that a software engineer can apply, especially in the design of algorithms and the determination of the appropriate trade-offs

between accuracy and fairness. Fairness is a human, not a mathematical, determination, grounded in shared ethical beliefs. Thus, algorithmic decisions that may have a serious consequence for people will require human involvement.

For example, while the training data discrepancies in the COMPAS algorithm can be corrected, human interpretation of fairness still matters. For that reason, while an algorithm such as COMPAS may be a useful tool, it cannot substitute for the decision-making that lies within the discretion of the human arbiter.[41] We believe that subjecting the algorithm to rigorous testing can challenge the different definitions of fairness, a useful exercise among companies and other operators of algorithms.

"It's important for algorithm operators and developers to always be asking themselves: Will we leave some groups of people worse off as a result of the algorithm's design or its unintended consequences?"

In the decision to create and bring algorithms to market, the ethics of likely outcomes must be considered—especially in areas where governments, civil society, or policymakers see potential for harm, and where there is a risk of perpetuating existing biases or making protected groups more vulnerable to existing societal inequalities. That is why it's important for algorithm operators and developers to always be asking themselves: *Will we leave some groups of people worse off as a result of the algorithm's design or its unintended consequences?*

We suggest that this question is one among many that the creators and operators of algorithms should consider in the design, execution, and evaluation of algorithms, which are described in the following mitigation proposals. Our first proposal addresses the updating of U.S. nondiscrimination laws to apply to the digital space.

Mitigation proposals

Nondiscrimination and other civil rights laws should be updated to interpret and redress online disparate impacts

To develop trust from policymakers, computer programmers, businesses, and other operators of algorithms must abide by U.S. laws and statutes that currently forbid discrimination in public spaces. Historically, nondiscrimination laws and statutes unambiguously define the thresholds and parameters for the disparate treatment of protected classes. The 1964 Civil Rights Act "forbade discrimination on the basis of sex as well as race in hiring, promoting, and firing." The 1968 Fair Housing Act prohibits discrimination in the sale, rental, and financing of dwellings, and in other housing-related transactions to federally protected classes. Enacted in 1974, the Equal Credit Opportunity Act stops any creditor from discriminating against any applicant from any

type of credit transaction based on protected characteristics. While these laws do not necessarily mitigate and resolve other implicit or unconscious biases that can be baked into algorithms, companies and other operators should guard against violating these statutory guardrails in the design of algorithms, as well as mitigating their implicit concern to prevent past discrimination from continuing.

Roundtable participant Wendy Anderson from the Office of Congresswoman Val Demings stated, "[T]ypically, legislators only hear when something bad happens. We need to find a way to protect those who need it without stifling innovation." Congress can clarify how these nondiscrimination laws apply to the types of grievances recently found in the digital space, since most of these laws were written before the advent of the internet.[42] Such legislative action can provide clearer guardrails that are triggered when algorithms are contributing to legally recognizable harms. Moreover, when creators and operators of algorithms understand that these may be more or less non-negotiable factors, the technical design will be more thoughtful in moving away from models that may trigger and exacerbate explicit discrimination, such as design frames that exclude rather than include certain inputs or are not checked for bias.[43]

Operators of algorithms must develop a bias impact statement

Once the idea for an algorithm has been vetted against nondiscrimination laws, we suggest that operators of algorithms develop a bias impact statement, which we offer as a template of questions that can be flexibly applied to guide them through the design, implementation, and monitoring phases.

As a self-regulatory practice, the *bias impact statement* can help probe and avert any potential biases that are baked into or are resultant from the algorithmic decision. As a best practice, operators of algorithms should brainstorm a core set of initial assumptions about the algorithm's purpose prior to its development and execution. We propose that operators apply the bias impact statement to assess the algorithm's purpose, process and production, where appropriate. Roundtable participants also suggested the importance of establishing a cross-functional and interdisciplinary team to create and implement the bias impact statement.

- **New York University's AI Now Institute**

New York University's AI Now Institute has already introduced a model framework for governmental entities to use to create algorithmic impact assessments (AIAs), which evaluate the potential detrimental effects of an algorithm in the same manner as environmental, privacy, data, or human rights impact statements.[44] While there may be differences in implementation given the type of predictive model, the AIA encompasses multiple rounds of review from internal, external, and public audiences. First, it assumes that after this review, a company will develop a list of potential harms or biases in their self-assessment, with the assistance of more technical outside experts. Second, if bias appears to have occurred, the AIA pushes for notice to be given to impacted populations and a comment period opened for response. And third, the AIA process looks to federal and other entities to support users' right to challenge algorithmic decisions that feel unfair.

While the AIA process supports a substantive feedback loop, what may be missing is both the required forethought leading up to the decision and the oversight of the algorithm's provisions. Moreover, our proposed bias impact statement starts with a framework that identifies *which* automated decisions should be subjected to such scrutiny, operator incentives, and stakeholder engagement.

- **Which automated decisions?**

In the case of determining which automated decisions require such vetting, operators of algorithms should start with questions about whether there will be a possible negative or unintended outcome resulting from the algorithm, for whom, and the severity of consequences for members of the affected group if not detected and mitigated. Reviewing established legal protections around fair housing, employment, credit, criminal justice, and health care should serve as a starting point for determining which decisions need to be viewed with special caution in designing and testing any algorithm used to predict outcomes or make important eligibility decisions about access to a benefit. This is particularly true considering the legal prescriptions against using data that has a likelihood of disparate impact on a protected class or other established harms. Thus, we suggest that operators should be constantly questioning the potential legal, social, and economic effects and potential liabilities associated with that choice when determining which decisions should be automated and how to automate them with minimal risks.

- **What are the user incentives?**

Incentives should also drive organizations to proactively address algorithmic bias. Conversely, operators who create and deploy algorithms that generate fairer outcomes should also be recognized by policymakers and consumers who will trust them more for their practices. When companies exercise effective algorithmic hygiene before, during, and after introducing algorithmic decision-making, they should be rewarded and potentially given a public-facing acknowledgement for best practices.

- **How are stakeholders being engaged?**

Finally, the last element encapsulated in a bias impact statement should involve the engagement of stakeholders who could help computer programmers in the selection of inputs and outputs of certain automated decisions. "Tech succeeds when users understand the product better than its designers," said Rich Caruana from Microsoft. Getting users engaged early and throughout the process will prompt improvements to the algorithms, which ultimately leads to improved user experiences.

Stakeholder responsibilities can also extend to civil society organizations who can add value in the conversation on the algorithm's design. "Companies [should] engage civil society," shared Miranda Bogen from Upturn. "Otherwise, they will go to the press and regulators with their complaints." A possible solution for operators of algorithms could be the development of an advisory council of civil society organizations that, working alongside companies, may be helpful in defining the scope of the procedure and predicting biases based on their ground-level experiences.

- **The template for the bias impact statement**

These three foundational elements for a bias impact statement are reflected in a discrete set of questions that operators should answer during the design phase to filter out potential biases (Table 1). As a self-regulatory framework, computer programmers and other operators of algorithms can construct this type of tool prior to the model's design and execution.

Table 1. Design questions template for bias impact statement

What will the automated decision do?
Who is the audience for the algorithm and who will be most affected by it?
Do we have training data to make the correct predictions about the decision?
Is the training data sufficiently diverse and reliable? What is the data lifecycle of the algorithm?
Which groups are we worried about when it comes to training data errors, disparate treatment, and impact?
How will potential bias be detected?
How and when will the algorithm be tested? Who will be the targets for testing?
What will be the threshold for measuring and correcting for bias in the algorithm, especially as it relates to protected groups?
What are the operator incentives?
What will we gain in the development of the algorithm?
What are the potential bad outcomes and how will we know?
How open (e.g., in code or intent) will we make the design process of the algorithm to internal partners, clients, and customers?
What intervention will be taken if we predict that there might be bad outcomes associated with the development or deployment of the algorithm?
How are other stakeholders being engaged?
What's the feedback loop for the algorithm for developers, internal partners and customers?
Is there a role for civil society organizations in the design of the algorithm?
Has diversity been considered in the design and execution?
Will the algorithm have implications for cultural groups and play out differently in cultural contexts?
Is the design team representative enough to capture these nuances and predict the application of the algorithm within different cultural contexts? If not, what steps are being taken to make these scenarios more salient and understandable to designers?
Given the algorithm's purpose, is the training data sufficiently diverse?
Are there statutory guardrails that companies should be reviewing to ensure that the algorithm is both legal and ethical?

Diversity-in-design

Operators of algorithms should also consider the role of diversity within their work teams, training data, and the level of cultural sensitivity within their decision-making processes. Employing diversity in the design of algorithms upfront will trigger and potentially avoid harmful discriminatory effects on certain protected groups, especially racial and ethnic minorities. While the immediate consequences of biases in these areas may be small, the sheer quantity of digital interactions and inferences can amount to a new form of systemic bias. Therefore, the operators of algorithms should not discount the possibility or prevalence of bias and should seek to have a diverse workforce developing the algorithm, integrate inclusive spaces within their products, or employ "diversity-in-design," where deliberate and transparent actions will be taken to ensure that cultural biases and stereotypes are addressed upfront and appropriately. Adding inclusivity into the algorithm's design can potentially vet the cultural inclusivity and sensitivity of the algorithms for various groups and help companies avoid what can be litigious and embarrassing algorithmic outcomes.

The bias impact statement should not be an exhaustive tool. For algorithms with more at stake, ongoing review of their execution should be factored into the process. The goal here is to monitor for disparate impacts resulting from the model that border on unethical, unfair, and unjust decision-making. When the process of identifying and forecasting the purpose of the algorithm is achieved, a robust feedback loop will aid in the detection of bias, which leads to the next recommendation promoting regular audits.

Other self-regulatory best practices

Operators of algorithms should regularly audit for bias

The formal and regular auditing of algorithms to check for bias is another best practice for detecting and mitigating bias. On the importance of these audits, roundtable participant Jon Kleinberg from Cornell University shared that "[a]n algorithm has no choice but to be premeditated." Audits prompt the review of both input data and output decisions, and when done by a third-party evaluator, they can provide insight into the algorithm's behavior. While some audits may require technical expertise, this may not always be the case. Facial recognition software that misidentifies persons of color more than whites is an instance where a stakeholder or user can spot biased outcomes, without knowing anything about how the algorithm makes decisions. "We should expect computers to have an audit trail," shared roundtable participant Miranda Bogen from Upturn. Developing a regular and thorough audit of the data collected for the algorithmic operation, along with responses from developers, civil society, and others impacted by the algorithm, will better detect and possibly deter biases.

"Developing a regular and thorough audit of the data collected for the algorithmic operation, along with responses from developers, civil society, and others impacted by the algorithm, will better detect and possibly deter biases."

The experience of government officials in Allegheny County reflects the importance of third-party auditing. In 2016, the Department of Human Services launched a decision support tool, the Allegheny Family Screening Tool (AFST), to generate a score for which children are most likely to be removed from their homes within two years, or to be re-referred to the county's child welfare office due to suspected abuse. The county took ownership of its use of the tool, worked collaboratively with the developer, and commissioned an independent evaluation of its direct and indirect effects on the maltreatment screening process, including decision accuracy, workload, and consistency. County officials also sought additional independent research from experts to determine if the software was discriminating against certain groups. In 2017, the findings did identify some statistical imbalances, with error rates higher across racial and ethnic groups. White children who were scored at the highest-risk of maltreatment were less likely to be removed from their homes compared to African-American children with similar risk scores.[45] The county responded to these findings as part of the rebuild of the tool, with version two implemented in November 2018.[46]

Facebook recently completed a civil rights audit to determine its handling of issues and individuals from protected groups.[47] After the reveal of how the platform was handling a variety of issues, including voter suppression, content moderation, privacy, and diversity, the company has committed to an updated audit around its internal infrastructure to handle civil rights grievances and address diversity in its products' designs by default. Recent actions by Facebook to ban white nationalist content or address disinformation campaigns are some of the results of these efforts.[48]

Operators of algorithms must rely upon cross-functional work teams and expertise

Roundtable participants largely acknowledged the notion that organizations should employ cross-functional teams. But movement in this direction can be difficult in already-siloed organizations, despite the technical, societal, and possibly legal implications associated with the algorithm's design and execution. Not all decisions will necessitate this type of cross-team review, but when these decisions carry risks of real harm, they should be employed. In the mitigation of bias and the management of the risks associated with the algorithm, collaborative work teams can compensate for the blind-spots often missed in smaller, segmented conversations and reviews. Bringing together experts from various departments, disciplines, and sectors will help facilitate accountability standards and strategies for mitigating online biases, including from engineering, legal, marketing, strategy, and communications.

Cross-functional work teams–whether internally driven or populated by external experts–can attempt to identify bias before and during the model's rollout. Further, partnerships between the private sector, academics, and civil society organizations can also facilitate greater transparency in AI's application to a variety of scenarios, particularly those that impact protected classes or are disseminated in the public interest. Kate Crawford, AI researcher and founder of the AI Now Partnership, suggested that "closed loops are not open for algorithmic auditing, for review, or for public debate" because they generally exacerbate the problems that they are trying to solve.[49] Further on this point, roundtable participant Natasha Duarte from the Center for Democracy and Technology spoke to Allegheny's challenge when she shared, "[C]ompanies should be more forthcoming with describing the limits of their tech, and government should know what questions to ask in their assessments," which speaks to the importance of more collaboration in this area.

Increase human involvement in the design and monitoring of algorithms

Even with all the precautionary measures listed above, there is still some risk that algorithms will make biased decisions. People will continue to play a role in identifying and correcting biased outcomes long after an algorithm is developed, tested, and launched. While more data can inform automated decision-making, this process should complement rather than fully replace human judgement. Roundtable participant Alex Peysakhovich from Facebook shared, "[W]e don't need to eliminate human moderators. We need to hire more and get them to focus on edge cases." Such sentiment is growing increasingly important in this field as the comparative advantages of humans and algorithms become more distinguishable and the use of both improves the outcomes for online users.

People will continue to play a role in identifying and correcting biased outcomes long after an algorithm is developed, tested, and launched. (Credit: Gabrielle Lurie/Reuters)

However, privacy implications will arise when more humans are engaged in algorithm management, particularly if more sensitive information is involved in the model's creation or in testing the algorithm's predictions for bias. The timing of the roundtables, which also transpired around the adoption of the EU's GDPR, spoke to the need for increased consumer privacy principles where users are empowered over what data they want to share with companies. As the U.S. currently debates the need for federal privacy legislation, access to and use of personal data may become even more difficult, potentially leaving algorithmic models prone to more bias. Because the values of creators and users of algorithms shift over time, humans must arbitrate conflicts between outcomes and stated goals. In addition to periodical audits, human involvement provides continuous feedback on the performance of bias mitigation efforts.

Other public policy recommendations

As indicated throughout the paper, policymakers play a critical role in identifying and mitigating biases, while ensuring that the technologies continue to make positive economic and societal benefits.

Congress should implement regulatory sandboxes and safe harbors to curb online biases

Regulatory sandboxes are perceived as one strategy for the creation of temporary reprieves from regulation to allow the technology and rules surrounding its use to evolve together. These policies could apply to algorithmic bias and other areas where the technology in question has no analog covered by existing regulations. Rather

than broaden the scope of existing regulations or create rules in anticipation of potential harms, a sandbox allows for innovation both in technology and its regulation. Even in a highly regulated industry, the creation of sandboxes where innovations can be tested alongside with lighter touch regulations can yield benefits.

"Rather than broaden the scope of existing regulations or create rules in anticipation of potential harms, a sandbox allows for innovation both in technology and its regulation."

For example, companies within the financial sector that are leveraging technology, or fintech, have shown how regulatory sandboxes can spur innovation in the development of new products and services.[50] These companies make extensive use of algorithms for everything from spotting fraud to deciding to extend credit. Some of these activities mirror those of regular banks, and those would still fall under existing rules, but new ways of approaching tasks would be allowed within the sandbox.[51] Because sandboxes give innovators greater leeway in developing new products and services, they will require active oversight until technology and regulations mature. The U.S. Treasury recently reported not only on the benefits that countries that have adopted fintech regulatory sandboxes have realized, but recommended that the U.S. adopt fintech sandboxes to spur innovation.[52] Given the broad usefulness of algorithms to spur innovation in various regulated industries, participants in the roundtables considered the potential usefulness of extending regulatory sandboxes to other areas where algorithms can help to spur innovations.

Regulatory safe harbors could also be employed, where a regulator could specify which activities do not violate existing regulations.[53] This approach has the advantage of increasing regulatory certainty for algorithm developers and operators. For example, Section 230 of the Communications Decency Act removed liability from websites for the actions of their users, a provision widely credited with the growth of internet companies like Facebook and Google. The exemption later narrowed to exclude sex trafficking with the passage of the Stop Enabling Online Sex Trafficking Act and Fight Online Sex Trafficking Act. Applying a similar approach to algorithms could exempt their operators from liabilities in certain contexts while still upholding protections in others where harms are easier to identify. In line with the previous discussion on the use of certain protected attributes, safe harbors could be considered in instances where the collection of sensitive personal information is used for the specific purposes of bias detection and mitigation.

Consumers need better algorithmic literacy

Widespread algorithmic literacy is crucial for mitigating bias. Given the increased use of algorithms in many aspects of daily life, all potential subjects of automated decisions would benefit from knowledge of how these systems function. Just as computer literacy is now considered a vital skill in the modern economy, understanding how algorithms use their data may soon become necessary.

The subjects of automated decisions deserve to know when bias negatively affects them, and how to respond when it occurs. Feedback from users can share and anticipate areas where bias can manifest in existing and future algorithms. Over time, the creators of algorithms may actively solicit feedback from a wide range of data subjects and then take steps to educate the public on how algorithms work to aid in this effort. Public agencies that regulate bias can also work to raise algorithmic literacy as part of their missions. In both the public and private sector, those that stand to lose the most from biased decision-making can also play an active role in spotting it.

Conclusion

In December 2018, President Trump signed the First Step Act, new criminal justice legislation that encourages the usage of algorithms nationwide.[54] In particular, the system would use an algorithm to initially determine who can redeem earned-time credits—reductions in sentence for completion of educational, vocational, or rehabilitative programs—excluding inmates deemed higher risk. There is a likelihood that these algorithms will perpetuate racial and class disparities, which are already embedded in the criminal justice system. As a result, African-Americans and poor people in general will be more likely to serve longer prison sentences.

"When algorithms are responsibly designed, they may avoid the unfortunate consequences of amplified systemic discrimination and unethical applications."

As outlined in the paper, these types of algorithms should be concerning if there is not a process in place that incorporates technical diligence, fairness, and equity from design to execution. That is, when algorithms are responsibly designed, they may avoid the unfortunate consequences of amplified systemic discrimination and unethical applications.

Some decisions will be best served by algorithms and other AI tools, while others may need thoughtful consideration before computer models are designed. Further, testing and review of certain algorithms will also identify, and, at best, mitigate discriminatory outcomes. For operators of algorithms seeking to reduce the risk and complications of bad outcomes for consumers, the promotion and use of the mitigation proposals can create a pathway toward algorithmic fairness, even if equity is never fully realized.

Amazon, Facebook, Google, IBM, and Microsoft provide general, unrestricted support to The Brookings Institution. Paul Resnick is also a consultant to Facebook, but this work is independent and his views expressed here are his own. The findings, interpretations, and conclusions posted in this piece are not influenced by any donation. Brookings recognizes that the value it provides is in its absolute commitment to quality, independence, and impact. Activities supported by its donors reflect this commitment.

Appendix: List of Roundtable Participants

Participant	Organization
Wendy Anderson	Office of Congresswoman Val Demings
Norberto Andrade	Facebook
Solon Barocas	Cornell University
Genie Barton	Privacy Genie
Ricardo Baeza-Yates	NTENT
Miranda Bogen	Upturn
John Brescia	Better Business Bureau
Julie Brill	Microsoft
Rich Caruana	Microsoft Research
Eli Cohen	Brookings Institution
Anupam Datta	Carnegie Mellon
Deven Desai	Georgia Tech
Natasha Duarte	Center for Democracy and Technology
Nadia Fawaz	LinkedIn
Laura Fragomeni	Walmart Global eCommerce
Sharad Goel	Stanford University
Scott Golder	Cornell University
Aaron Halfaker	Wikimedia
Sarah Holland	Google
Jack Karsten	Brookings Institution
Krishnaram Kenthapadi	LinkedIn and Stanford University
Jon Kleinberg	Cornell University
Isabel Kloumann	Facebook
Jake Metcalf	Ethical Resolve
Alex Peysakhovich	Facebook
Paul Resnick	University of Michigan
William Rinehart	American Action Forum

Participant	Organization
Alex Rosenblat	Data and Society
Jake Schneider	Brookings Institution
Jasjeet Sekhon	University of California-Berkeley
Rob Sherman	Facebook
JoAnn Stonier	Mastercard Worldwide
Nicol Turner Lee	Brookings Institution
Lucy Vasserman	Jigsaw's Conversation AI Project / Google
Suresh Venkatasubramanian	University of Utah
John Verdi	Future of Privacy Forum
Heather West	Mozilla
Jason Yosinki	Uber
Jinyan Zang	Harvard University
Leila Zia	Wikimedia Foundation

References

Angwin, Julia, and Terry Parris Jr. "Facebook Lets Advertisers Exclude Users by Race." Text/html. ProPublica, October 28, 2016. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race.

Angwin, Julia, Jeff Larson, Surya Mattu, and Laura Kirchner. "Machine Bias." ProPublica, May 23, 2016. Available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last accessed April 19, 2019).

Barocas, Solon, and Andrew D. Selbst, "Big Data's Disparate Impact," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2016. Available at https://papers.ssrn.com/abstract=2477899.

Blass, Andrea, and Yuri Gurevich. Algorithms: A Quest for Absolute Definitions. Bulletin of European Association for Theoretical Computer Science 81, 2003. https://www.microsoft.com/en-us/research/wp-content/uploads/2017/01/164.pdf (last accessed April 12, 2019).

Brennan, Tim, William Dieterich, and Beate Ehret. "Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System." Criminal Justice and Behavior 36 (2009): 21–40.

Chessell, Mandy. "Ethics for Big Data and Analytics." IBM, n.d. Available at https://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf (last accessed April 19, 2019).

Chodosh, Sara. “Courts use algorithms to help determine sentencing, but random people get the same results.” Popular Science, January 18, 2018. Available at https://www.popsci.com/recidivism-algorithm-random-bias (last accessed October 15, 2018).

Corbett-Davies, Sam, Emma Peirson, Avi Feller, and Sharad Goel. “A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It’s Actually Not That Clear.” Washington Post (blog), October 17, 2016. Available at https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ (last accessed April 19. 2019).

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. “Algorithmic Decision Making and the Cost of Fairness.” ArXiv:1701.08230 [Cs, Stat], January 27, 2017. https://doi.org/10.1145/3097983.309809.

Courtland, Rachel. “Bias Detectives: The Researchers Striving to Make Algorithms Fair,” Nature 558, no. 7710 (June 2018): 357–60. Available at https://doi.org/10.1038/d41586-018-05469-3 (last accessed April 19, 2019).

DeAngelius, Stephen F. “Artificial intelligence: How algorithms make systems smart,” Wired Magazine, September 2014. Available at https://www.wired.com//insights/2014/09/artificial-intelligence-algorithms-2/ (last accessed April 12, 2019).

Elejalde-Ruiz, Alexia. “The end of the resume? Hiring is in the midst of technological revolution with algorithms, chatbots.” Chicago Tribune (July 19, 2018). Available at http://www.chicagotribune.com/business/ct-biz-artificial-intelligence-hiring-20180719-story.html.

Eubanks, Virginia. “A Child Abuse Prediction Model Fails Poor Families,” Wired, January 15, 2018. Available at https://www.wired.com/story/excerpt-from-automating-inequality/ (last accessed April 19, 2019).

FTC Hearing #7: The Competition and Consumer Protection Issues of Algorithms, Artificial Intelligence, and Predictive Analytics, § Federal Trade Commission (2018). https://www.ftc.gov/system/files/documents/public_events/1418693/ftc_hearings_session_7_transcript_day_2_11-14-18.pdf.

Garbade, Michael J. “Clearing the Confusion: AI vs. Machine Learning vs. Deep Learning Differences,” Towards Data Science, September 14, 2018. Available at https://towardsdatascience//clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb (last accessed April 12, 2019).

Griggs v. Duke Power Company, Oyez. Available at https//www.oyez.org/cases/1970/124 (last accessed October 1, 2018.

Guerin, Lisa. “Disparate Impact Discrimination.” www.nolo.com. Available at https://www.nolo.com/legal-encyclopedia/disparate-impact-discrimination.htm (last accessed April 24, 2019).

Hadhazy, Adam. “Biased Bots: Artificial-Intelligence Systems Echo Human Prejudices.” Princeton University, April 18, 2017. Available at https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices (last accessed April 20, 2019).

Hamilton, Isobel Asher. “Why It’s Totally Unsurprising That Amazon’s Recruitment AI Was Biased against Women.” Business Insider, October 13, 2018. Available at https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10 (last accessed April 20, 2019).

Hardesty, Larry. “Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems.” MIT News, February 11, 2018. Available at http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (last accessed April 19, 2019).

High-level Expert Group on Artificial Intelligence. “Ethics Guidelines for Trustworthy AI (Draft).” The European Commission, December 18, 2018.

Ingold, David, and Spencer Soper. “Amazon Doesn’t Consider the Race of Its Customers. Should It?” Bloomberg.com, April 21, 2016. http://www.bloomberg.com/graphics/2016-amazon-same-day/.

Kearns, Michael. “Data Intimacy, Machine Learning and Consumer Privacy.” University of Pennsylvania Law School, May 2018. Available at https://www.law.upenn.edu/live/files/7952-kearns-finalpdf (last accessed April 12, 2019).

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan, “Inherent Trade-Offs in the Fair Determination of Risk Scores.”In Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017. Available at https://arxiv.org/pdf/1609.05807.pdf (last accessed April 19, 2019).

Larson, Jeff, Surya Mattu, and Julia Angwin. “Unintended Consequences of Geographic Targeting.” Technology Science, September 1, 2015. Available at https://techscience.org/a/2015090103/ (last accessed April 19, 2019).

Locklear, Mallory. “Facebook Releases an Update on Its Civil Rights Audit.” Engadget (blog), December 18, 2018. Available at https://www.engadget.com/2018/12/18/facebook-update-civil-rights-audit/ (last accessed April 19, 2019).

Lopez, German. “The First Step Act, Congress’s Criminal Justice Reform Bill, Explained.” Vox, December 3, 2018. Available at https://www.vox.com/future-perfect/2018/12/3/18122392/first-step-act-criminal-justice-reform-bill-congress (last accessed April 16, 2019).

Mnuchin, Steven T., and Craig S. Phillips. “A Financial System That Creates Economic Opportunities – Nonbank Financials, Fintech, and Innovation.” Washington, D.C.: U.S. Department of the Treasury, July 2018.Available at https://home.treasury.gov/sites/default/files/2018-08/A-Financial-System-that-Creates-Economic-Opportunities—Nonbank-Financials-Fintech-and-Innovation_0.pdf (last accessed April 19, 2019).

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. “Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.” New York: AI Now, April 2018.

Romei, Andrea, and Salvatore Ruggieri. “Discrimination Data Analysis: A Multi-Disciplinary Bibliography.” In Discrimination and Privacy in the Information Society, edited by Bart Custers, T Calders, B Schermer, and T Zarsky, 109–35. Studies in Applied Philosophy, Epistemology and Rational Ethics. Springer, Berlin, Heidelberg,

2013. Available at https://doi.org/10.1007/978-3-642-30487-3_6 (last accessed April 19, 2019).

Schatz, Brian. AI in Government Act of 2018, Pub. L. No. S.B. 3502 (2018). https://www.congress.gov/bill/115th-congress/senate-bill/3502.

Spielkamp, Matthias. "We Need to Shine More Light on Algorithms so They Can Help Reduce Bias, Not Perpetuate It." MIT Technology Review. Accessed September 20, 2018. Available at https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/ (last accessed April 19, 2019).

Stack, Liam. "Facebook Announces New Policy to Ban White Nationalist Content." The New York Times, March 28, 2019, sec. Business. Available at https://www.nytimes.com/2019/03/27/business/facebook-white-nationalist-supremacist.html (last accessed April 19, 2019).

Sweeney, Latanya, and Jinyan Zang. "How appropriate might big data analytics decisions be when placing ads?" Powerpoint presentation presented at the Big Data: A tool for inclusion or exclusion, Federal Trade Commission conference, Washington, DC. September 15, 2014. Available at https://www.ftc.gov/systems/files/documents/public_events/313371/bigdata-slides-sweeneyzang-9_15_14.pdf (last accessed April 12, 2019).

Sweeney, Latanya. "Discrimination in online ad delivery." Rochester, NY: Social Science Research Network, January 28, 2013. Available at https://papers.ssrn.com/abstract=2208240 (last accessed April 12, 2019).

Sydell, Laura. "It Ain't Me, Babe: Researchers Find Flaws In Police Facial Recognition Technology." NPR.org, October 25, 2016. Available at https://www.npr.org/sections/alltechconsidered/2016/10/25/499176469/it-aint-me-babe-researchers-find-flaws-in-police-facial-recognition (last accessed April 19, 2019).

"The Global Data Ethics Project." Data for Democracy, n.d. https://www.datafordemocracy.org/project/global-data-ethics-project (last accessed April 19, 2019).

Tobin, Ariana. "HUD sues Facebook over housing discrimination and says the company's algorithms have made the problem worse." ProPublica (March 28, 2019). Available at https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms (last accessed April 29, 2019).

Turner Lee, Nicol. "Inclusion in Tech: How Diversity Benefits All Americans," § Subcommittee on Consumer Protection and Commerce, United States House Committee on Energy and Commerce (2019). Also available on Brookings web site, https://www.brookings.edu/testimonies/inclusion-in-tech-how-diversity-benefits-all-americans/ (last accessed April 29, 2019).

Turner Lee, Nicol. Detecting racial bias in algorithms and machine learning. Journal of Information, Communication and Ethics in Society 2018, Vol. 16 Issue 3, pp. 252-260. Available at https://doi.org/10.1108/JICES-06-2018-0056/ (last accessed April 29, 2019).

"Understanding bias in algorithmic design," Impact.Engineered, September 5, 2017. Available at https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e (last accessed April 12, 2019).

Vincent, James. "Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women." The Verge, October 10, 2018. Available at https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report (last accessed April 20, 2019).

Zafar, Muhammad Bilal, Isabel Valera Martinez, Manuel Gomez Rodriguez, and Krishna Gummadi. "Fairness Constraints: A Mechanism for Fair Classification." In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, FL, 2017.

Zarsky, Tal. "Understanding Discrimination in the Scored Society." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 15, 2015. https://papers.ssrn.com/abstract=2550248.

Report Produced by **Center for Technology Innovation**

Footnotes

1. 1 Nicol Turner Lee, Fellow, Center for Technology Innovation, Brookings Institution; Paul Resnick, Michael D. Cohen Collegiate Professor of Information, Associate Dean for Research and Faculty Affairs, Professor of Information and Interim Director of Health Informatics, School of Information at the University of Michigan; Genie Barton, President, Institute for Marketplace Trust, Better Business Bureau and Member, Research Advisory Board, International Association of Privacy Professionals. The authors also acknowledge the input from the current leadership of the Better Business Bureau's Institute for Marketplace Trust and Jinyan Zang, Harvard University.
2. 2 The concepts of AI, algorithms and machine learning are often conflated and used interchangeably. In this paper, we will follow generally understood definitions of these terms as set out in publications for the general reader. See, e.g., Stephen F. DeAngelius. "Artificial intelligence: How algorithms make systems smart," Wired Magazine, September 2014. Available at https://www.wired.com//insights/2014/09/artificial-intelligence-algorithms-2/ (last accessed April 12, 2019). See also, Michael J. Garbade. "Clearing the Confusion: AI vs. Machine Learning vs. Deep Learning Differences," Towards Data Science, September 14, 2018. Available at https://towardsdatascience//clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb (last accessed April 12, 2019).
3. 3 Andrea Blass and Yuri Gurevich. Algorithms: A Quest for Absolute Definitions. Bulletin of European Association for Theoretical Computer Science 81, 2003. https://www.microsoft.com/en-us/research/wp-content/uploads/2017/01/164.pdf (last accessed April 12, 2019).
4. 4 Kearns, Michael. "Data Intimacy, Machine Learning and Consumer Privacy." University of Pennsylvania Law School, May 2018. Available at https://www.law.upenn.edu/live/files/7952-kearns-finalpdf (last accessed April 12, 2019).
5. 5 Technically, this describes what is called "supervised machine learning."
6. 6 Chodosh, Sara. "Courts use algorithms to help determine sentencing, but random people get the same results." Popular Science, January 18, 2018. Available at https://www.popsci.com/recidivism-algorithm-random-bias (last accessed October 15, 2018).
7. 7 Blog. "Understanding bias in algorithmic design," Impact.Engineered, September 5, 2017. Available at https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e (last accessed April 12, 2019). This definition is intended to include the concepts of disparate treatment and disparate impact, but the legal definitions were not designed with AI in mind. For example, the demonstration of disparate treatment does not describe the ways in which an algorithm can learn to treat similarly situated groups differently, as will be discussed later in the paper.
8. 8 The recommendations offered in the paper are those of the authors and do not represent the views or a consensus of views among roundtable participants.
9. 9 Hamilton, Isobel Asher. "Why It's Totally Unsurprising That Amazon's Recruitment AI Was Biased against Women." Business Insider, October 13, 2018. Available at https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10 (last accessed April 20, 2019).
10. 10 Vincent, James. "Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women." The Verge, October 10, 2018. Available at https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report (last accessed April 20, 2019). Although Amazon scrubbed the data of the particular references that appeared to discriminate against female candidates, there was no guarantee that the algorithm could not find other ways to sort and rank male candidates higher so it was scrapped by the company.
11. 11 Hadhazy, Adam. "Biased Bots: Artificial-Intelligence Systems Echo Human Prejudices." Princeton University, April 18, 2017. Available at https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices (last accessed April 20, 2019).

12. 12 Sweeney, Latanya. "Discrimination in online ad delivery." Rochester, NY: Social Science Research Network, January 28, 2013. Available at https://papers.ssrn.com/abstract=2208240 (last accessed April 12, 2019).
13. 13 Sweeney, Latanya and Jinyan Zang. "How appropriate might big data analytics decisions be when placing ads?" Powerpoint presentation presented at the Big Data: A tool for inclusion or exclusion, Federal Trade Commission conference, Washington, DC. September 15, 2014. Available at https://www.ftc.gov/systems/files/documents/public_events/313371/bigdata-slides-sweeneyzang-9_15_14.pdf (last accessed April 12, 2019).
14. 14 "FTC Hearing #7: The Competition and Consumer Protection Issues of Algorithms, Artificial Intelligence, and Predictive Analytics," § Federal Trade Commission (2018),
15. 15 Hardesty, Larry. "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems." MIT News, February 11, 2018. Available at http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (last accessed April 19, 2019). These companies were selected because they provided gender classification features in their software and the code was publicly available for testing.
16. 16 Ibid.
17. 17 COMPAS is a risk-and needs-assessment tool originally designed by Northpointe, Inc., to assist state corrections officials in making placement, management, and treatment decisions for offenders. Angwin, Julia, Jeff Larson, Surya Mattu, and Laura Kirchner. "Machine Bias." ProPublica, May 23, 2016. Available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last accessed April 19, 2019).
See also, Brennan, Tim, William Dieterich, and Beate Ehret. "Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System." Criminal Justice and Behavior 36 (2009): 21–40.
18. 18 Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic Decision Making and the Cost of Fairness." ArXiv:1701.08230 [Cs, Stat], January 27, 2017. https://doi.org/10.1145/3097983.309809.
19. 19 Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2016), https://papers.ssrn.com/abstract=2477899.
20. 20 Turner Lee, Nicol. "Inclusion in Tech: How Diversity Benefits All Americans," § Subcommittee on Consumer Protection and Commerce, United States House Committee on Energy and Commerce (2019). Also available on Brookings web site, https://www.brookings.edu/testimonies/inclusion-in-tech-how-diversity-benefits-all-americans/ (last accessed April 29, 2019).
21. 21 Ibid. See also, Turner Lee, Nicol. Detecting racial bias in algorithms and machine learning. Journal of Information, Communication and Ethics in Society 2018, Vol. 16 Issue 3, pp. 252-260. Available at https://doi.org/10.1108/JICES-06-2018-0056/ (last accessed April 29, 2019).
22. 22 Sydell, Laura. "It Ain't Me, Babe: Researchers Find Flaws In Police Facial Recognition Technology." NPR.org, October 25, 2016. Available at https://www.npr.org/sections/alltechconsidered/2016/10/25/499176469/it-aint-me-babe-researchers-find-flaws-in-police-facial-recognition (last accessed April 19, 2019).
23. 23 Guerin, Lisa. "Disparate Impact Discrimination." www.nolo.com. Available at https://www.nolo.com/legal-encyclopedia/disparate-impact-discrimination.htm (last accessed April 24, 2019). See also, Jewel v. NSA where the Electronic Frontier Foundation argues that massive (or dragnet) surveillance is illegal. Information about case available at https://www.eff.org/cases/jewel (last accessed April 19, 2019).
24. 24 This is often called an anti-classification criterion that the algorithm cannot classify based on membership in the protected or sensitive classes.
25. 25 Zarsky, Tal. "Understanding Discrimination in the Scored Society." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 15, 2015. https://papers.ssrn.com/abstract=2550248.
26. 26 Larson, Jeff, Surya Mattu, and Julia Angwin. "Unintended Consequences of Geographic Targeting." Technology Science, September 1, 2015. Available at https://techscience.org/a/2015090103/ (last accessed April 19, 2019).
27. 27 Terry Parris Jr Julia Angwin, "Facebook Lets Advertisers Exclude Users by Race," text/html, ProPublica, October 28, 2016. Available at https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race (last accessed April 19, 2019).
28. 28 Amazon doesn't consider the race of its customers. Should It? Bloomberg.com. Available at http//www.bloomberg.com/graphics/2016-amazon-same-day (last accessed April 19, 2019).
29. 29 Corbett-Davies et al., "Algorithmic Decision Making and the Cost of Fairness."
30. 30 Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2016. Available at https://papers.ssrn.com/abstract=2477899.
31. 31 See, Zafar, Muhammad Bilal, Isabel Valera Martinez, Manuel Gomez Rodriguez, and Krishna Gummadi. "Fairness Constraints: A Mechanism for Fair Classification." In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, FL, 2017. *See also*, Spielkamp, Matthias. "We Need to Shine More Light on Algorithms so They Can Help Reduce Bias, Not Perpetuate It." MIT Technology Review. Accessed September 20, 2018. Available at https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/ (last accessed April 19, 2019). See also Corbett-Davies, Sam, Emma Peirson, Avi Feller, and Sharad Goel. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." Washington Post (blog), October 17, 2016. Available at https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ (last accessed April 19. 2019).
32. 32 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores."In Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017. Available at https://arxiv.org/pdf/1609.05807.pdf (last accessed April 19, 2019).
33. 33 This notion of disparate impact has been legally tested dating back to the 1971 U.S. Supreme Court decision, Griggs v. Duke Power Company where the defendant was found to be using intelligence test scores and high school diplomas as factors to hire more white applicants over people of color. As determined by the court decision, there was no correlation between the tests or educational requirements for the jobs in question. See, Griggs v. Duke Power Company, Oyez. Available at https//www.oyez.org/cases/1970/124 (last accessed October 1, 2018.

34. 34 Various computer models are being created to combat the discriminatory effects of algorithmic bias. *See*, Romei, Andrea, and Salvatore Ruggieri. "Discrimination Data Analysis: A Multi-Disciplinary Bibliography." In Discrimination and Privacy in the Information Society, edited by Bart Custers, T Calders, B Schermer, and T Zarsky, 109–35. Studies in Applied Philosophy, Epistemology and Rational Ethics. Springer, Berlin, Heidelberg, 2013. Available at https://doi.org/10.1007/978-3-642-30487-3_6 (last accessed April 19, 2019).
35. 35 Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic Decision Making and the Cost of Fairness." ArXiv:1701.08230 [Cs, Stat], January 27, 2017. Available at https://doi.org/10.1145/3097983.309809 (last accessed April 19, 2019).
36. 36 Ibid.
37. 37 Schatz, Brian. AI in Government Act of 2018, Pub. L. No. S.B. 3502 (2018). https://www.congress.gov/bill/115th-congress/senate-bill/3502.
38. 38 At its February meeting, the OECD announced that it had approved its expert group's guidelines and hoped to (C.
39. 39 See European Union, Digital Single Market, Ethics Guidelines for Trustworthy AI, available for download from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (last accessed April 19, 2019).
40. 40 *See* High-level Expert Group on Artificial Intelligence. "Ethics Guidelines for Trustworthy AI (Draft)." The European Commission, December 18, 2018. *See also*, Chessell, Mandy. "Ethics for Big Data and Analytics." IBM, n.d. Available at https://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf (last accessed April 19, 2019).
https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_draft_ethics_guidelines_18_december.pdf. *See also* "The Global Data Ethics Project." Data for Democracy, n.d. https://www.datafordemocracy.org/project/global-data-ethics-project (last accessed April 19, 2019).
41. 41 Spielkamp, Matthias. "We need to shine more light on algorithms so they can help reduce bias, not perpetuate It." MIT Technology Review. Available at https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/ (last accessed September 20, 2018).
42. 42 Tobin, Ariana. "HUD sues Facebook over housing discrimination and says the company's algorithms have made the problem worse." ProPublica (March 28, 2019). Available at https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms (last accessed April 29, 2019).
43. 43 Elejalde-Ruiz, Alexia. "The end of the resume? Hiring is in the midst of technological revolution with algorithms, chatbots." Chicago Tribune (July 19, 2018). Available at http://www.chicagotribune.com/business/ct-biz-artificial-intelligence-hiring-20180719-story.html.
44. 44 Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability." New York: AI Now, April 2018.
45. 45 Alexandra Chouldechova et al., "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions," *1st Conference on Fairness, Accountability and Transparency*, n.d., 15.
46. 46 Rhema Vaithianathan et al., "Section 7: Allegheny Family Screening Tool: Methodology, Version 2," April 2019.
47. 47 Locklear, Mallory. "Facebook Releases an Update on Its Civil Rights Audit." Engadget (blog), December 18, 2018. Available at https://www.engadget.com/2018/12/18/facebook-update-civil-rights-audit/ (last accessed April 19, 2019).
48. 48 Stack, Liam. "Facebook Announces New Policy to Ban White Nationalist Content." The New York Times, March 28, 2019, sec. Business. Available at https://www.nytimes.com/2019/03/27/business/facebook-white-nationalist-supremacist.html (last accessed April 19, 2019).
49. 49 Qtd. in Rachel Courtland, "Bias Detectives: The Researchers Striving to Make Algorithms Fair," *Nature* 558, no. 7710 (June 2018): 357–60. Available at https://doi.org/10.1038/d41586-018-05469-3 (last accessed April 19, 2019).
50. 50 Fintech regulatory sandboxes in UK, Singapore, and states in the U.S. are beginning to authorize them. They allow freedom to offer new financial products and use new technologies such as blockchain.
51. 51 In March, the state of Arizona became the first U.S. state to create a "regulatory sandbox" for fintech companies, allowing them to test financial products on customers with lighter regulations. The U.K. has run a similar initiative called Project Innovate since 2014. The application of a sandbox can allow both startup companies and incumbent banks to experiment with more innovative products without worrying about how to reconcile them with existing rules.
52. 52 Mnuchin, Steven T., and Craig S. Phillips. "A Financial System That Creates Economic Opportunities - Nonbank Financials, Fintech, and Innovation." Washington, D.C.: U.S. Department of the Treasury, July 2018.Available at https://home.treasury.gov/sites/default/files/2018-08/A-Financial-System-that-Creates-Economic-Opportunities---Nonbank-Financials-Fintech-and-Innovation_0.pdf (last accessed April 19, 2019).
53. 53 Another major tech-related Safe Harbor is the EU-US Privacy Shield after the previous Safe Harbor was declared invalid in the EU. Available at https://en.wikipedia.org/wiki/EU%E2%80%93US_Privacy_Shield (last accessed April 19. 2019).
54. 54 Lopez, German. "The First Step Act, Congress's Criminal Justice Reform Bill, Explained." Vox, December 3, 2018. Available at https://www.vox.com/future-perfect/2018/12/3/18122392/first-step-act-criminal-justice-reform-bill-congress (last accessed April 16, 2019).