Title

The Sins of the Parents Are to Be Laid Upon the Children: Biased Humans, Biased Data, Biased Models.

Permalink
https://escholarship.org/uc/item/8mg3f95f

Authors
Osborne, Merrick R
Omrani, Ali
Dehghani, Morteza

Publication Date
2023-07-18

DOI
10.1177/17456916231180099

The Sins of the Parents Are to Be Laid Upon the Children: Biased Humans, Biased Data, Biased Models

Merrick R. Osborne[1], Ali Omrani[2,3], and Morteza Dehghani[2,3,4]

[1] Haas School of Business, University of California, Berkeley;
[2] Department of Computer Science, University of Southern California;
[3] Brain and Creativity Institute, University of Southern California;
[4] Department of Psychology, University of Southern Californi

Abstract

Technological innovations have become a key driver of societal advancements. Nowhere is this more evident than in the field of machine learning (ML), which has developed algorithmic models that shape our decisions, behaviors, and outcomes. These tools have widespread use, in part, because they can synthesize massive amounts of data to make seemingly objective recommendations. Yet, in the past few years, the ML community has been drawing attention to the need for caution when interpreting and using these models. This is because these models are created by humans, from data generated by humans, whose psychology allows for various biases that impact how the models are developed, trained, tested, and interpreted. As psychologists, we thus face a fork in the road: Down the first path, we can continue to use these models without examining and addressing these critical flaws and rely on computer scientists to try to mitigate them. Down the second path, we can turn our expertise in bias toward this growing field, collaborating with computer scientists to reduce the models' deleterious outcomes. This article serves to light the way down the second path by identifying how extant psychological research can help examine and curtail bias in ML models.

Keywords

machine learning, bias, prejudice, fairness

Introduction

Machine learning (ML) models are now a quintessential part of everyday life. It is easy to understand why these models have become so valuable: They are helpful tools that can identify and synthesize complex patterns in large data sets, and they often outperform traditional statistical methods in prediction and classification across a variety of tasks (e.g., Goretzko & Bühner, 2020; Yeomans, 2021). Their ability to simplify decision- making in a seemingly unbiased manner makes them essential for many power holders and laypeople alike. After all, ML models appear to be cold, objective algorithms that—supposedly—are not hindered by the same heuristics and biases that plague human psychology. However, in reality, these models are developed by imperfect humans using imperfect data generated in imperfect societies. These imperfections are consequential because ML models both reflect and amplify the same types of biases that humans have (e.g., Angwin & Larson, 2016; Bender et al., 2021; Bolukbasi et al., 2016; Caliskan et al., 2017; Hovy & Spruit, 2016).

Determining how bias arises in ML is critical to reducing the potential for the model's problematic outputs. Indeed, research in ML—as well as in social and cognitive psychology—has demonstrated that the biases present in these models reflect human biases in the real world (Caliskan et al., 2017; Charlesworth et al., 2021; Garg et al., 2018; Koenecke et al., 2020); for instance, they have generated decisions that ensure Black and Brown people (relative to White people) are more likely to experience unfavorable outcomes when applying for a loan (Oneto & Chiappa, 2020), receiving sentencing decisions (Angwin & Larson, 2016), or obtaining health related care (Obermeyer et al., 2019). Unfortunately, these models can perpetuate extant societal inequalities that further disadvantage marginalized people by unjustly ruling against their favor. Preventing bias from seeping into these models, and decreasing their capacity to output biased decisions, is critical to developing a society in which advancements in technology run parallel with advancing societal equality.

Fortunately, ML scientists and engineers have started examining how bias permeates the construction of ML models. They have highlighted how a variety of these sources—such as biases baked into the training data itself (e.g., Bolukbasi et al., 2016), biases in the annotations of the training data (e.g., Davani et al., 2021), and the underrepresentation of certain populations in the data (e.g., Lucy & Bamman, 2021; Mehrabi et al., 2021; Shankar et al., 2017)—contribute to biases in the model output. By bringing awareness to how these factors amplify extant biases, they intend to highlight potential biases in the decisions made by the models and propose mechanisms to reduce bias in the model's outputs. This area of inquiry, called *fairness in machine learning*, has generated a variety of tools (e.g., Mehrabi et al., 2021) and insights (e.g., Barocas et al., 2019) to help ML researchers and practitioners identify, and subsequently reduce, bias in ML models.

Fairness in machine learning is a relatively new field. The novelty of fairness originates, in part, from ML's past dominant objective of maximizing models' predictive accuracy. Although a model's ability to correctly classify an input is important, researchers have recently sought to understand how they can create "fair" models that do not disproportionately benefit one group of people over others. Although some work has noted that fairness and accuracy may not always be in tension with one another when constructing ML models (e.g., Pessach & Shmueli, 2021; Wick et al., 2019), work in ML largely acknowledges that increasing fairness has the risk of negatively impacting a model's accuracy (Menon & Williamson, 2018; Zafar et al., 2019).

To that end, ML researchers have raised a signpost to other research communities indicating that there is work to be done in understanding how bias can be mitigated in ML (Blodgett et al., 2020). Yet, there is relatively little work outside of ML, including within psychology, that explicitly identifies and explains this new frontier of bias manifestation. We would thus like to draw psychologists' attention to the computer scientists' aforementioned signpost, indicating that we face a fork in the road.

Down the first path lies psychologists' continued usage of ML models, which are currently developed and implemented without significant involvement or informed scrutiny from psychologists. These models have appeared in research examining how facial- recognition software predicts users' sexuality (Wang & Kosinski, 2018), what jobs people are likely to opt into (Song et al., 2022), how moral rhetoric online predicts future arrests during protests (Mooijman et al., 2018) and even in theory development (e.g., Leavitt et al., 2021; Yarkoni & Westfall, 2017). However, this path is rockier than it seems. Topics like bias and prejudice have long been a core research focus in psychology (e.g., Banaji & Hardin, 1996; Greenwald & Banaji, 1995; Roberts & Rizzo, 2021; Seaton et al., 2018); yet, although ML models are a new manifestation and representation of bias, our field, with a few exceptions, has not sufficiently incorporated these advancements into our understanding of how modern forms of bias manifest in contemporary life. Nonetheless, the theories and insights that psychologists have developed over the past century not only are applicable to computer scientists and fairness scholars but could be their guiding light in detecting—and preventing—bias in technology.

This brings us to the second path: psychologists collaborating closely with computer scientists to identify and reduce bias in machine learning models. This path will serve to be both theoretically fruitful and practically important. Thus far, the methods developed in computer science have led computer scientists to generate solutions that address bias as it arises in the process of developing ML models. These solutions are not always completely satisfactory, as they are more inclined to hide the bias in the model rather than eliminate it (e.g., Gonen & Goldberg, 2019). In contrast, over the better part of the past century, psychologists have illuminated a variety of strategies to directly eliminate bias at the source. In other words, currently, both psychologists and computer scientists look at different pieces of the problem without necessarily accounting for the entire process. Psychology has focused its efforts on understanding how bias forms in people, whereas computer science has been more interested in how models become biased.

Without examining how human biases find their ways into ML models, and how such biases can get eliminated, we will continue to lack a full picture of how technology can be a purveyor of bias. The aim of this article, then, is to light the way down the second path for psychologists and encourage them to turn their attention toward investigating both how bias arises in ML models and how it can be mitigated. By highlighting how psychological biases arise in ML models, psychologists can further inform fairness researchers' efforts to develop generative and interesting solutions that systematically address the issue.

Sources of Bias in Machine Learning Models

In this section, we review three potential sources of bias for ML models: First, we discuss how biases arise in the composition and content of training data. Second, we describe how ML engineers' own internal biases affect model design. Finally, we note that there is a historical lack of knowledge about the possibility of bias in ML because the field of fairness research is relatively new. See Figure 1 for an overview of our proposed framework.

Source of bias: Training data

The first source of bias stems from the training data. At a broad level, it is reasonable to think that data sets just serve as reflections of the current state of our world. Yet, in reality, they are reflections of the processes that went into collecting the data and curating the data set. In other words, psychological characteristics encoded in the training data affect how the model makes judgments—thus, the way that these characteristics are represented and accounted for in constructing the data set is essential to generating equitable outcomes (Kilbertus et al., 2018). We highlight two important sources of bias that ML engineers may overlook when compiling training data sets, specifically, how much bias already exists in the data set and how such bias can get exacerbated when constructing ground truths data using human annotations.

Extant biases in the data set. Identifying extant biases in the data is an important part of detecting and eliminating bias in the model's outputs. Indeed, data are a product of the historical context in which they were generated (Suresh & Guttag, 2019), leading to *historical bias* that can perpetuate extant inequalities. For instance, Bolukbasi et al. (2016) and Caliskan et al. (2022) demonstrate that word embeddings contain biases that reflect gender stereotypes in broader society. Because word embeddings are frequently used in real-world applications, Bolukbasi et al. argue that they both reflect and perpetuate these stereotypes.

Some data sets may also hold biases because of inadequate representation of people with *protected traits*. Protected traits are facets of people's identities that could lead to marginalization, such as their race, gender, sexuality, and so on (Corbett-Davies & Goel, 2018). Although ML models do not always prioritize protected traits, such traits can be disproportionately weighted in the model's decision-making process. Oftentimes, the infrastructures and institutions that house the source data make it difficult to collect data sets where those with protected traits are well represented— that is, where the same percentage of people who hold intersecting protected traits in the general population are represented in the data set. For instance, platforms like Twitter, Facebook, and Reddit—as well as resources like the Common Crawl Corpus—oftentimes do not reflect the composition of the general population across various protected traits (Luccioni & Viviano, 2021; Odabas¸, 2022). As a result, there may be too few observations to accurately represent those with underrepresented protected traits or unrepresentative sampling that can limit a model's generalizability (Robinson et al., 2020). Indeed, computer scientists have frequently indicated the importance of considering the issue of underrepresentation in data sets (see the literature on "representation bias"; e.g., Lucy & Bamman, 2021; Mehrabi et al., 2021; Shankar et al., 2017), noting that it can arise from nonrandom sampling (Mehrabi et al., 2021) and lead the model to make decisions that perpetuate further marginalization. Thus, the absence of marginalized populations is as fundamental to the model's output and decision-making as their presence would be.

Another potential reason for models' biased decisions is engineers' reliance on *proxy labels*, markers representing the operationalizations of constructs that need to be captured but can be captured only indirectly. For instance, the engineer may want to assess employee performance (the true label); if there is no direct and objective way to capture this, they would instead use a proxy variable (i.e., human assessments of performance). Proxy labels capture only "a particular aspect of what we want to measure" (Suresh & Guttag, 2019; p. 4), as determining what is captured in this proxy label is up to the engineers' discretion and thus may also be subject to the engineers' biases. Unfortunately, these labels could inadvertently perpetuate inequality by capturing external factors that skew the model's output (Stock & Cisse, 2018; Suresh & Guttag, 2019; Van Miltenburg, 2016). For instance, Correctional Offender Management Profiling for Alternative Sanction (COMPAS; Angwin & Larson, 2016) was a tool created to predict recidivism risk in convicted criminals using a variety of proxy labels. However, to predict future recidivism, the tool relied on a variety of proxy labels, like the felon's own prior arrests as well as their family's. These labels inadvertently perpetuated bias; many of those who were predicted to have a high risk of recidivism came from communities that were initially overpoliced, which in turn impacted the felon's prior arrests and family arrests.

Biases in the annotators. Supervised ML models need ground-truth data for training and fine-tuning. To create ground-truth data, model engineers employ the help of annotators to go through a portion of the data and label them for the desired set of categories. Some annotation tasks require distinguishing between clearly defined, and objective, categories (e.g., cats vs. dogs, verbs vs. adjectives). Increasingly, however, ML models are also being used to make subjective decisions. For example, the authors of this article often train ML models for moral-sentiment classification. This is a subjective task where the background of the annotators plays an important role in detecting and categorizing different facets of morality in text. However, the people who annotate our data, like the model engineers themselves, are imperfect people who hold their own biases (Davani et al., 2021; Van Miltenburg, 2016). Given that the annotators' inputs serve as learning and tuning guidelines for the model, their biases directly influence the model's decisions. Moreover, how model engineers select annotators—or account for annotator's responses— could potentially erase the input from annotators with underrepresented identities, further perpetuating bias in the model's output (Prabhakaran et al., 2021).

Source of bias: ML engineers

The second source of bias is the people who develop and design the models themselves. We argue that model engineers are imbued with a sense of power that is derived either from the psychological consequences of managing a highly valuable resource (i.e., ML models) and/or is granted by the privileges that society affords those who possess their demographic characteristics. Model engineers have a sense of psychological power (Anderson et al., 2012; Anderson & Galinsky, 2006; Tost, 2015) because ML models are valued resources; their output is oftentimes vital to the day-to-day operations of many, making the scope of their role in contemporary life challenging to overstate. Model engineers are also likely to possess racial and gender identities that are societally privileged; indeed, they are overwhelmingly White and male (Blodgett & O'Connor, 2017; Jaccheri, 2022). In other words, model engineers are more likely to possess societal power that is derived from the model engineer's demographic characteristics (Fiske & Berdahl, 2007).

Both experiencing an elevated psychological sense of power and possessing societal power have been associated with reduced inhibitions (Cho & Keltner, 2020; Fiske, 1993; Keltner et al., 2003). Consequently, model engineers are positioned to feel fewer psychological constraints preventing them from acting on, or even detecting, their own biases—thus obfuscating their ability to effectively interrogate how their models could impact important societal outcomes for different groups of users. Overriding these biases and stereotypes can be done by exercising self-control (Guinote, 2017), but those in power do not always have the resources or the motivation to exercise the necessary self-control to overcome their biases (Fiske & Berdahl, 2007; Guinote, 2017). As a result, model engineers' heightened experiences of (psychological or societal) power may lead them to perpetuate biases in their ML models.

Source of bias: negligence

The third source of bias stems from model engineers' ambivalence and negligence about the biased outcomes of their models. Given the technology's novelty and its promises of objectivity (Bogert et al., 2022), scientists did not direct much attention to the ways that ML models could make systematically biased decisions until recent years (e.g., Barocas et al., 2019). As a result, many of these models have been developed without thorough scrutiny—permitting some models to produce biased output. Even now, as more scientists across different fields raise increasingly visible (and urgent) signposts highlighting the deleterious outcomes associated with unfair models, not all model engineers may be paying as much attention as this problem requires. For instance, although many top industry labs are among the leaders in fairness in artificial intelligence research, for smaller companies—with much more limited resources—fairness might not be a high priority.

Fairness in ML

Broadly, fairness research in the ML community has been focused on defining mathematical formalizations of fairness criteria—often inspired from social science literature—and developing methods and models to satisfy these criteria. This is a relatively nascent field, but over the past few years, there has been a sharp increase in fairness research (Chouldechova & Roth, 2020). A popular focus of the field is on natural language processing and seeks to solve problems such as debiasing embeddings (Bolukbasi et al., 2016). However, more recently, fairness research has extended into other domains, such as speech (Koenecke et al., 2020) and face (Buolamwini & Gebru, 2018) recognition.

In what follows, we first review the prominent formalizations of fairness in ML, focusing on allocational harm as a consequence of the decisions made by the model in the context of classification. Then, we discuss quantifications of bias captured in representations learned by these models. It should be noted, though, that a complete review of all operationalizations of fairness and bias in ML is beyond the scope of this work. For a more comprehensive review of measures of fairness, we refer the reader to Mehrabi et al. (2021).

Allocational harms

Classification is one of the major applications of ML. Formally, the model's goal in classification is to predict a target variable Y (e.g., a hiring decision), given an observation X (e.g., one's resume). When these applications involve allocating resources or opportunities, the model can discriminate against certain groups and bring about *allocational harms* (i.e., harms caused by a model denying a certain group access to a resource or an opportunity). Measures of algorithmic fairness in classification can be divided into three categories of group, subgroup, and individual fairness. Across all definitions discussed next, we use S to denote the protected attribute (e.g., gender), Y to denote the model prediction, and Y to denote the ground truth. Moreover, let $S=1$ denote the majority group (e.g., males in science, technology, engineering, and mathematics) and $Y=1$ be the preferred outcome (e.g., getting hired).

Group fairness. This family of fairness measures focuses on treating different groups equally. There are multiple ways to assess group fairness: through disparate impact, statistical parity, equality of opportunity, and equality of odds.

Disparate impact. Viewed as a mathematical formalization of the "80% rule" in the legal notion of disparate impact (Feldman et al., 2015), which states that the allocation rate of a desired resource for the protected group should be at least 80% of the allocation rate of that desired resource for the nonprotected group, the measure of disparate impact assesses the extent to which members of one group experience the desired outcome relative to other groups. Intuitively, this measure ensures that the ratio of assigning desired outcomes to groups remains close. Formally, disparate impact is defined as

$$\frac{P\left[\hat{Y}=1 \mid S \neq 1\right]}{P\left[\hat{Y}=1 \mid S=1\right]} \geq 1-\epsilon. \qquad (1)$$

Relying on this equation, model engineers are able to assess the differences in desired outcome rates across groups ($PY[\ = 1|S \neq 1]$ versus $PY[\ = 1|S = 1]$). Lower values of disparate impact point to a difference in desired outcome rates across groups (unfair), and higher values show similar desired outcome rates across groups (fair).

Statistical (or demographic) parity. This measure ensures that majority and minority groups have equal probability of getting assigned the desired outcome from the model. Dwork et al. (2012) define statistical parity as

$$| P[\hat{Y} = 1 \mid S = 1] - P[\hat{Y} = 1 \mid S \neq 1] | \leq \epsilon. \qquad (2)$$

The two terms in the equation denote the probability of a positive outcome ($\hat{Y}$=1) for different groups (S =1 and S $\neq$ 1). Lower values of this measure indicate similar positive outcome rates across groups (fair), and higher values are evidence of unequal rates (unfair).

Equality of opportunity and equality of odds. Hardt et al. (2016) introduced equality of opportunity, which requires the model outputs to assign the desired outcome ($\hat{Y}$=1) to people in the desired ground truth (Y =1) of different groups ( )S with equal probability. An unfair model, according to this definition, will have unequal rates of assigning the desired outcome to people with the desired ground truth for different groups. Formally, this fairness criterion is defined as

$$P\left(\hat{Y} = 1 | S = 0, Y = 1\right) = P\left(\hat{Y} = 1 | S = 1, Y = 1\right). \qquad (3)$$

However, in the equality-of-odds measure, Hardt et al. (2016) extend this definition beyond the desired ground truth (Y = 1), requiring the model to have equal probability of assigning the desired outcome to different groups (e.g., S =1 or S = 0) for all possible values of ground truth (Y y=). For example, an unfair model according to this definition could have different false positive rates ($P[\hat{Y}$= 1|Y = 0]) for different groups. Formally, equality of odds is defined as

$$P(\hat{Y} = 1 \mid S = 0, Y = y) = P(\hat{Y} = 1 \mid S = 1, Y = y). \qquad (4)$$

Both equality of odds and opportunity assess the extent to which people are classified relative to the desired outcome. Because both measures consider some notion of equality between groups and ignore individuals, pushing models to satisfy either of them may result in two similar individuals receiving different treatments, which often is prohibited by law. Additionally, a perfect classifier, meaning a classifier that predicts the correct outcome ($\hat{Y}$ = Y) for all instances, may still be considered unfair according to both of these measures. Note that a perfect classifier would essentially reproduce the ground-truth labels Y given the input features X. Therefore, if the input data set (X,Y) is biased according to some definition, the perfect classifier trained on it would also be biased.

Subgroup fairness. Kearns et al. (2018, 2019) extend the definition of group fairness for subgroups. Instead of asking definitions of fairness to hold for a number of coarse groups, they ask for the definitions of fairness to hold for an infinitely large collection of subgroups defined as g : X→ {0 , 1} ∈G, where g(x) =1 denotes an individual's membership in group g. For example, false positive subgroup fairness requires the model to have equal false-positive rates overall and for any group *g*. Formally, λ-false-positive fairness with respect to *G* is achieved when, for all *g* ∈*G* ,

$$\alpha_{FP} . \beta_{FP} \leq \lambda, \tag{5}$$

$$\alpha_{FP}(g) = P[g(x) = 1, y = 0], \text{ and} \tag{6}$$

$$\beta_{FP}(g) = |P[\hat{Y} = 1 | y = 0] - P[\hat{Y} = 1 | g(x) = 1, y = 0]| . \tag{7}$$

Individual fairness. Although group notions of fairness are desirable, they are not strong enough in all scenarios. For example, group fairness puts no constraints over which members of each group are granted an opportunity. An undesirable model could satisfy group fairness criteria but offer the opportunity only to individuals within the minority group who are less likely to benefit from it. Unlike the measures discussed so far that focus on groups, individual fairness posits that "similar" individuals should be treated similarly. Next, we list three strategies for achieving individual fairness.

Fairness through awareness. Fairness through awareness formalizes the core idea of similar treatment of similar individuals by imposing constraints (e.g., a Lipschitz condition) on the classifier (Dwork et al., 2012).

Fairness through unawareness. Kusner et al. (2017) states, "An algorithm is fair as long as any protected attributes . . . are not explicitly used in decision-making process" (p. 2). However, more recent work has shown that fairness through unawareness may not be as effective at advancing fairness as previously thought, given the model's ability to uncover protected attributes without explicitly relying on them (Pedreshi et al., 2008; Zemel et al., 2013). For example, race could be correlated to, and therefore inferred from, neighborhood.

Counterfactual fairness. Kusner et al. (2017) formalizes the idea that a decision for an individual is fair if it does not change in a counterfactual world where the individual belongs to a different group.

Representational harms

Instead of focusing on resource allocation or opportunity granting, *representational harms* are concerned with providing fair representations of groups and individuals at the same time. Most of the success in ML today can be attributed to deep learning models' ability to build informative representations of the world. However, these representations often capture and amplify historical inequalities. Therefore, various measures have recently been devised to quantify the extent to which ML models capture and potentially amplify our society's stereotypical biases. The seminal work by Caliskan et al. (2017) demonstrated that word-embedding models replicate human-like stereotypes and proposed the Word Embedding

Association Test (WEAT) derived from the Implicit Association Test (IAT; Greenwald & Banaji, 1995): Formally, let *X* and *Y* be two equal-sized sets of target words (e.g., science-related words and art-related words) and *AB*, be two sets of attribute words (e.g., male and female words); then the effect size and *p* value of WEAT are defined as

$$\text{effect size} = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(x, A, B)}{\text{std} - \text{dev}_{w \in X \cup Y} s(w, A, B)} \text{ and } \quad (8)$$

$$p = P_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)], \quad (9)$$

where $\{(X\ Y_i,\ _i)\}$ is used to denote partitions of $X \cup Y$ to equal size sets and

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}). \quad (10)$$

More recent work has extended the preceding formalization from word embeddings to language models while attempting to provide more coverage on the variety of stereotypes covered; for example, May et al. (2019) attempted to extend WEAT to sentences, and Nadeem et al. (2020) proposed a measure to quantify the stereotypes in state-of-the-art language models. Unsurprisingly, they demonstrate that language models encode human stereotypes to varying degrees.

Impossibility of fairness

Ideally, we want models that are fair with respect to all the aforementioned definitions. However, Chouldechova (2017) and Kleinberg et al. (2016) demonstrate *impossibility of fairness* by proving that it is impossible for a model to satisfy all definitions of fairness in decision-making simultaneously. Green (2022) argues this can be alleviated by moving toward substantive evaluations of the role of models in promoting justice in practical settings.

How Psychologists Can Help Reduce Algorithmic Bias

As psychologists uncover the various ways that psychological bias is prevalent in ML models, they may begin to generate solutions to reduce this bias. Central to the psychological literature of understanding bias and prejudice is identifying ways to reduce it (e.g., see Mallett & Monteith, 2019; Paluck et al., 2021; Paluck & Green, 2009). In this section, we point psychologists to some ways that their research can help model engineers identify, and reduce, extant biases. We use psychological theories to address two issues model engineers may run into when designing these models: identifying existing biases and accounting for protected characteristics.

Intergroup contact theory

Some fairness researchers recommend that model engineers should interact with people who are affected by their models (e.g., Blodgett et al., 2020), as doing so can help the engineers understand how their models impact the lived experiences of different communities— especially marginalized communities. Work in psychology provides empirical support for this recommendation: Namely, under some conditions, engaging with outgroup members can reduce prejudice and discrimination (Pettigrew & Tropp, 2006; Reimer & Sengupta, 2022). For instance, *incidental* intergroup contact, where people engage with outgroup members fleetingly, has the potential to reduce prejudice (Anicich et al., 2021b). Indeed, this may be achieved when these interactions are structured by the model engineers' employer, as institutionally supported intergroup contact is an effective way to improve intergroup relations (Anicich et al., 2021a).

Color blindness in constructing models

One of the key goals of fairness in machine learning is to derive mathematical formulations of nondiscrimination in decision-making to reduce bias expressed in the models. These formulations set constraints on the shapes of distributions relative to different protected traits (Oneto & Chiappa, 2020). There is active discussion within the ML community to understand the best way to represent protected traits. Some point out that model engineers could design models that do not account for protected traits (Chouldechova & Roth, 2020; Kusner et al., 2017) but note that this is not always adequate for removing bias (Veale & Binns, 2017; Žliobaite & Custers, 2016). This neatly dovetails into the psychological research on color blindness, or the belief that racial group membership should not be accounted for in decision-making (Apfelbaum et al., 2012). However, higher levels of color-blindness (vs. awareness) ideologies is associated with increased bias (Plaut et al., 2018) and other negative outcomes. For instance, among White people, higher levels of color blindness have been associated with reduced understanding of marginalized people's unique realities (Neville et al., 2013), greater apathy to racism (Tynes & Markoe, 2010), and less willingness to support antidiscrimination efforts (Awad et al., 2005). This is because color blindness can correct for visible sources of bias (e.g., making a colorblind hiring decision can reduce biases derived from names) but cannot take into account embedded biases (e.g., lack of access to resources and necessary support to make a resume or curriculum vitae strong and attractive). Thus, reducing model engineers' color blindness in the model construction process may be instrumental to reducing bias in ML. Fairness scholars have identified the value of acknowledging protected characteristics, too: For instance, Bender et al. (2021) note the power of culturally appropriate training data, and Kilbertus et al. (2018) state that including protected characteristics can help model engineers understand if a model is really fair. Moreover, race-neutral approaches to model development ultimately have poorer prediction rates for racial-minority populations (Robinson et al., 2020).

Of course, psychology can help extend computer scientists' understanding of the notion of fairness itself. For example, recent research in moral psychology has argued that the notion of fairness is too broad and vague, and it may be better captured by breaking it down further into notions of equality and proportionality (Atari et al., 2023; Rai & Fiske, 2011). Other lines of work could look at cultural differences in perceptions and concerns about fairness; for instance, psychology's extant work on procedural justice could help fairness scholars understand how engineers' desire to voice concerns about unfair model inputs varies by culture (Brockner et al., 2001). Additionally, psychologists can investigate how psychological bias gets transferred from engineers to models, as research on implicit social cognition would argue that model engineers nonconsciously transfer beliefs and attitudes into the model (Greenwald & Banaji, 1995), impeding their ability to input fairness constraints into the model.

Other major theories in psychology, such as the stereotype content model (SCM; Cuddy et al., 2008, 2009), can also help reduce stereotypical biases learned by ML models. SCM posits that the content of stereotypes, as opposed to their process, can be decomposed along two fundamental dimensions of warmth and competence (Cuddy et al., 2008). Recent research from our own team has applied SCM to characterize how the stereotypical biases of annotators transfers to ML models (Davani et al., 2021) and demonstrated the efficacy of SCM in mitigating representational harms in natural language processing across a range of social groups (Omrani et al., 2022).

Without direct collaborations, much is lost between the two fields. Research in ML, and investigations into fairness in ML, provides an exciting new context wherein psychologists can test and expand their theories. Research in psychology, on the other hand, can inform tools, strategies, and processes to help ML model engineers identify and mitigate their own bias and bias in their models.

Discussion and Conclusion

ML models are not created in a vacuum; they are designed and built by humans and are trained and optimized using data from a thin slice of the human population. Moreover, humans' psychological biases beset ML—these biases arise, in part, from existing biases in the data set, heightened feelings of power in model engineers, and a negligence about how biases manifest in ML. As a result of these biases, ML models—and consequently, the humans who rely on them—can make problematic decisions that further perpetuate social inequalities. Our goal in writing this article is to encourage psychologists to think about how their work could contribute to our understanding, and mitigating, of bias in ML.

Specifically, we call for more collaborations between psychologists and ML researchers to develop frameworks established in psychological theories that can be used to explain how bias leaks into the models. The complexity of today's dominant ML pipelines can prohibit a deep understanding of their exact operating. Therefore, leveraging the rich body of psychological research on human biases has the potential to help ML researchers and practitioners understand the sources of bias and mitigate them in ML pipelines at a deeper level.

In this work, we have focused primarily on strategies that individual model engineers can take rather than strategies that organizations can implement. Investigating institution-level biases is likely a fruitful alternative approach for psychologists to understand how reducing bias at different levels (e.g., the organizational level) could have downstream consequences for the degree of bias the models exhibit. As much as people perpetuate bias, organizations can institutionalize it, particularly if they prioritize prediction accuracy—and consequently, profit—over fairness.

Birhane and Prabhu (2021), inspired by Benjamin (2019), reminded us that "feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty, is a fantasy" (p. 1540). In essence, the biases manifested in the ML models are a reflection of the psychological states of the creators of the models; this includes not just the model engineers but our society as a whole. In that way, the models are biased because “the sins of the parents are laid upon the children” (Shakespeare, 2022). It is important for us to use the insights that we have developed in psychology to shine our lights on the various inequities that ML models perpetuate and reduce bias in the model construction process. Yet, rather than reinventing the wheel, psychologists should be aware of the progress that fairness researchers have already made and build upon their innovations.

Figures

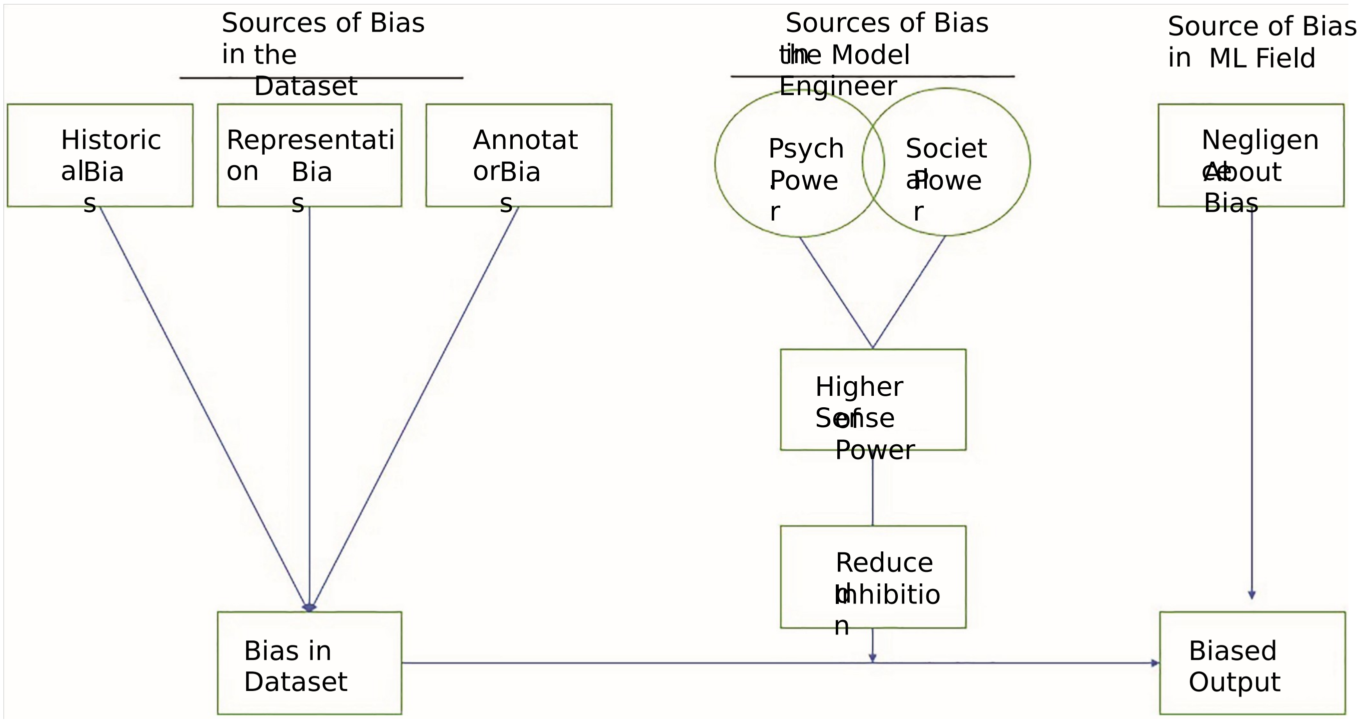

Fig. 1. Bias in the data and bias of model engineers influence bias in the model output.

References

Anderson, C., & Galinsky, A. D. (2006). Power, optimism, and risk-taking. *European Journal of Social Psychology*, *36*(4), 511–536.

Anderson, C., John, O. P., & Keltner, D. (2012). The personal sense of power. *Journal of Personality*, *80*(2), 313–344.

Angwin, J., & Larson, J. (2016). Bias in criminal risk scores is mathematically inevitable, researchers say. In *Ethics of data and analytics* (pp. 265–267). Auerbach.

Anicich, E. M., Jachimowicz, J. M., Osborne, M. R., & Phillips, L. T. (2021a). Design physical and digital spaces to foster inclusion. *Harvard Business Review*. Advance online publication. https://hbr.org/2021/08/design-physical-anddigital-spaces-to-foster-inclusion

Anicich, E. M., Jachimowicz, J. M., Osborne, M. R., & Phillips, L. T. (2021b). Structuring local environments to avoid racial diversity: Anxiety drives whites' geographical and institutional self-segregation preferences. *Journal of Experimental Social Psychology*, *95*, Article 104117.

Apfelbaum, E. P., Norton, M. I., & Sommers, S. R. (2012). Racial color blindness: Emergence, practice, and implications. *Current Directions in Psychological Science*, *21*(3), 205–209.

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2023). Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*. https://doi .org/10.31234/osf.io/q6c9r

Awad, G. H., Cokley, K., & Ravitch, J. (2005). Attitudes toward affirmative action: A comparison of color-blind versus modern racist attitudes. *Journal of Applied Social Psychology*, *35*(7), 1384–1399.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. http://www.fairmlbook.org

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In L. Irani, S. Kannan, M. Mitchell, & D. Robinson (Eds.), *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). Association for Computing Machinery.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity.

Birhane, A., & Prabhu, V. U. (2021). Large image datasets: A pyrrhic win for computer vision? In R. Farrell, C. Canton, L. Leal-Taixe, & J. Yu (Eds.), *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1536– 1546). The Computer Vision Foundation. https://doi .org/10.1109/WACV48630.2021.00158

Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (2020). *Language (technology) is power: A critical survey of "bias" in NLP*. ArXiv. arXiv:2005.14050.

Blodgett, S. L., & O'Connor, B. (2017). Racial disparity in natural language processing: A case study of social media African-American English. ArXiv. arXiv:1707.00061.

Bogert, E., Lauharatanahirun, N., & Schecter, A. (2022). Human preferences toward algorithmic advice in a word association task. *Scientific Reports*, *12*(1), 1–9.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*, 4349–4357. http://papers.nips.cc/paper/6228-man-istocomputer-programmer-as-woman-is-to-homemakerdebiasing-wordembeddings.pd

Brockner, J., Ackerman, G., Greenberg, J., Gelfand, M. J., Francesco, A. M., Chen, Z. X., Leung, K., Bierbrauer, G., Gomez, C., Kirkman, B. L., & Shapiro, D. (2001). Culture and procedural justice: The influence of power distance on reactions to voice. *Journal of Experimental Social Psychology*, *37*(4), 300–315.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. Friedler, & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). https://proceedings.mlr.press/v81/buolam wini18a.html

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., & Banaji, M. R. (2022). Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In M. Scheutz, R. Calo, M. Mara, & A. Zimmermann (Eds.), *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (pp. 156–170). Association for Computing Machinery. https://doi.org/10.1145/3514094 .3534162

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

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240.

Cho, M., & Keltner, D. (2020). Power, approach, and inhibition: Empirical advances of a theory. *Current Opinion in Psychology*, *33*, 196–200.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, *63*(5), 82–89.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. ArXiv. arXiv:1808.00023.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in Experimental Social Psychology*, *40*, 61–149.

Cuddy, A. J., Fiske, S. T., Kwan, V. S., Glick, P., Demoulin, S., Leyens, J.-P., Bond, M. H., Croizet, J.-C., Ellemers, N., Sleebos, E., Htun, T. T., Kim, H. J., Maio, G., Perry, J., Petkova, K., Todorov, V., Rodríguez-Bailón, R., Morales, E., Moya, M., . . . Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1–33.

Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2021). *Hate speech classifiers learn human-like social stereotypes*. ArXiv. arXiv:2110.14839.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In S. Goldwasser (Ed.), *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). Association for Computing Machinery. https://doi.org/10.1145/209 0236.2090255

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In T. Joachims, G. Webb, D. D. Margineantu, & G. Williams (Eds.), *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259–268). Association for Computing Machinery. https://doi.org/10 .1145/2783258.2783311

Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. *American Psychologist*, *48*(6), 621–628.

Fiske, S. T., & Berdahl, J. (2007). Social power. In A. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 678–692). Guilford.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, USA*, *115*(16), E3635–E3644.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. ArXiv. arXiv:1903.03862.

Goretzko, D., & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Psychological Methods*, *25*(6), 776–786.

Green, B. (2022). Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology*, *35*(4), Article 90.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

Guinote, A. (2017). How power affects people: Activating, wanting and goal seeking. *Annual Review of Psychology*, *68*, 353–381.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, (Vol. 29, pp. 3315– 3323). Neural Information Processing Systems Foundation.

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In K. Erk, & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the Association for Computational Linguistics* (Vol. 2, pp. 591–598). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096

Jaccheri, L. (2022). Gender issues in computer science research, education, and society. In E. Barendsen, & Simon (Eds.), *Proceedings of the 27th ACM conference on innovation and technology in computer science education* (Vol. 1, p. 4). Association for Computing Machinery. https://doi .org/10.1145/3502718.3534204

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2564–2572). Proceedings of Machine Learning Research.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. In A. Chouldechova, & F. Diaz (Eds.), *Proceedings of the 2019 conference on fairness, accountability, and transparency* (pp. 100–109). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287592

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review*, *110*(2), 265–284.

Kilbertus, N., Gascón, A., Kusner, M., Veale, M., Gummadi, K., & Weller, A. (2018). Blind justice: Fairness with encrypted sensitive attributes. In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2630–2639). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v80/kilbertus18a.html

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. ArXiv. arXiv:1609.05807.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, USA*, *117*(14), 7684–7689.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Leavitt, K., Schabram, K., Hariharan, P., & Barnes, C. M. (2021). Ghost in the machine: On organizational theory in the age of machine learning. *Academy of Management Review*, *46*(4), 750–777.

Luccioni, A. S., & Viviano, J. D. (2021). What's in the box? A preliminary analysis of undesirable content in the common crawl corpus. ArXiv. arXiv:2105.02732.

Lucy, L., & Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. In N. Akoury, F. Brahman, S. Chaturvedi, E. Clark, M. Iyyer, & L. J. Martin (Eds.), *Proceedings of the third workshop on narrative understanding* (pp. 48–55). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nuse-1.5

Magee, J. C., & Galinsky, A. D. (2008). 8 social hierarchy: The self-reinforcing nature of power and status. *Academy of Management Annals*, *2*(1), 351–398.

Mallett, R., & Monteith, M. (2019). Confronting prejudice and discrimination: The science of changing minds and behaviors. Academic Press.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). *On measuring social biases in sentence encoders*. ArXiv. arXiv:1903.10561.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. In S. Friedler, & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 107–118). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v81/ menon18a.html

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violent protests. *Nature Human Behavior*, *2*, 389–396.

Nadeem, M., Bethke, A., & Reddy, S. (2020). *Stereoset: Measuring stereotypical bias in pretrained language models*. ArXiv. arXiv:2004.09456.

Neville, H. A., Awad, G. H., Brooks, J. E., Flores, M. P., & Bluemel, J. (2013). Color-blind racial ideology: Theory, training, and measurement implications in psychology. *American Psychologist*, *68*(6), 455–466.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Odabas¸, M. (2022). *10 facts about Americans and Twitter*. https://www.pewresearch.org/fact-tank/2022/05/05/10facts-about-americans-and-twitter/

Omrani, A., Kennedy, B., Atari, M., & Dehghani, M. (2022). *Social-group-agnostic word embedding debiasing via the stereotype content model*. ArXiv. arXiv:2210.05831.

Oneto, L., & Chiappa, S. (2020). Fairness in machine learning. In L. Oneto, N. Navarin, A. Sperduti, & D. Anguita (Eds.), *Recent trends in learning from data* (pp. 155–196). Springer.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? a review and assessment of research and practice. *Annual Review of Psychology*, *60*(1), 339–367.

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, *72*, 533–560.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discriminationaware data mining. In B. Liu, & S. Sarawagi (Eds.), *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 560–568).

Pessach, D., & Shmueli, E. (2021). Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. *Expert Systems with Applications*, *185*, Article 115667.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783.

Plaut, V. C., Thomas, K. M., Hurd, K., & Romano, C. A. (2018). Do color blindness and multiculturalism remedy or foster discrimination and racism? *Current Directions in Psychological Science*, *27*(3), 200–206.

Prabhakaran, V., Davani, A. M., & Diaz, M. (2021). On releasing annotator-level labels and information in datasets. ArXiv. arXiv:2110.05699.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75.

Reimer, N. K., & Sengupta, N. K. (2022). Meta-analysis of the “ironic” effects of intergroup contact. *Journal of Personality and Social Psychology*, *124*(2), 362–380.

Roberts, S. O., & Rizzo, M. T. (2021). The psychology of American racism. *American Psychologist*, *76*(3), 475–487.

Robinson, W. R., Renson, A., & Naimi, A. I. (2020). Teaching yourself about structural racism will improve your machine learning. *Biostatistics*, *21*(2), 339–344.

Seaton, E. K., Gee, G. C., Neblett, E., & Spanierman, L. (2018). New directions for racial discrimination research as inspired by the integrative model. *American Psychologist*, *73*(6), 768–780.

Shakespeare, W. (2022). The merchant of Venice: The complete play with annotations, audio and knowledge organisers (CGP Books, Ed.). Coordination Group.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). *No classification without representation: Assessing geodiversity issues in open data sets for the developing world*. https://doi.org/10.48550/ARXIV.1711.08536

Song, Q. C., Shin, H. J., Tang, C., Hanna, A., & Behrend, T. (2022). Investigating machine learning's capacity to enhance the prediction of career choices. *Personnel Psychology*. Advance online publication. https://doi.org/10 .1111/peps.12529

Stock, P., & Cisse, M. (2018). Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.) *Proceedings of the European conference on computer vision (ECCV)* (pp. 498–512). Springer.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. ArXiv. arXiv:1901.10002.

Tost, L. P. (2015). When, why, and how do powerholders "feel the power"? Examining the links between structural and psychological power and reviving the connection between power and responsibility. *Research in Organizational Behavior*, *35*, 29–56.

Tynes, B. M., & Markoe, S. L. (2010). The role of colorblind racial attitudes in reactions to racial discrimination on social network sites. *Journal of Diversity in Higher Education*, *3*(1), 1–13.

Van Miltenburg, E. (2016). *Stereotyping and bias in the flickr30k dataset*. ArXiv. arXiv:1605.06083.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 2053951717743530.

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257.

Wick, M., Swetasudha, P., & Tristan, J.-B. (2019). Unlocking fairness: A trade-off revisited. In *Advances in Neural Information Processing Systems* (Vol. 32). http://papers.nips.cc/paper/9082-unlocking-fairness-a-trade-off-revisited

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100–1122.

Yeomans, M. (2021). A concrete example of construct construction in natural language. *Organizational Behavior and Human Decision Processes*, *162*, 81–94.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, *20*(1), 2737–2778.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In S. Dasgupta, & D. McAllester (Eds.), *International conference on machine learning* (pp. 325–333).

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