Disambiguating Algorithmic Bias: From Neutrality to Justice

Elizabeth Edenberg
elizabeth.edenberg@baruch.cuny.edu
Department of Philosophy, Baruch College, The City University of New York
New York, NY, USA

Alexandra Wood
awood@cyber.harvard.edu
Berkman Klein Center for Internet & Society, Harvard University
Cambridge, MA, USA

ABSTRACT

As algorithms have become ubiquitous in consequential domains, societal concerns about the potential for discriminatory outcomes have prompted urgent calls to address algorithmic bias. In response, a rich literature across computer science, law, and ethics is rapidly proliferating to advance approaches to designing fair algorithms. Yet computer scientists, legal scholars, and ethicists are often not speaking the same language when using the term 'bias.' Debates concerning whether society can or should tackle the problem of algorithmic bias are hampered by conflations of various understandings of bias, ranging from neutral deviations from a standard to morally problematic instances of injustice due to prejudice, discrimination, and disparate treatment. This terminological confusion impedes efforts to address clear cases of discrimination.

In this paper, we examine the promises and challenges of different approaches to disambiguating bias and designing for justice. While both approaches aid in understanding and addressing clear algorithmic harms, we argue that they also risk being leveraged in ways that ultimately deflect accountability from those building and deploying these systems. Applying this analysis to recent examples of generative AI, our argument highlights unseen dangers in current methods of evaluating algorithmic bias and points to ways to redirect approaches to addressing bias in generative AI at its early stages in ways that can more robustly meet the demands of justice.

CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**.

KEYWORDS

algorithms, bias, discrimination, fairness, justice, generative AI, large language models, vision-language models, law, philosophy

ACM Reference Format:
Elizabeth Edenberg and Alexandra Wood. 2023. Disambiguating Algorithmic Bias: From Neutrality to Justice. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3600211.3604695

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604695

1 INTRODUCTION

Algorithms exert influence over an increasingly wide range of social domains, including criminal justice, health care, finance, employment, and education [21, 39, 43, 86, 91, 118, 122, 147]. In both the academic literature and at a societal level, there is a growing awareness of the potential for bias or discrimination in the use of algorithms in sociotechnical systems [21, 22, 28, 37, 57, 109, 115, 119, 137]. When claims of discriminatory effects from such systems arise, public calls to address algorithmic fairness follow closely behind [107], leading to considerable attention from regulatory bodies around the world in recent years [14, 15, 17, 66, 85, 132, 143].

While there may be an emerging consensus that algorithms embedded in sociotechnical systems should be designed to be fair, computer scientists, legal scholars, and ethicists are not speaking the same language when using terms such as bias, fairness, and discrimination [108]. In a 2018 survey, Narayanan highlights twenty-one common technical definitions of fairness, a subset that is illustrative, not exhaustive, of mathematical approaches to bias [111], and Suresh and Guttag observe at least seven distinct sources of downstream harm that can arise at different stages of the machine learning lifecycle [136]. Technical definitions of fairness are often incompatible with one another, and the choice of which to employ when designing or evaluating an algorithm for fairness has an enormous influence on outcomes [42, 60, 97]. Further, the relationships between the various technical definitions of fairness and the legal and ethical notions of antidiscrimination, equality, and justice are not well understood.

These considerations are of vital consequence for legislative and enforcement efforts, such as the EU Artificial Intelligence Act [17], the White House's Blueprint for an AI Bill of Rights [143], and the US Federal Trade Commission's call for businesses to test their algorithms regularly to ensure they do not discriminate on the basis of a protected attribute [85, 132]. When assessing algorithmic bias from a legal or policy perspective, one confronts challenges associated with definition, detection, and enforcement [92]. Longstanding antidiscrimination doctrine, for example, protects individuals against discrimination on the basis of certain protected classes tied to social identities of race, sex, and religion and, accordingly, prohibits the consideration of these protected characteristics in decisions that influence economic opportunity [92]. It is well-recognized, however, that, even in cases where algorithms explicitly exclude protected characteristics in their labeling, such features can continue to influence algorithmically-informed decisions [92]. For example, because race is highly correlated with ZIP code, information about an individual's location can serve as a proxy for this protected characteristic even when race is explicitly excluded from consideration [49].

Given the vast quantities of personal data analyzed by platforms, the fact that they do not explicitly use protected characteristics does little to establish confidence that such characteristics do not influence their results [24, 47]. Further, it is unlikely that, if proxies for protected characteristics do influence the results, they will do so in ways that can be proven to be unlawfully discriminatory without access to substantial additional statistical evidence [92].

In this paper, we analyze current approaches to algorithmic bias with respect to their potential as well as continued challenges for addressing harms to individuals, groups, and society. First (§ 2), we analyze differences between technical, legal, and ethical approaches to defining fairness in order to lay the groundwork for a broader understanding of the underlying harms and the values that, as a society, we should seek to protect in designing and enforcing fair algorithms. Then (§ 3), we outline two common approaches to addressing algorithmic bias: disambiguating different notions of bias and designing for justice. We argue that, while promising, both approaches carry risks: disambiguating bias can neutralize the term in ways that undermine public calls for justice and can also be used to avoid accountability for addressing algorithmic harms. Designing for justice and equity seeks to capture the broad range of social harms but also risks collapsing into debates similar to those that plague the algorithmic fairness literature. Lastly (§§ 4 and 5), we apply our analysis to generative AI to demonstrate methods of identifying and disambiguating bias in such systems, and conclude by suggesting forward-looking approaches to ensuring accountability for algorithmically-driven injustices and inequities.

2 WHAT DOES IT MEAN FOR AN ALGORITHM TO BE FAIR?

The literature reflects a broad range of philosophical, legal, and technical approaches to defining fairness that may provide a basis for designing and evaluating fair algorithms. Notably, philosophical and legal notions provide underpinnings for an expansive view of algorithmic harms and algorithmic justice. However, in the technical literature, concerns about harms from sociotechnical systems are often reduced to measuring various forms of bias in algorithmic results.

2.1 Philosophical and Legal Notions of Justice and Antidiscrimination

John Rawls, one of the most influential philosophers on defining justice, contends in his 1971 *A Theory of Justice* that "justice is the first virtue of social institutions" [125]. Justice is the primary normative criterion we should use in evaluating social institutions and, thus, "laws and institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust" [125]. Rawls argues that principles of justice should apply to the core social institutions that "distribute fundamental rights and duties and determine the division of advantages from social cooperation" [125]. Such political, social, and economic arrangements are the principal focus of justice because they define people's basic rights and duties and "influence their life prospects" [125]. Given the profound influence of algorithms in significant social institutions, including credit, housing, employment, and criminal justice decisions, it is therefore essential to evaluate the justice of algorithmic decisions.

Rawls defends a principle of justice as fairness, not as equivalent terms, but to clarify that principles of justice should specify fair terms of cooperation in society. Determining whether terms are fair is not merely treating similar cases similarly. Instead, Rawls leverages a thought experiment, referred to as the 'original position' [125], to determine which principles could be embraced by people "as free and equal," rather than from a position of domination or subordination [125, 126]. The thought experiment asks people to consider whether they could accept principles of justice no matter where within the social order they fell. Imagine they do not know specifics about their particular situation, but do know general facts about people and society (including facts about racial and gender discrimination, economic inequality, and scarcity). In this scenario, people should choose principles of justice that could be embraced even if they ended up in the least advantaged position in society. If they meet this test and protect the basic dignity of each person, we have some confidence in believing the principles of justice are fair.

Rawls' specific principles of justice have been influential (and controversial) in contemporary theorizing about justice. While applying his theory to algorithms is beyond the scope of this paper, we introduce Rawls because his core methods of thinking about the justice of social institutions illustrate the expansive nature of philosophical approaches that extend beyond applying bias metrics or treating like cases alike. We argue that taking this more expansive view is essential to understanding what is missing when evaluations of fairness are restricted to more limited notions of bias.

Like philosophical notions, antidiscrimination law takes an expansive view of justice in society. Antidiscrimination law developed to address systematic patterns of disadvantage, such as those rooted in the institutions of slavery and Jim Crow segregation in the United States. The Civil Rights Act of 1964 outlawed segregation in public places and prohibited employment discrimination on the basis of race, color, religion, sex, or national origin—with the aim of "dismantling systems of segregation that were endemic to American economic and political systems" [91]. In this way, antidiscrimination law is arguably designed to embody a form of distributive justice, operationalizing principles such as that morally irrelevant characteristics like race and sex should not determine one's opportunities in life [83]. Antidiscrimination law's recognition of harms from discrimination can also be understood through a Rawlsian lens, as rules one might choose in the original position [83]. For instance, the Supreme Court's interpretation of Title VII of the Civil Rights Act of 1964 as prohibiting "not only overt discrimination, but also practices that are fair in form, but discriminatory in operation," including criteria exhibiting a discriminatory preference for or excluding any group and cannot be shown to be related to job performance [6], may reflect what a rational person would choose as a rule for society, not knowing what their own social standing would be [83].

2.2 Technical Measurements of Bias in Algorithms

An expansive and growing number of definitions of algorithmic fairness have been presented and evaluated in the computer science literature [106, 128]. In one survey, Narayanan identifies a broad

collection of fairness definitions, including various notions of statistical bias, group fairness, blindness, individual fairness, process fairness, diversity, and representational harms [111]. Among these, a frequently used fairness definition is statistical bias, i.e., the difference between an estimator's expected value and the true value of the parameter being estimated. Yet, it is widely recognized that statistical bias is inherently limited as a fairness definition because it does not account for biases in the underlying data [111].

For example, the COMPAS risk assessment algorithm employed by many US courts to assess recidivism risk was calibrated to meet fairness defined in terms of statistical bias. It was shown to produce recidivism scores that were only slightly less predictive for black men than white men, seemingly satisfying fairness when understood as statistical bias. However, COMPAS also produced many more false positives among black defendants and false negatives among white defendants due to the data reflecting differences in recidivism prevalence between these groups, which are a product of biases in society [21]. Black men are more likely to live in neighborhoods that have a greater police presence, are subjected to racial profiling by officers, and as a result are re-arrested more often than white men. In this context, the high rate of false positives for black men is particularly concerning, especially when compared to the elevated false negatives for white men, exacerbating the over-incarceration of black men in our criminal justice system [18].

Group fairness definitions measure bias in a model in terms of systematic differences between groups [111]. As one example, the equalized odds definition requires protected and unprotected groups to have equal rates for true positives and false positives [103]. For an algorithm used in support of making loan decisions, for instance, this aims to address potential bias against certain groups that may be learned from the training data, such as that members of historically marginalized groups are denied loans despite being creditworthy. Equalized odds requires ensuring that the fractions of non-defaulters and defaulters approved for loans are equal across groups.

Use of such definitions has limitations, as research has shown that it is impossible to achieve three or more (and, in some cases, even two) group fairness definitions simultaneously [42, 97]. In addition, satisfying even one fairness definition can result in a significant loss in accuracy [45]. For example, in the case of equalized odds, an algorithm must achieve equally high accuracy across all groups, so an algorithm performs only as well as it does on the hardest-to-classify group [75]. Further, where there are disparities in prevalence between groups—resulting, e.g., from measurement bias or historical prejudice—balancing outcomes across different groups requires treating people from different groups differently [23, 103, 111].

The technical literature also introduces tools for mitigating algorithmic bias. Researchers have shown that designing algorithms to be blind to sensitive attributes does not eliminate bias against protected groups, as a sensitive feature such as race may be redundantly encoded in other features such as place of residence [75]. One approach is to explicitly recognize differences in prevalence, such as with Dwork et al.'s *fairness through awareness*, which is based on the principle that "similar individuals should be treated similarly," using a metric that defines how similar two individuals are in the context of a particular decision-making task [55].

2.3 Limitations, Trade-offs, and Gaps Between Definitions

The limitations of technical definitions and the impossibility of satisfying multiple definitions simultaneously requires explicitly addressing the trade-offs between different definitions, as well as between fairness and other considerations such as accuracy [111]. Quantitative definitions also overlook how inequality compounds over time, even through generations, and they cannot resolve conflicts between different values, among other concerns [112]. In practice, the application of such approaches may be limited due to privacy concerns and data minimization policies [93]. Further, there are challenges with respect to measuring bias throughout different stages of the machine learning lifecycle, as measures of bias at one stage may not be reliably correlated with measures in downstream tasks [65].

Some scholars have argued that quantitative approaches are limited in their ability to combat oppression due to being overly formal and limited to isolated decision-making procedures [30, 69, 71, 129]. For example, Green argues that "efforts to formulate mathematical definitions of fairness overlook the contextual and philosophical meanings of fairness" [69] (citing [30, 70, 84, 100, 129]). The various fairness definitions rely on a wide range of understandings of the concept of bias, whether conceptualized as differences between the prediction and the world, different treatment for different groups, human prejudice in the data collection, or other factors [16, 25, 103]. As we will argue below, the wide variety in understandings of bias can impede well-meaning efforts at correcting problematic forms of prejudice, unjust treatment, and discrimination.

Consider, for instance, the relationships between philosophical, legal, and technical definitions of bias in the context of antidiscrimination law. Many scholars have argued that fairness definitions have a role to play in auditing algorithms for evidence of unlawful discrimination (see, e.g., [79, 87]). While antidiscrimination law aims to protect individuals from harmful discrimination stemming from longstanding prejudice, current doctrine is applied more narrowly, in cases where demonstrable harm is shown in a regulated context with a deep history of discrimination, such as employment, housing, education, credit, and public accommodation (see, e.g., [2, 4, 7]). Antidiscrimination law explicitly prohibits discrimination in ads for housing and job opportunities based on protected attributes such as race, sex, age, religion, disability status, and more [1, 3, 5]. This carries through to algorithmic decisions, as recent findings of discrimination have led online platforms to implement changes to address discriminatory targeting and delivery of certain ads [19, 140]. Discrimination may manifest as disparate treatment (see, e.g., [8, 10]), in the case of an algorithm that explicitly considers a protected attribute or where it is intended to classify on the basis of a protected attribute, or as disparate impact (see, e.g., [6, 12]), in the case of an algorithm that has a disproportionate effect on a protected group without a business justification.

Narayanan argues that disparate impact has emerged as the prevailing definition of unintentional algorithmic discrimination in part because it can be readily measured using quantitative tools using existing datasets from a single setting at a single point in time, and that "[i]njustices other than disparate impact seem illegible to regulators" [112]. It is notoriously difficult to establish the intent

behind a human decision, as, for instance, individuals are often not even aware of their own intentions, but it is especially challenging to establish with the rise of algorithmic decision-making. The shift to algorithmic decision-making risks deflecting accountability for addressing harm by companies offsetting their own responsibilities by pointing to the decision-making powers of an algorithmic system—potentially using questions about whether AI systems can have intentions to keep discussions of legal liability stuck in abstract philosophy of mind discussions while avoiding addressing accountability for harms perpetrated through these systems.

Challenges in understanding the relationships between different technical notions of fairness and legal conceptions of fairness, such as those operating within antidiscrimination law [79, 87], do not exhaust the complications in the fairness debate. Crawford argues that most of the technical work on fairness aims to confront instances of allocative harm, leaving unaddressed a category of harms stemming from the use of biased algorithms called representational harms [47].

Allocative harms, according to Crawford, arise when systems allocate or withhold resources or opportunities to people on the basis of their group identity (e.g., when a woman is offered a lower credit limit than her husband despite a shared financial history) [47]. Because allocative harms are discrete, transactional, and easily quantifiable, they lend themselves to technical analysis and intervention through application of the various types of technical definitions of fairness that have been proposed [47]. In contrast, representational harms occur when systems reinforce the subordination of certain groups on the basis of their social identity. Representational harms are difficult to formalize because they are long-term, diffuse, and tied to how people are represented and understood socially [47]. Crawford identifies numerous examples of representational harms such as those involving stereotyping, failures of recognition, harms of denigration, underrepresentation, or ex-nomination (where certain groups are framed as the norm by not giving them names, such as the use of 'athlete' for men vs. 'female athlete' for women) [47].

One way of thinking about allocative harms is through the lens of distributive justice, although this notion extends far beyond the narrow protections in current antidiscrimination law and quantitative definitions of fairness. Likewise, representational harms can be understood through a broader philosophical lens of epistemic injustice, introduced by Fricker [59] to describe ways in which people can be harmed in their capacity as epistemic agents, because these harms of representation reflect and reinforce problematic epistemological frameworks through which we understand and interpret our experiences. Philosophical discussions of justice and fairness aim to capture and analyze both what an ideal theory of justice requires and the ways our existing systems fall short. This broad lens is a useful metric for analyzing instances where algorithmic fairness and antidiscrimination law fall short of their goals.

Applying a philosophical lens for evaluating justice and fairness reveals even broader gaps in the literature. Technical notions of fairness and justice are often "conflated," bearing "the consequence that distributive justice concerns are not addressed explicitly" [98]. Scholars argue that they risk "mirroring some of antidiscrimination discourse's most problematic tendencies" [81] and "often exacerbate oppression and legitimize unjust institutions" [69] (citing [50, 68, 88, 116, 117, 123]). For these reasons, some call for rejecting fairness in favor of alternative frames of justice, equity, or reparation [69]. Bui and Noble argue that "simply striving for fairness in the face of these [unjust] systems of power does little to address" the unjust power structures themselves, and argue for deeply interrogating the underlying power structures and inequalities of such systems [36]. Similarly, D'Ignazio and Klein show how intersectional feminist theories can be applied towards tackling unjust power structures through data science and data ethics [54]. Likewise, Costanza-Chock's principles of design justice call for designers to critically examine how existing practices contribute to the reproduction of systemic oppression and to transform design's values to better meet the aims of social justice [46]. Drawing from intersectional critical theorists, Davis et al. develop a principle of algorithmic reparations, which they argue can name, unmask, and undo both allocative and representational harms in algorithms [50].

We support calls to move beyond discussions of bias to broader notions of justice, but we argue that the challenge is to do so in a way that can actually address unjust power structures. Complex social phenomena and normative goals can be challenging to formalize in ways that can be built into mathematical systems and translated into clear laws and policies [113, 114]. Developing approaches that interface well with both normative and technical understandings will be necessary to ensure protection for individuals, groups, and society, but it must be done with care. It may be appealing for both regulators and technologists to focus on the most readily quantifiable measures of bias, as they can seemingly render abstract problems more concrete. However, there is a risk of narrowing the scope of analysis in ways that can obscure the broader social context that is crucial to understanding algorithmic harm. As Tukey posited with his maxim for data analysis, "[f]ar better an approximate answer to the *right* question, which is often vague, than the *exact* answer to the wrong question, which can always be made precise" [138]. It is critical to focus on developing approaches that address fundamental normative concerns regarding algorithmic harms, even if they might seem vague, rather than focusing on notions of bias just because they lend themselves to quantification and not because they capture what is important.

3 DISAMBIGUATING ALGORITHMIC BIAS: FROM NEUTRALITY TO JUSTICE

Discussions of algorithmic fairness often focus on unpacking specific quantitative definitions of fairness that measure forms of bias arising at certain stages in the development and deployment of algorithmic systems. We refer to the tendency to reduce questions of fairness to discussions of bias metrics as the normative reduction claim. As we outlined in section 2, the reduction of fairness to a bias metric omits broader considerations of justice that the public means to call attention to when critiquing algorithms for the ways they contribute to and perpetuate injustices in society. Although scholars often acknowledge the limitations of normative reduction as tackling a more tractable subset of the problem, the gaps between technical, legal, and ethical approaches to algorithmic bias can undermine even our best efforts to address this problem.

In this section, we discuss current approaches to algorithmic bias and highlight two potential solutions, as well as challenges that arise with each approach. The first approach seeks to disambiguate

different notions of bias. While helpful, this approach lends itself towards neutralizing the term bias in ways that can undermine efforts to address bias and can be used to deflect accountability for addressing algorithmic harms. The second approach seeks to substitute discussions of bias or fairness with discussions of justice, thereby explicitly addressing unjust power structures. This approach is promising insofar as it seeks a broader lens, but, in moving to develop broader discussions of algorithmic justice, theorists must take care not to replicate some of the problems that have beset discussions of algorithmic fairness and bias.

3.1 Current Approaches to Algorithmic Bias

As early as 1996, Friedman and Nissenbaum's normative work on bias identified technical bias, resulting from specific technical constraints, as one of three types of bias that can arise in computer systems [61]. Also important, they argue, are preexisting social biases and emergent biases, which arise out of particular use cases [61]. Since then, increasing attention has been drawn to other significant ways social prejudice can be embedded in data sets, as well as the ways biases can arise when algorithms are trained on data sets that are not representative of the population to which they are applied [56, 115, 119]. Friedman and Nissenbaum call for "freedom from bias" as one of the important criteria by which to judge the acceptability of automated systems [61]. In more recent work, Nissenbaum has disavowed the "seductive diversion" of attempts to "solve bias" in AI systems in ways that can distract from asking whether the systems should be built or used in the first place [123]. Yet within computer science, 'bias' can mean many different things—not all of which can be effectively eliminated given the very nature of algorithmic design. As David Weinberger argues, "bias is machine learning's original sin" because it is embedded into its very essence [142]. By looking for patterns in the data, machine learning systems may find "biased patterns so subtle and complex that they hide from the best-intentioned human attention" [142].

Further, despite the development of a rich body of technical scholarship on algorithmic bias, the term 'bias' is generally not wielded with same degree of precision as other terms used in the computer science literature. Instead, 'bias' is often used as a catch-all to refer to a wide range of behaviors which, in turn, are associated with diverse types of harms, each of which has different types of impacts on different groups of individuals. For instance, in a 2020 analysis of the body of papers on bias in natural language processing, Blodgett et al. found that "the majority of them fail to engage critically with what constitutes 'bias' in the first place," often referring to 'bias' using vague descriptions—or no description at all—and relying instead on unstated assumptions about what makes a system harmful, to whom, and why [33]. Nanayakkara et al. reviewed the broader impact statements for research presented at high-impact AI research conferences and found that, while 'bias' is "frequently mentioned," it is "not always clear whether authors are referring to bias in a societal or technical sense, or whether technical forms of bias are related to social inequalities" [110].

The term 'bias' (or, similarly, 'algorithmic bias') lends more confusion than clarity to the complex array of harms to individuals, groups, and society stemming from the use of algorithms. Algorithmic bias has been used to refer to 'biased' data inputs (including importing social prejudice as well as under- or over-representation of certain groups), 'biased' algorithmic design (including optimization tasks), and 'biases' that result from the algorithms designed in one context being inappropriately used in different contexts [48], tracking the three 'types of bias' Friedman and Nissenbaum highlight [61]. Danks and London go beyond this early work to identify five different meanings of bias: training data bias, algorithmic focus bias, algorithmic processing bias, transfer context bias, and interpretation bias [48]. These biases can also be the result of a deliberate choice, for example, when statistical biases are used to ensure that an algorithm is unbiased relative to a moral standard [48]. Danks and London offer a "taxonomy of different types and sources of algorithmic bias," distinguishing between (i) "neutral or unobjectionable forms of algorithmic bias" and (ii) biases that are "problematic" and therefore demand a response [48]. They argue "there is no coherent notion of 'algorithmic bias"' because the one term refers to statistical, ethical, and legal biases [48]. These different notions of bias can also be separated. It is possible for an algorithm to satisfy technical specifications of fairness (e.g., by offering statistically unbiased predictions) while remaining morally problematic. It is also possible for statistical bias to be morally neutral [58].

Notwithstanding the value in unpacking the many different instances, types, and sources of bias, we argue that the term 'bias' is at best unhelpful and at worst can mask deep injustices. It also can create a false sense that the barriers to addressing bias are insurmountable. This brief survey of meanings of the term illustrates the confusion likely to arise when aiming to mitigate algorithmic bias.

3.2 Two Potential Solutions and Their Challenges

In this section, we highlight two promising approaches to algorithmic bias. One approach, which we call disambiguating algorithmic bias, is to move from broad discussions of bias in favor of identifying the specific notion of bias that is used in a particular instance, as well as the groups affected and where within the lifecycle of the algorithm it occurs. We see examples of this approach when scholars specify the ways bias can arise at different points in the development and deployment of an algorithm [48, 58, 61, 63, 136]. There is great value in specificity in order to track to whom and where the problem occurs, as well as what corrective measures are being used. Efforts to disambiguate and specify the meaning of bias will go a long way towards clarity across disciplines when discussing the wide range of problems and proposed solutions to instances of algorithmic bias. However, this must be done with care, as it also lends itself to using the broad concept of bias in more neutral terms so that it can appropriately capture any instance of deviation from a norm, including technical measures alongside the moral notion.

The second approach to problems arising from the vast array of referents captured by the broad term of algorithmic bias, which we call designing for justice and equity, is to move away from discussions of bias and fairness towards explicit discussions of equity and justice (see, e.g., [36, 46, 50, 67]). We see this approach reflected by scholars who have critiqued the limitations of quantitative measures of bias or fairness in algorithms (see, e.g., [36, 46, 50, 54, 67, 69, 98]).

While we think that there is much to be gained by turning towards more "substantive" understandings of justice and fairness [69], efforts to move in this direction continue to import some of the same challenges that beset efforts to disambiguate questions of bias.

3.2.1 Disambiguating bias—and recognizing the dangers of neutrality. The first approach to disambiguating the range of biases that can arise in algorithms treats bias as a neutral umbrella term capturing any deviation from a norm. Take, for example, Danks and London's argument that sometimes we can use 'neutral' technical forms of bias to help address and correct for the morally problematic forms of bias [48]. Furthermore, they acknowledge the frequent negative connotation of the term 'bias' in English, but they explicitly use the term in "an older and more neutral way" in which "'bias' simply refers to deviation from a standard" [48]. This broader notion in which bias is used to mark deviations from the standard is meant to encapsulate statistical bias (in which estimates deviate from a standard), cases they label "moral bias in which a judgment deviates from a moral norm," and legal, social and psychological biases, all defined in terms of deviation from a norm [48].

However, it is not simply that bias means one (or several) things in the technical literature that are innocently different from the normative uses of the term 'bias' at play when the public expresses concerns of bias. The use of the term 'bias' for all of these different senses can actually hamper our best efforts to try to address problematic forms of injustice, prejudice, and discrimination that underlie public concerns. Most scholars who seek to clarify the many different meanings of algorithmic bias share the normative goal of ensuring that algorithms can live up to our moral standards. However, the broader and more seemingly 'neutral' use of the term bias in all of these instances leads to significant moral confusion. Danks and London suggest that we should avoid calling for an end to algorithmic bias because not all bias is bad and, in fact, some biases are neutral and others can be beneficial to achieving our normative goals, such as in the case where a biased algorithm could be used to "reduce a moral *societal* bias" [48]. Chander, for example, makes the case for designing algorithms to be conscious of protected characteristics, employing algorithmic affirmative action to remedy harms engendered by a "world permeated with the legacy of discriminations past and the reality of discriminations present" [41].

While we agree that additional clarity on how and where deviations from a certain standard arise in the process of designing, training, and deploying algorithms, we do not believe the right move is to neutralize the term bias. Doing so will likely undercut efforts to address real social injustices that can arise. This neutralization of the term bias does help explain why it arises in so many different contexts with so many different implications—but, in so doing, it undercuts the normative force of calls to eliminate bias. After all, if 'bias' is a mere deviation from a standard, bias will never be eliminated and those who resist social change can point to the public naiveté of technical matters and easily dismiss calls for 'unbiased' algorithms as if people were calling for round squares.

Whether or not the public is naive about different statistical measures or how algorithms are optimized to produce the desired result has little bearing on the very real injustices that arise in connection with the pervasive influence of algorithms on various aspects of our modern lives. When the public calls for an end to algorithmic bias, typically this is meant as a call for social justice and to end prejudice, and is tied to long and well-documented histories of racism, sexism, ageism, classism, and other longstanding social prejudices.

Although scholars examining questions of bias or fairness recognize the limitations of technical work and avoid claiming that technical specifications of bias or fairness capture the full complexity of these real-world social problems (see, e.g., [30, 58, 69–71, 81, 84, 100, 129]), standard strategies focus on quantitative measures to identify, quantify, and correct for biases in algorithms in ways that are nevertheless largely divorced from normative understandings of harm [31, 33, 110]. Furthermore, the method of disambiguating bias can also be exploited in ways that can undermine the aims of justice, as we explore through several case studies in the next section.

3.2.2 Designing for justice and equity. The second approach of avoiding the term 'bias' and shifting towards language of equity and justice explicitly seeks to take a broader evaluation of the harms wrought by algorithms in society. However, here too we may collapse into an ever expanding set of debates about how best to specify justice or ways to mathematically formalize and measure philosophical theories of justice. In other words, we risk falling into the same problems that plague the literature on algorithmic bias. For example, when people call for equality, it can quickly lead to debates about what we are trying to make equal and why. We can anticipate that this will repeat the same issues that arise in fairness debates with respect to specifying metrics according to which people should be treated equally. Likewise, for questions of justice: some might worry about how to specify justice mathematically and in a way that is not itself subject to overwhelming disagreement. For both of these claims, the philosophical literature on justice and egalitarianism can provide useful insights—but it will also be easy to conclude there is a lot of continued disagreement and debate about what, e.g., justice requires and how equality should best be measured and protected in society. Such debates may be cited as an excuse to avoid accountability for clear instances of harm.

However, specification of what justice and equality require is worthwhile as a way to get to the heart of the problem in the spirit of Tukey's maxim for data analysis (§2.3). Despite continued debates about which theory of justice is best, there is in fact substantial agreement with respect to some clear instances of harm. As Rawls suggests in his later work seeking to grapple with the continued disagreements in society, in any society that protects freedom of thought and expression, continued disagreement about key normative questions should be expected [126]. Despite continued disagreement, many views are reasonable, and there are a number of rationales by which morally decent people who are reasoning responsibly may come to hold different views. However, this disagreement need not undermine attempts to develop principles of justice that can apply to society broadly. As a society, we can and do find fair terms of social cooperation without requiring everyone to agree to the same moral view.

Rawls argues that there are certain core areas of agreement that any view of justice that could be considered reasonable should be able to capture [125, 126]. Our considered convictions of justice

include ideas like "religious intolerance and racial discrimination are unjust" and, if a theory of justice cannot show why, e.g., racial and gender discrimination are wrong, it should be revised because it fails to capture our considered convictions about justice [125, 126]. Extending this intuition in his later work tackling the broad set of moral and religious disagreements in society [126], Rawls argues that there is substantial agreement that principles of justice should treat people as free, equal moral persons and social institutions should be arranged so that they are substantively (not merely formally) fair. We can leverage these points of agreement to secure legitimate social structures by appealing to these shared areas of agreement (which he calls public reasons) when justifying coercive power. While we cannot expect everyone to agree on every law or policy, their legitimacy depends on whether they are justified in terms of public reasons that can be recognized as reasons of the right kind—i.e., grounded in appeals to freedom, equality, justice, and fairness [126]. We believe a similar lesson can be applied to make progress in addressing clear injustices in algorithmic systems and for adjudicating continued areas of disagreement. Examining algorithmic harms in terms of power imbalances, inequalities, and oppression frame these issues in ways that demand revising unjust structures to better meet the demands of justice.

4 BIAS AND ACCOUNTABILITY IN GENERATIVE AI

Most of the existing literature on algorithmic bias focuses on applications to predictive algorithms, wherein mathematical formulations of bias can be developed, implemented, and tuned for specific tasks and deployments. With the shift towards a new paradigm of generative AI, models are trained on broad data and applied to an extremely wide range of tasks, magnifying the potential for harm. Although the harm to an individual from a single generative AI output might be small, these harms are multiplied dramatically across a large number of users, especially if they are all using the same small set of foundation models [34, 35, 96]. The general public has frequent direct interactions with generative AI models across a broad range of social contexts and the potential for harm is difficult to anticipate. Additionally, generative AI models are used for wide-ranging tasks they are not explicitly trained for, and their characteristics are not well understood. Consequently, their large-scale use presents complex and pressing challenges for addressing algorithmic harms.

Recent scholarship and media coverage of generative AI has uncovered a wide range of examples of harmful representations and associations that are described as evidence of algorithmic bias. In this section, we outline several prominent cases illustrating different types of bias arising in generative AI, including both large language models and vision-language models, to explore what lessons existing approaches to bias developed for the predictive setting have for generative AI. We outline challenges created by the lack of clarity in discussions of algorithmic bias. Disambiguating the kinds of bias involved can be helpful in better identifying and addressing these challenges, but will not go far enough and can too easily be used to deflect accountability for addressing injustices in generative AI. We also show the need to develop tools for evaluating the justice and fairness of algorithms in ways that can capture clear cases of harm, while leaving open productive methods of continued contestation with respect to what justice requires without risking deflection of accountability for algorithmic harms.

4.1 Identifying Bias in Generative AI Models

We highlight a collection of examples from the small but growing body of work exploring harms with respect to large language models and AI image generators, in order to show that, while helpful, disambiguating notions of bias in neutral terminology will not be sufficient guidance to help address normative concerns. We focus on generative AI models because the biases in these models both replicate problems identified by existing research on biased algorithms but also introduce new challenges.

4.1.1 Bias in large language models. Large language models that power popular generative AI services like ChatGPT and Bard have the potential to replicate and amplify existing harmful instances of biased use of language in ways that sustain oppression [26, 27, 51, 53, 72, 94, 141]. The idea that language can be used to perpetuate harm, particularly against marginalized identities, is well established in the philosophical literature on speech and harms (see, e.g., [102]) and has been recognized by law (see, e.g., laws prohibiting hate speech in many countries [9, 13], and the International Criminal Court linking the use of slurs to instances of genocide [11]). A finding from early research on algorithmic bias is that biases in the training set have an enormous influence on the resulting model (see, e.g. [25]). By pulling their training data from the open internet, companies are training the AI systems in ways that amplify racist, misogynistic, and otherwise toxic content that is prevalent on the internet.

Because language models are designed to mirror patterns in natural language, they will predictably encode, reinforce, and perpetuate harmful stereotypes and biases present in the training data, whether due to historical injustice or underrepresentation in a data set [141]. These harms extend beyond allocative harms to, predominantly, more expansive, harder to identify representational harms and instances of epistemic injustice [33]. For example, in focus groups, people with disabilities characterized outputs from large language models as mirroring and reinforcing "perceptions of disability that participants encountered in their lives and dominant media," by emphasizing themes such as visible disability, passivity, lack of autonomy, sadness, and a desire to be "fixed" [62]. Weidinger et al. taxonomize social harms that can arise from large language models producing discriminatory, exclusionary, or toxic language, or performing worse for certain languages and groups [141]. Further, language models have been shown to encode "stereotypical associations," "negative sentiment towards specific groups," and intersectionality effects (i.e., "more bias against identities marginalized along more than one dimension than would be expected based on just the combination of the bias along each of the axes") [27].

4.1.2 Bias in vision-language models. Research has likewise uncovered extensive evidence of bias in vision-language AI models trained on internet-scale data [29, 145, 146]. Data sets used to train vision-language models have been found to contain "troublesome and explicit images and text pairs of rape, pornography, malign

stereotypes, racist and ethnic slurs, and other extremely problematic content" [32]. Vision-language models reflect and magnify this problematic content in various ways that amplify representational harms and epistemic injustices to marginalized populations. Katzman et al. identify and categorize six types of representational harms in image captioning systems: denying people the opportunity to self-identify, reifying social groups, stereotyping, erasing, demeaning, and alienating [90]. As one example, image captioning systems reflect harms of *ex-nomination* [47] by way of a tendency to associate white men, aged 20-59, with being the norm and labeling other groups according to their deviation from this perceived norm [145]. Image captioning systems also demonstrate significant disparities in performance, sentiment, and word choice in captioning of lighter versus darker-skinned individuals [148].

Katzman et al. illustrate how different types of representational harms are in tension with different interventions for bias mitigation in image captioning. For example, they argue that removing potentially sensitive terms such as 'hijab' could result in a system mistagging people wearing hijabs in ways that disrespect or demean them, or in not tagging them at all, "thereby erasing their identities" [90]. Existing research on bias seems to suggest that disambiguating specific instances and types of biases may not go far enough in helping us to mitigate harms caused by these systems. As Zhao et al. highlight, modern systems actually perform less well than older systems, reflecting greater disparities between groups [148]. Despite increased attention to algorithmic bias in recent years, this is not translating into net improvements for marginalized populations.

AI image generation systems appear to amplify these concerns through the ways that vision-language models problematically import and amplify common stereotypes of people. These stereotypes range from who is assumed to hold particular jobs to the reduction of women and girls to sexualized objects. For example, comparing images generated by Stable Diffusion in response to prompts for different professions described by an adjective reveals stereotypes in the model, such as an "assertive firefighter" represented as a white male and a "committed janitor" represented as a person of color [80]. Vision-language models also amplify the sexual objectification of women in society, which Wolfe et al. label *sexual objectification bias*, by "associating images of professional women with sexualized descriptions," "disassociating emotion from images of objectified women," and "generating sexualized images of underage girls" [146]. Harmful associations such as these can influence beliefs and behaviors in real-world contexts, as research has demonstrated that repeated exposure to stereotypical images can be correlated with "discrimination, hostility, and justification of violence against stereotyped peoples" [29] (citing [20, 38, 64, 131]).

4.1.3 Illustration: Bias in the generation of "magic avatars". In 2022, Prisma Labs introduced a "magic avatar" feature for its popular digital retouching app Lensa AI. Employing the open-source Stable Diffusion deep learning model that was trained on a database of over five billion image-text pairs of images and captions scraped from the web, Lensa uses a user's self portraits to retrain the model and generate a collection of digital portraits in different art styles [44].

Reports of social biases, including sexism, misogyny, sexual objectification, racism, and the compounding intersectional nature of oppression, surfaced immediately. One reporter observed that, while her male colleagues' photos were used to generate avatars such as "astronauts" and "fierce warriors," hers, as an Asian woman, generated avatars that were "topless" or with "extremely skimpy clothes and overtly sexualized poses" [77]. Women have long been subject to sexual objectification in society, and this is reflected in online images of women that are sexually objectifying and demeaning. Searching for the term "Asian" on the image databases used to train models such as the one used by Lensa generates results that are "almost exclusively porn" [76, 77]. In similar datasets, the language "an 18 year old girl" is associated with images that "often depict only sexual body parts, with the face omitted, commensurate with findings that objectified female bodies are represented and recognized by their sexual parts" [145, 146]. It is therefore unsurprising that many women have reported similar experiences with Lensa producing highly sexualized avatars based on their photos [104, 133, 134].

Other users have raised concerns over Lensa's tendency to lighten the skin tones and anglicize the features of people of color [133, 134], and to make people's bodies appear thinner [134]. Such representations can contribute to well-documented harms to body image and mental health in connection with social media use, especially for teenage girls. The perception that AI-generated avatars present a "more objective" representation "as if some external, all-knowing being has generated this image of what you [should] look like" has the potential to heighten their impact on a user's body image [95]. Journalists have reported on anecdotal accounts from plastic surgeons and psychologists about patients seeking cosmetic surgery to alter their appearance to more closely resemble their digital avatars, or experiencing distress when confronted with the fact that their actual appearance differs from the AI-tuned photos they have posted on social media [74]. The social biases embedded in such tools can also make it possible for bad actors to easily generate photorealistic nude or otherwise problematic images of a victim using photos often accessible from their social media profile [89]. Drawing from the classifications of harms in [27, 90, 141], these examples engender demeaning representations, stereotypical associations, exclusionary norms, reifying of social groups, intersectionality effects, and worse performance for some groups than others. They also illustrate how harms resulting from biased datasets are magnified and made deeply personal and intimately violating when a generative AI model retrained on an individual's personal data produces harmful images in their likeness.

4.2 Disambiguating Bias and Seeking Accountability in AI

Despite progress on disambiguating various notions of algorithmic bias at different stages of the machine learning lifecycle, a wide range of social biases are persistently magnified and reinforced by generative AI systems. Prior research on fairness in predictive settings provides lessons on the promises and pitfalls of approaches to disambiguating bias and seeking justice that may be instructive towards addressing harms in generative AI systems.

4.2.1 Refusing queries with potentially harmful outputs. In response to public concerns of bias, companies developing generative AI systems have implemented various changes, including, for example, removing offensive content from pre-training datasets and refusing

certain queries that are deemed likely to produce outputs that are explicitly biased or prejudicial [121].

The refusal tactic has been implemented for large language models such as OpenAI's GPT-4 and for vision-language models such as Midjourney's AI image generator. For example, OpenAI seeks to mitigate bias by training for refusals [120], and Midjourney reportedly blocks the use of certain words, such as references to female anatomy, in user prompts, to help prevent the generation of potentially offensive content [78]. Researchers have demonstrated that Midjourney's model learned to associate certain parts of the human anatomy—particularly those related to the female anatomy—with sexual or violent content [78]. Nevertheless, these associations are deeply embedded in the model, and simple workarounds, such as the use of British English spelling, can easily evade the safeguards and produce outputs reflecting such associations [78].

This approach, though useful, is incomplete and backwards-looking—patching instances of problematic outputs as they are discovered. It can also contribute to injustice and inequity, if the model learns from the training for refusal to associate certain marginalized communities with prohibited content [120]. As we will suggest below, we should also adopt a forward-looking approach to designing algorithms that reflect a more just picture of the world we would like rather than piecemeal corrections for the world as it is.

It is critical to identify and address the harm where it operates. Depending on where the bias arises, different interventions may be suitable, such as collecting additional data to balance the training dataset or employing various approaches to measure representational bias in and debias language models [82]. Suresh and Guttag identify seven sources of harm in machine learning, including historical, representation, measurement, aggregation, learning, evaluation, and deployment bias [136]. In the case of AI image generation systems, each harm that is identified could operate at one or more stages. For example, the demeaning representations of Asian women in the Lensa case could reflect *historical bias* due to misogynist and racist beliefs embedded in society and reflected in online content, *representation bias* due to predominantly pornographic representations of Asian women in the training data, *learning bias* if the algorithm amplifies performance disparities between different groups such as Asian women vs. white men, or *evaluation bias* if the bias stems, in part, from the performance of the model being judged with respect to images of white men but underperformance for other groups was not discovered or addressed. This becomes more challenging in generative AI systems actively deployed, whose learning will be subject to malicious actors and prejudices reflected in online content and in the real world. Situating the biases identified in AI systems in the broader social context will help to ensure we are attentive to the broad range of harms and can identify root causes of the harms.

4.2.2 Disambiguating bias but deflecting accountability. Disambiguating bias can help to clarify the types of harms perpetrated, who is impacted, and which aspects of an AI system contribute to this harm. However, it comes with a thus far unrecognized danger—that it can be used to deflect accountability for harm.

One danger is that by pointing to the ways algorithmic bias embeds preexisting social biases into algorithmic systems, the moral problems are deflected away from the technology and instead point back to intractable social problems that have long plagued society. While it is important to acknowledge the persistence of various forms of discrimination and prejudice in the world, the overemphasis on these preexisting biases as the root cause of the problem can reinforce an idea that the AI systems themselves are neutral and are not contributing to the problem.

As Langdon Winner argued in 1980, it is necessary to examine not only the social and economic systems from which a technology arose, but the political qualities of technologies themselves [144]. Often, "the very process of technical development is so thoroughly biased in a particular direction that it regularly produces results counted as wonderful breakthroughs by some social interests and crushing setbacks by others" [144]. In this case, identifying underlying social systems as the cause of bias in AI image generation is only part of the analysis; it is also critical to examine the technology itself, what harms it perpetrates, who is harmed, and how. It is readily foreseeable that a model trained on data scraped indiscriminately from the web will regularly lead to misogynistic and racist imagery. It is likewise foreseeable that white men aged 20-59, who are represented as the norm in vision-language models, would disproportionately receive benefits from this technology, while women and other marginalized groups that are the subjects of denigrating imagery online, would be disproportionately harmed by it.

In response to complaints about pornographic and objectifying images of women, Prisma Labs updated its system to make it more difficult to generate adult-oriented content (i.e., the refusal tactic) and revised its web site to acknowledge the potential for harm to women. It now features a frequently-asked question of "Why do female users tend to get results featuring an over sexualised look?" [99]. They note that "occasional sexualization is observed across all gender categories, although in different ways," and provide an explanation that "[t]he stable Diffusion model was trained on unfiltered Internet content. So it reflects the biases humans incorporate into the images they produce. Creators acknowledge the possibility of societal biases. So do we." [99]. In their acknowledgement of the harm, they point to the social biases embedded in the training data—thereby deflecting the problem from their proprietary algorithm toward the well-known misogyny in the world. Yet the decision to train the algorithm on unfiltered internet content was a deliberate choice on the part of human agents and one that could predictably result in disproportionate harms for marginalized groups given the extensive and well-documented prevalence of racist and misogynistic pornographic content on the web (see, e.g., [115]).

Prisma Lab's framing suggests that its model can only reflect the world as it is. However, technology does not merely shine a neutral mirror on our world. It actively shapes our future by creating new possibilities that extend our imagination about what is possible. This active role is often celebrated by technology innovators—until it attracts negative press. Then, the claim is that the technology is not the cause; it simply reflects broader societal problems.

4.2.3 Technology's role in shaping our future. This brings us to the third core danger we want to highlight. Many discussions of algorithmic bias are set against a false binary that we either have (albeit imperfect) algorithms or the status quo [69]. When presented

with this choice, algorithms can seem preferable. As Miller explains, while algorithms are clearly biased, "the humans they are replacing are significantly more biased" [105]. At least with algorithms, biases can be documented, measured, and adjusted to work toward improving the status quo. There is also evidence that, in some contexts, algorithms can perform better than human decisionmakers at reducing racial disparities and gender inequities [105].

The choice between biased algorithms and the biases of the status quo sets up a false binary grounded in a static picture of the world. However, this relationship is far more dynamic. We can distinguish between two different epistemic contexts from which to assess biases built into algorithms. At times, we can seek to understand the world as it is. For example, when Lensa's or Midjourney's image generation produces predominantly pornographic images of women, this underscores the prevalence of misogynistic images on the web. It can give a new reason to call attention to the deep roots of misogyny and the persistence of this problem in contemporary society.

There is another important epistemic lens that can be adopted by those who create, deploy, and evaluate algorithms—a forward-looking lens that seeks to build a more just future. For interventions to secure algorithmic accountability to effectively address harms in a rapidly evolving landscape of generative AI and other advances, it will be necessary to design sociotechnical systems and regulations with a forward-looking lens rather than to react to instances of harm as they arise. Developers should be aware of social injustice so they can make specific choices for correcting injustice, towards building algorithms that do not reflect the world as it is but instead start to reflect and build the ideals of a more just future.

4.3 Designing Interventions for Accountability

Designing notions of justice that can be embedded in sociotechnical systems and regulations will be critical to ensuring robust interventions for algorithmic accountability. Better training for human labelers, for example, is not enough to ensure robustly just systems, as research has shown that harmful associations along the lines of race, gender, and the intersection of race and gender can be automatically learned by unsupervised vision-language models [135].

Sociotechnical systems require proactive and continual monitoring for algorithmic injustices, and research to develop bias metrics to measure and reduce representational harms such as stereotyping in algorithmic systems is crucial towards developing proactive interventions (see, e.g., [40, 53, 65, 73, 101, 127, 148]). For instance, Dev et al. introduce a framework of representational harms as well as a set of heuristics that can be used to align bias measures in natural language processing with specific harms [52]. Returning to Rawls [126], we can ground conceptions of algorithmic harm in terms of power imbalances, inequalities, and oppression, and demand that sociotechnical systems and the regulatory frameworks that govern their use be designed based on principles of justice—to treat people as free, equal moral persons, and to require social institutions to be arranged so that they are substantively fair.

Although the adoption of sociotechnnical interventions presents numerous practical challenges [128], various approaches that could incorporate a forward-looking lens are emerging, including participatory design, algorithmic auditing, and regulatory oversight.

4.3.1 Participatory design. Because marginalized communities are disproportionately harmed by representational biases, a critical component of interventions for ensuring justice and equity in generative AI is participatory design. Blodgett et al. call for researchers and practitioners to make explicit their normative reasoning and to take into account the "lived experiences of members of communities affected by NLP systems" [33]. In the case of Lensa, conversations with marginalized groups could have promptly alerted the developers to the predictable misogyny and racism that runs rampant on the open internet and is reflected in the generated images. As an illustration of the value of user participation, Gadiraju et al. demonstrated how focus groups with people with disabilities readily surfaced a wide range of ways in which outputs from a large language model, while not producing blatantly offensive outputs, mirrored subtle stereotypes that people with disabilities encounter in their daily lives and in popular media [62]. A better understanding of how communities experience oppression could point to different choices for data curation and model training to reflect ideals of gender and racial equality rather than marginalization and gender-based violence. With deliberate design choices informed by the lived experiences of marginalized groups, generative AI applications could work as well for Asian women as they do for their white male colleagues, and as well for persons with disabilities as for persons without disabilities.

Awareness of the importance of community involvement in the design of public-facing algorithmic systems is growing. In OpenAI's February 2023 response to user concerns about "outputs that they consider politically biased, offensive, or otherwise objectionable," the company announced plans to solicit public input from "as many perspectives as possible" and invest in research and engineering to address bias with improvements based on feedback from the user community [121]. Inviting public input is a step in the right direction, but it is often employed in ways that are more backwards looking to correct errors in deployment rather than forward-looking to anticipate and design more robustly just products.

4.3.2 Algorithmic auditing. Regular auditing and continuous monitoring is another component that is critical to ensuring they meet the demands of justice. One approach is to invite feedback from people who are using these algorithms in a variety of cultural setting. As Shen et al. highlight, regular auditing by everyday users who engage with algorithmic systems in real-world social and cultural contexts is crucial because these users are well-situated to detect algorithmic harms that may go unnoticed by design teams [130]. There are two ways to view algorithmic auditing: this auditing can be used to ensure compliance with clear standards as a method of enforcement, but it can also be used as a forward-looking mode of oversight. Researchers have called for internal audits of algorithms for adherence to ethical standards prior to deployment and for continually monitoring the model throughout its lifecycle [124]. Frameworks for ensuring justice and equity in sociotechnical systems, informed by legal requirements and philosophical theory, could be incorporated in the standards used in such processes.

4.3.3 Regulatory oversight. Another promising model for a forward-looking approach to oversight could put the burden on companies to prove the safety, efficacy, and adherence to justice to a regulatory body before it is deployed or released. This model is inspired

by the U.S. Food and Drug Administration (see, e.g., [139]), which requires pharmaceutical companies to prove that a drug meets strict standards for safety and efficacy before it is approved for use. Adopting a similar model for algorithms and generative AI systems could require organizations developing and deploying the system to demonstrate their safety, efficacy, and adherence to requirements for justice and equity prior to releasing these tools to the public. Stricter standards could be adopted for algorithms designed for use in highly-consequential domains or otherwise expected to have a significant impact on a large number of people.

5 CONCLUSION

Conversations around how to design, implement, and evaluate fair algorithms are impeded by a lack of common understanding of the term 'bias.' One step researchers and practitioners could take is to disambiguate the term bias and adopt instead a wider range of terminology, such as prejudice, discrimination, and statistical weighting, that more accurately expresses when and which types of injustices occur. Yet, even when researchers are precise in locating the specific harm, there is a real danger this can be used to deflect accountability away from the algorithm and its developers. Injustices persist in both the world and in the algorithms that reflect and amplify societal harms. But this need not mean we can hope for no better. We call attention to the interplay between modeling the world as it is and promoting a more just and equitable social order, and argue that the design and use of algorithms has a role to play in both aspects.

ACKNOWLEDGMENTS

The writing of this paper was partially supported by a gift to the McCourt School of Public Policy and Georgetown University. The authors thank Sílvia Casacuberta Puig, Aloni Cohen, Jörg Drechsler, Ayelet Gordon-Tapiero, Kobbi Nissim, Patrick Schenk, Leah von der Heyde, James Williams, the members of the Bridging Privacy Working Group, and the participants of the 15th annual Privacy Law Scholars Conference (PLSC 2022) for their helpful feedback on early drafts of this paper.

REFERENCES

[1] 1964. *Section 2000e-3(b) of Title VII of the Civil Rights Act of 1964.* 42 U.S.C. § 2000e-3(b).
[2] 1964. *Title VII of the Civil Rights Act of 1964.* 42 U.S.C. § 2000e et seq.
[3] 1967. *Section 623(e) of Age Discrimination in Employment Act of 1967.* 29 U.S.C. § 623(e).
[4] 1968. *Fair Housing Act.* 42 U.S.C. § 3601 et seq.
[5] 1968. *Section 3604(c) of the Fair Housing Act.* 42 U.S.C. § 3604(c).
[6] 1971. *Griggs v. Duke Power Co.* 401 U.S. 424 (1971).
[7] 1974. *Equal Credit Opportunity Act.* 15 U.S.C. § 1691 et seq.
[8] 1976. *Washington v. Davis.* 426 U.S. 229 (1976).
[9] 1986. *Public Order Act 1986 (c 64).* Parts III and 3A (UK).
[10] 1995. *Adarand Constructors, Inc. v. Pena.* 515 U.S. 200 (1995).
[11] 2007. *The media and the Rwanda genocide.* Pluto Press ; Fountain Publishers ; International Development Research Centre, London ; Ann Arbor, MI : Kampala, Uganda : Ottawa.
[12] 2009. *Ricci v. DeStefano.* 557 U.S. 557 (2009).
[13] 2015. *Criminal Code of Germany.* § 130 (Volksverhetzung) (Germany).
[14] 2019. *Algorithmic Accountability Act of 2019.* S.1108, 116th Cong.
[15] 2020. *Data Accountability and Transparency Act of 2020.* S.____, 116th Cong. (Discussion Draft).
[16] 2021. Algorithmic Bias in Education. , 1052–1092 pages.
[17] 2021. *European Commission.* Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM/2021/206 final).
[18] Michelle Alexander. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* The New Press, New York, NY.
[19] American Civil Liberties Union. 2019. In Historic Decision on Digital Bias, EEOC Finds Employers Violated Federal Law when they Excluded Women and Older Workers from Facebook Ads. https://www.aclu.org/press-releases/historic-decision-digital-bias-eeoc-finds-employers-violated-federal-law-when-they Press release.
[20] David Amodio and Patricia Devine. 2006. Stereotyping and Evaluation in Implicit Race Bias: Evidence for Independent Constructs and Unique Effects on Behavior. *Journal of personality and social psychology* 91 (11 2006), 652–61. https://doi.org/10.1037/0022-3514.91.4.652
[21] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (23 May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[22] Julia Angwin and Terry Parris, Jr. 2016. Facebook Lets Advertisers Exclude Users by Race. *ProPublica* (28 October 2016). https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race
[23] Solon Barocas. 2017. What is the Problem to Which Fair Machine Learning is the Solution?. Presentation at AI Now. (10 July 2017). https://ainowinstitute.org/symposia/videos/what-is-the-problem-to-which-fair-machine-learning-is-the-solution.html
[24] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. *Special Interest Group for Computing, Information and Society* (2017).
[25] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities.* fairmlbook.org. http://www.fairmlbook.org.
[26] Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. arXiv:2010.14534 [cs.CL]
[27] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. 610–623. https://doi.org/10.1145/3442188.3445922
[28] Yochai Benkler, Rob Faris, and Harold Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics.* Oxford University Press, New York, NY.
[29] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. https://arxiv.org/abs/2211.03759
[30] Reuben Binns. 2018. Fairness in machine learning: lessons from political philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (2018).
[31] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. arXiv:2106.15590 [cs.LG]
[32] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *CoRR* abs/2110.01963 (2021). arXiv:2110.01963 https://arxiv.org/abs/2110.01963
[33] Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP.
[34] Rishi Bommasani et al. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258
[35] Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? arXiv:2211.13972
[36] Matthew Le Bui and Safiya Umoja Noble. 2020. We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness. In *The Oxford Handbook of Ethics of AI.*
[37] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html
[38] Diana Burgess, Yingmei Ding, Margaret Hargreaves, Michelle van Ryn, and Sean Phelan. 2008. The Association between Perceived Discrimination and Underutilization of Needed Medical and Mental Health Care in a Multi-Ethnic Community Sample. *Journal of health care for the poor and underserved* 19 (09 2008), 894–911. https://doi.org/10.1353/hpu.0.0063
[39] Jenna Burrell and Marion Fourcade. 2021. The Society of Algorithms. *Annual Review of Sociology* 47, 1 (2021), 213–237. https://doi.org/10.1146/annurev-soc-090820-020800 arXiv:https://doi.org/10.1146/annurev-soc-090820-020800
[40] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR* abs/1608.07187 (2016). arXiv:1608.07187 http://arxiv.org/abs/1608.07187

[41] Anupam Chander. 2017. The Racist Algorithm? *Michigan Law Review* 115 (2017), 1023–1045.

[42] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[43] Danielle Keats Citron and Frank A. Pasquale. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89 (2014), 1–33.

[44] Nicole Clark. 2022. Lensa's viral AI art creations were bound to hypersexualize users: AI-generated art is rife with issues. *Polygon* (20 December 2022). https://www.polygon.com/23513386/ai-art-lensa-magic-avatars-artificial-intelligence-explained-stable-diffusion

[45] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *CoRR* abs/1701.08230 (2017). http://arxiv.org/abs/1701.08230

[46] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need.* MIT Press, Cambridge, MA.

[47] Kate Crawford. 2017. The Trouble with Bias. Keynote address. *Neural Information Processing Systems* (2017). https://www.youtube.com/watch?v=fMym_BKWQzk

[48] David Danks and Alex J. London. 2017. Algorithmic Bias in Autonomous Systems. *Proc. 26th Int'l Joint Conf. on Artificial Intelligence*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

[49] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Proxy Non-Discrimination in Data-Driven Systems. https://arxiv.org/abs/1707.08120

[50] Jenny L. Davis, Apryl Williams, and Michael W. Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2 (2021). https://doi.org/10.1177/20539517211044808

[51] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. arXiv:2108.12084 [cs.CL]

[52] Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022.* Association for Computational Linguistics, Online only, 246–267. https://aclanthology.org/2022.findings-aacl.24

[53] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21).* Association for Computing Machinery, New York, NY, USA, 862–872. https://doi.org/10.1145/3442188.3445924

[54] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* MIT Press, Cambridge, MA.

[55] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2011. Fairness Through Awareness. *CoRR* abs/1104.3913 (2011). arXiv:1104.3913 http://arxiv.org/abs/1104.3913

[56] Virginia Eubanks. 2017. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, New York, NY.

[57] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, Inc., USA.

[58] Sina Fazelpour and David Danks. 2021. Algorithmic Bias: Senses, sources, solutions. *Philosophy Compass*, 1–16. https://doi.org/10.1111/phc3.12760

[59] Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford University Press, New York, NY.

[60] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Commun. ACM* 64, 4 (March 2021), 136–143. https://doi.org/10.1145/3433949

[61] Batya Friedman and Helen Fay Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14 (1996), 330–347. Issue 3.

[62] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models.

[63] Bruce Glymour and Jonathan Herington. 2019. Measuring the Biases That Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19).* Association for Computing Machinery, New York, NY, USA, 269–278. https://doi.org/10.1145/3287560.3287573

[64] Phillip Goff, Jennifer Eberhardt, Melissa Williams, and Matthew Jackson. 2008. Not Yet Human: Implicit Knowledge, Historical Dehumanization, and Contemporary Consequences. *Journal of personality and social psychology* 94 (03 2008), 292–306. https://doi.org/10.1037/0022-3514.94.2.292

[65] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. 1926–1940. https://doi.org/10.18653/v1/2021.acl-long.150

[66] Government of Canada. 2021. Directive on automated decision-making. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

[67] Ben Green. 2018. Putting the J(ustice) in FAT. *Berkman Klein Center Collection - Medium* (26 February 2018). https://medium.com/berkman-klein-center/putting-the-j-ustice-in-fat-28da2b8eae6d

[68] Ben Green. 2020. The false promise of risk assessments: epistemic reform and the limits of fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2020). https://doi.org/10.1145/3351095.3372869

[69] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35, 90 (2022).

[70] Ben Green and Lily Hu. 2018. The myth in the methodology: towards a recontextualization of fairness in machine learning. *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning* (2018).

[71] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2020).

[72] Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21).* Association for Computing Machinery, New York, NY, USA, 122–133. https://doi.org/10.1145/3461702.3462536

[73] Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Dublin, Ireland, 1012–1023. https://doi.org/10.18653/v1/2022.acl-long.72

[74] Anna Haines. 2022. How AI Avatars And Face Filters Are Altering Our Conception Of Beauty. *Forbes* (19 December 2022). https://www.forbes.com/sites/annahaines/2022/12/19/how-ai-avatars-and-face-filters-are-affecting-our-conception-of-beauty/

[75] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[76] Melissa Heikkilä. 2022. The Algorithm: AI-generated art raises tricky questions about ethics, copyright, and security. *MIT Technology Review* (20 September 2022). https://www.technologyreview.com/2022/09/20/1059792/the-algorithm-ai-generated-art-raises-tricky-questions-about-ethics-copyright-and-security/

[77] Melissa Heikkilä. 2022. The viral AI avatar app Lensa undressed me—without my consent. *MIT Technology Review* (12 December 2022). https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

[78] Melissa Heikkilä. 2023. AI image generator Midjourney blocks porn by banning words about the human reproductive system. *MIT Technology Review* (24 February 2023). https://www.technologyreview.com/2023/02/24/1069093/ai-image-generator-midjourney-blocks-porn-by-banning-words-about-the-human-reproductive-system/

[79] Deborah Hellman. 2020. Measuring Algorithmic Fairness. *Virginia Law Review* 106 (2020), 811–866. https://virginialawreview.org/articles/measuring-algorithmic-fairness/

[80] Justin Hendrix. 2022. Researchers Find Stable Diffusion Amplifies Stereotypes. *Tech Policy Press* (9 November 2022).

[81] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

[82] Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass* 15, 8 (2021), e12432. https://doi.org/10.1111/lnc3.12432 arXiv:https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12432

[83] Julie Chi hye Suk. 2006. Antidiscrimination Law in the Administrative State. *University of Illinois Law Review* 2006 (2006), 405–474.

[84] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (2021), 375–385.

[85] Elisa Jillson. 2021. Aiming for truth, fairness, and equity in your company's use of AI. *Federal Trade Commission Business Blog* (19 April 2021).

[86] Kristin Johnson, Frank Pasquale, and Jennifer Chapman. 2019. Artificial Intelligence, Machine Learning, and Bias In Finance: Toward Responsible Innovation. *Fordham Law Review* 88, 2 (2019), 499–529.

[87] Sendhil Mullainathan Cass R. Sunstein Jon Kleinberg, Jens Ludwig. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2018), 113–174.

[88] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* (7 July 2020). https://www.nature.com/articles/d41586-020-02003-2

[89] Haje Jan Kamps. 2022. It's way too easy to trick Lensa AI into making NSFW images. *TechCrunch* (6 December 2022). https://techcrunch.com/2022/12/06/lensa-goes-nsfw

[90] Jared Katzman, Solon Barocas, Su Lin Blodgett, Kristen Laird, Morgan Klaus Scheuerman, and Hanna Wallach. 2021. Representational Harms in Image Tagging. In *Beyond Fair Computer Vision Workshop at CVPR 2021.*

[91] Pauline T. Kim. 2020. Manipulating Opportunity. *Virginia Law Review* 106 (2020), 867–935.

[92] Pauline T. Kim and Sharion Scott. 2018. Discrimination in Online Employment Recruiting. *St. Louis University Law Journal* 63 (2018), 93–118.

[93] Jennifer King, Daniel Ho, Arushi Gupta, Victor Wu, and Helen Webley-Brown. 2023. The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 492–505. https://doi.org/10.1145/3593013.3594015

[94] Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. arXiv:2102.04130 [cs.CL]

[95] Miles Klee. 2022. A Psychologist Explains Why Your 'Hot AI Selfies' Might Make You Feel Worse. *Rolling Stone* (12 December 2022). https://www.rollingstone.com/culture/culture-features/lensa-app-hot-ai-selfie-self-esteem-1234644965/

[96] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118. https://doi.org/10.1073/pnas.2018340118 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2018340118

[97] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR* abs/1609.05807 (2016). arXiv:1609.05807 http://arxiv.org/abs/1609.05807

[98] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. 2022. From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology* 7 (2022). https://doi.org/10.3389/fsoc.2022.883999

[99] Prisma Labs. 2023. Lensa's Magic Avatars Explained. *Live FAQ* (2023). https://prismalabs.notion.site/prismalabs/Lensa-s-Magic-Avatars-Explained-c08c3c34f75a42518b8621cc89fd3d3f [https://perma.cc/E65L-YT3A] (last visited Mar. 6, 2023).

[100] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2021. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics* 1 (2021), 529–544.

[101] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards Understanding and Mitigating Social Biases in Language Models. arXiv:2106.13219 [cs.CL]

[102] Ishani Maitra and Mary Kate McGowan (Eds.). 2012. *Speech and Harm: Controversies over Free Speech.* Oxford University Press.

[103] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning.

[104] Mia Mercado. 2022. Why Do All My AI Avatars Have Huge Boobs? *The Cut* (12 December 2022). https://www.thecut.com/2022/12/ai-avatars-lensa-beauty-boobs.html

[105] Alex P. Miller. 2018. Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review* (26 July 2018). https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

[106] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902 arXiv:https://doi.org/10.1146/annurev-statistics-042720-125902

[107] Loveday Morris, Elizabeth Dwoskin, and Hamza Shaban. 2021. Whistleblower Testimony and Facebook Papers Trigger Lawmaker Calls for Regulation. *Washington Post* (25 October 2021). https://www.washingtonpost.com/technology/2021/10/25/facebook-papers-live-updates

[108] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 119 (November 2019), 36 pages. https://doi.org/10.1145/3359221

[109] Cecilia Muñoz, Megan Smith, and DJ Patil. 2016. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.* Technical Report. Executive Office of the President, Washington, DC. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

[110] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 795–806. https://doi.org/10.1145/3461702.3462608

[111] Arvind Narayanan. 2018. 21 Fairness Definitions and Their Politics. *Tutorial for Conf. Fairness, Accountability & Transparency* (23 February 2018). https://www.youtube.com/watch?v=jIXIuYdnyyk

[112] Arvind Narayanan. 2022. The limits of the quantitative approach to discrimination. *2022 James Baldwin lecture, Princeton University* (11 October 2022). https://www.cs.princeton.edu/~arvindn/talks/baldwin-discrimination/baldwin-discrimination-transcript.pdf

[113] Kobbi Nissim and Alexandra Wood. 2018. Is privacy *privacy*? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (08 2018), 20170358. https://doi.org/10.1098/rsta.2017.0358

[114] K. Nissim and A. Wood. 2021. Foundations for Robust Data Protection: Co-designing Law and Computer Science. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA).* IEEE Computer Society, Los Alamitos, CA, USA, 235–242. https://doi.org/10.1109/TPSISA52974.2021.00026

[115] Safiya Umoja Noble. 2018. *Algorithms of oppression. How search engines reinforce racism.* New York University Press, New York. http://algorithmsofoppression.com/

[116] Rodrigo Ochigame. 2020. The Long History of Algorithmic Fairness. *Phenomenal World* (30 January 2020). https://www.nature.com/articles/d41586-020-02003-2

[117] Rodrigo Ochigame, Chelsea Barabas, Karthik Dinakar, Madars Virza, and Joichi Ito. 2018. Beyond Legitimation: Rethinking Fairness, Interpretability, and Accuracy in Machine Learning. *International Conference on Machine Learning* (2018).

[118] Ofqual. 2020. Awarding GCSE, AS, A Level, Advanced Extension Awards and Extended Project Qualifications in Summer 2020: Interim Report. (2020).

[119] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown, New York, NY.

[120] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[121] OpenAI. 2023. How should AI systems behave, and who should decide? *OpenAI Blog* (16 February 2023). https://openai.com/blog/how-should-ai-systems-behave

[122] Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information.* Harvard University Press, Cambridge, MA.

[123] Julia Powles and Helen Nissenbaum. 2018. The seductive diversion of 'solving' bias in artificial intelligence. *Medium* (7 December 2018). https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

[124] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. arXiv:2001.00973 [cs.CY]

[125] John Rawls. 1971. *A Theory of Justice.* Belknap Press of Harvard University Press, Cambridge, Mass.

[126] John Rawls. 2005. *Political Liberalism: Expanded Edition.* Columbia University Press, New York.

[127] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. arXiv:2104.06001 [cs.CL]

[128] Reva Schwartz, Apostol Vassilev, Kristen K. Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. https://doi.org/10.6028/NIST.SP.1270

[129] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (2019), 59–68.

[130] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (oct 2021), 1–29. https://doi.org/10.1145/3479577

[131] Morgan P. Slusher and Craig A. Anderson. 1987. When reality monitoring fails: The role of imagination in stereotype maintenance. *Journal of Personality and Social Psychology* 52 (04 1987), 653–662. https://doi.org/10.1037//0022-3514.52.4.653

[132] Andrew Smith. 2020. Using Artificial Intelligence and Algorithms. *Federal Trade Commission Business Blog* (8 April 2020).

[133] Olivia Snow. 2022. 'Magic Avatar' App Lensa Generated Nudes From My Childhood Photos. *Wired* (7 December 2022). https://www.wired.com/story/lensa-artificial-intelligence-csem/

[134] Zoe Sottile. 2022. What to know about Lensa, the AI portrait app all over social media. *CNN Style* (11 December 2022). https://www.cnn.com/style/article/lensa-ai-app-art-explainer-trnd/index.html

[135] Ryan Steed and Aylin Caliskan. 2021. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM. https://doi.org/10.1145/3442188.3445932

[136] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm in the Machine Learning Life Cycle. *Proc. ACM Equity & Access in Algorithms, Mechanisms & Optimization* (2021). http://doi.org/10.1145/3465416.

3483305
[137] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising. *Queue* 11, 3 (March 2013), 10–29. https://doi.org/10.1145/2460276.2460278
[138] John W. Tukey. 1962. The Future of Data Analysis. *The Annals of Mathematical Statistics* 33, 1 (1962), 1 – 67. https://doi.org/10.1214/aoms/1177704711
[139] Andrew Tutt. 2017. An FDA for Algorithms. *Administrative Law Review* 69 (2017), 83–123.
[140] U.S. Department of Housing and Urban Development. 2019. Charge of Discrimination, FHEO No. 01-18-0323-8.
[141] Laura Weidinger et al. 2021. Ethical and social risks of harm from Language Models. *DeepMind Report* (2021).
[142] David Weinberger. 2019. How Machine Learning Pushes Us to Define Fairness. *Harvard Business Review* (6 November 2019). https://hbr.org/2019/11/how-machine-learning-pushes-us-to-define-fairness
[143] White House Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf
[144] Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus* 109, 1 (1980), 121–136.
[145] Robert Wolfe and Aylin Caliskan. 2022. Markedness in Visual Semantic AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (2022).
[146] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2022. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. (2022). https://arxiv.org/abs/2212.11261
[147] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719.
[148] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. arXiv:2106.08503 [cs.CV]

Received 15 March 2023; revised 15 March 2023; accepted 15 March 2023