An Epistemic Lens on Algorithmic Fairness

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

In this position paper, we introduce a new epistemic lens for analyzing algorithmic harm. We argue that the epistemic lens we propose herein has two key contributions to help reframe and address some of the assumptions underlying inquiries into algorithmic fairness.

First, we argue that using the framework of epistemic injustice helps to identify the root causes of harms currently framed as instances of representational harm. We suggest that the epistemic lens offers a theoretical foundation for expanding approaches to algorithmic fairness in order to address a wider range of harms not recognized by existing technical or legal definitions.

Second, we argue that the epistemic lens helps to identify the epistemic goals of inquiries into algorithmic fairness. There are two distinct contexts within which we examine algorithmic harm: at times, we seek to understand and describe the world as it is, and, at other times, we seek to build a more just future. The epistemic lens can serve to direct our attention to the epistemic frameworks that shape our interpretations of the world as it is and the ways we envision possible futures. Clarity with respect to which epistemic context is relevant in a given inquiry can further help inform choices among the different ways of measuring and addressing algorithmic harms. We introduce this framework with the goal of initiating new research directions bridging philosophical, legal, and technical approaches to understanding and mitigating algorithmic harms.

CCS CONCEPTS

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

KEYWORDS

algorithms, bias, discrimination, fairness, epistemic injustice

ACM Reference Format:
Elizabeth Edenberg and Alexandra Wood. 2023. An Epistemic Lens on Algorithmic Fairness. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30–November 01, 2023, Boston, MA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3617694.3623248

1 INTRODUCTION

In recent years, the potential for algorithms to influence highly consequential social domains in unfair and unjust ways has come into sharp relief [17, 19, 22, 40, 54, 60, 62]. High-profile controversies have brought attention to the role of algorithms in criminal justice decisions [10, 11], eligibility determinations for public assistance programs [30], and prejudicial search engine results [57, 64], among other contexts. Risks of discriminatory effects of the use of such algorithms have motivated the development of an expansive body of technical scholarship investigating algorithmic fairness [21, 23, 29, 33, 37, 46, 51]. Much of the technical research has concentrated around advancing fairness metrics for evaluating algorithms, but there has been substantially less progress made with respect to establishing connections between mathematical notions and the legal and philosophical foundations of fairness.

1.1 Understanding Algorithmic Harms: Allocative vs. Representational Harms

Scholars have observed that legal and technical definitions of fairness generally address a subset of algorithmic harms that are understood as allocative harms [14, 24]. Allocative harms arise when systems allocate or withhold resources or opportunities on the basis of one's group identity—for example, when a woman is offered a lower credit limit than her husband despite a shared financial history. Current antidiscrimination doctrine is applied rather narrowly in cases of demonstrable allocative harm in the context of employment, housing, education, and credit decisions (see, e.g., [2, 5, 7]), though the law's more expansive roots aim to protect people from harmful social discrimination based on longstanding prejudice. Because allocative harms are discrete, transactional, and easily quantifiable, they also lend themselves more readily to technical analysis and intervention through application of the various types of technical definitions of fairness that have been proposed in the literature (see, e.g., [29, 55, 56, 63]).

Left unaddressed by most definitions are the underlying representational harms that arise when systems reinforce the subordination of groups on the basis of their social identity [14, 24, 42]. A prominent example was identified in 2015 by a user of the Google Photos app, who called attention to the image recognition software's labeling of his Black friends as gorillas [15]. Another example has been demonstrated with respect to language translation software that imports gendered assumptions into the translation, such as translating "she is a doctor" and "he is a nurse" into a gender neutral language and then back into English producing "he is a doctor" and "she is a nurse" [20]. Unlike discrete allocative harms, representational harms are difficult to formalize because they are long-term, diffuse, and produce harms "upstream" in terms of how people are represented and understood socially [24]. Kate 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

EAAMO '23, October 30–November 01, 2023, Boston, MA, USA

ACM ISBN 979-8-4007-0381-2/23/10.
https://doi.org/10.1145/3617694.3623248

framed as the norm by not giving them names, such as the use of "athlete" vs. "female athlete") [24]. Understanding and addressing this class of harms requires a broader understanding of the social context and history of discrimination and marginalization of certain communities.

1.2 A New Epistemic Lens for Analyzing Algorithmic Harms

Prior research distinguishing between allocative and representational harms has brought to light a category of harms not captured by many existing technical definitions of fairness. We argue that developing approaches to address this gap will require a theoretical foundation for evaluating and understanding representational harms. To do so, we propose a new epistemic lens for analyzing algorithmic harms and demonstrate its application through an analysis of representational harms. We argue that this epistemic lens can help us understand representational harms because many such harms concern the ways in which prejudice influences the assumptions people make about others on the basis of their social identity. This prejudice manifests not only in harms related to resources and opportunities, but also in the way we build our epistemic frameworks through which we make sense of the world [27, 34, 61, 65].

As the basic mental structures that underlie our ability to gain knowledge and find meaning [34, 61], epistemic frameworks operate in the background to help us interpret new evidence and experiences. Some epistemic frameworks may be generally harmless (such as those enabling our ability to interpret cause and effect relationships), while others are more pernicious. When prejudicial systems influence the ways we interpret the world and find meaning, this is a case of epistemic injustice because it is an injustice done to us in our capacity as knowers [32]. We argue that many representational harms are instances of epistemic injustice [32, 53] and that this framing helps to explain why these types of harms are particularly resilient and difficult to address [27, 65]. Further, we argue that identifying the root causes of algorithmic harms is essential to ensuring that changes to sociotechnical systems address problems at their source rather than merely shift them to new domains. In philosophical terms, this implicates two intimately related challenges for securing justice. Prejudicial discrimination on the basis of social identities is both a problem of distributive injustice (relating to allocative harms) and a problem of epistemic injustice (relating to representational harms). While prejudicial discrimination underlies both, epistemic injustice [32, 53] captures more precisely how prejudice influences the ways we view and the assumptions we make about people on the basis of their social identity [27, 61, 65].

1.3 The Goals of This Position Paper

This paper introduces a new epistemic lens for analyzing algorithmic harms and demonstrates how this lens can reframe the algorithmic fairness literature in ways that can help us make progress towards addressing algorithmic harms. We begin by outlining the framework (§ 2), where we define epistemic frameworks, epistemic injustice, and epistemic goals.

Then, we argue that the epistemic lens we propose herein has two key contributions to help reframe some of the assumptions underlying inquiries into algorithmic fairness. First, we argue that using the framework of epistemic injustice helps to identify the underlying causes of harms currently framed as instances of representational harm. Analyzing algorithmic systems through the epistemic lens (§ 3) can also help provide a basis for extending protections against a wider class of harms and ensuring that interventions target the source of the problem. Second, we argue that the epistemic lens helps us identify distinct epistemic goals of inquiries into algorithmic fairness. These distinct aims suggest new pathways for addressing algorithmic harm (§ 4). We conclude (§ 5), by suggesting that a full application of this epistemic lens to existing technical measures of algorithmic fairness be a direction for future research.

Throughout this position paper, we highlight the interplay between epistemic frameworks for evaluating algorithmic fairness, existing scholarship on allocative and representational harms, their relationships to legal theories of antidiscrimination and antisubordination, and the role that the design and use of algorithms could play both in understanding the world as it is and in building a world that is more just and equitable.

2 A NEW EPISTEMIC LENS

Drawing on the social epistemology literature, we introduce the idea of epistemic frameworks as a helpful tool for analyzing algorithmic harms. We argue that this epistemic lens offers a new way to characterize a form of injustice produced by algorithmic harms: epistemic injustice. Analyzing algorithmic harms through the lens of epistemic injustice captures categories of harm that are not currently addressed by existing definitions of fairness and discrimination. It also helps to reveal the root causes of what makes particular representations harmful and, therefore, helps to explain why representational harms have been difficult to address through existing measures of algorithmic fairness. Further, we direct our epistemic lens to consider the aims of inquiries into algorithmic fairness and to distinguish between two distinct epistemic contexts relevant to analyzing and addressing algorithmic harms through three categories of interventions: investigation, substantiation, and amelioration. Clarity with respect to which epistemic context is relevant in a given inquiry can help inform choices among different ways of measuring and addressing algorithmic harm.

2.1 Epistemic Frameworks

The representational harms identified by previous analyses of algorithmic harm focus on harms originating from prejudicial representations that stereotype, under-represent, denigrate, or ex-nominate people on the basis of their group identity [14, 24, 42, 67]. But why are such representations harmful? We argue that at least one way in which these representational harms function is by reinforcing prejudicial assumptions about people in ways that impact the epistemic frameworks people use to understand and navigate the world.

Epistemic frameworks refer to the epistemological systems that shape and constrain what individuals are in a position to know. They are the basic set of standards, norms and principles that structure our ways of making sense of the world [34]. Our epistemic frameworks influence which parts of the world we pay attention to, which types of explanations we deem explanatorily relevant, and

which beliefs are licensed as legitimate. These frameworks shape the expectations, interpretations, and responses we consider when faced with any situation [61].

In this way, epistemic frames can be thought of as a camera lens that limits "the otherwise unbounded and undelimited character of experience and restrict[s] one's scope of attention—not because one sees the frame, but because what one sees is seen through it" [61]. For example, the epistemic framework of naturalistic science favors naturalistic and causal explanations for a phenomenon based in the chemical and biological properties of objects. By contrast, a religious epistemic framework may ground the foundational interpretations of certain phenomena on religious texts and offer religious explanations grounded in God's agency rather than naturalistic explanations of science. As we see from this example, society is made up of individuals who hold different epistemic frameworks, and the incompatibility of different epistemic systems may be one explanation for the widespread and intractable disagreements we observe in political discourse [34]. Epistemologists refer to those disagreements that hinge on foundational differences between epistemic systems as being "deep disagreements" because the parties to the disagreement endorse different epistemic principles that license different beliefs and there is no further principle accepted by both parties that can resolve the dispute [52].

2.2 Epistemic Injustice

By attending to the epistemic frameworks that structure how we interpret and find meaning in the world, we are now well-placed to identify a specific type of injustice individuals may experience as a result of algorithmic harm—epistemic injustice. Miranda Fricker introduced the concept of epistemic injustice, which is often closely connected to social and political injustices, but focuses on the wrongs done to people in their capacity as knowers [32]. Fricker identifies two specific forms of epistemic injustice: testimonial and hermeneutical injustice [32]. Testimonial injustice refers to the credibility deficit tied to systemic prejudicial assumptions that cause people to deflate the credibility given to a speaker (e.g., discounting someone's expertise due to their gender or race). Hermeneutical injustice occurs when there is a gap in the collective interpretive resources that undermines the ability of certain groups to make sense of their social experiences. For example, prior to the introduction of the concept of sexual harassment, women who experienced it could not readily explain the harm nor see it as a systemic issue faced by many women. Instead, their discomfort was likely to be dismissed as overblown, humorless reactions to 'innocent flirtation' and 'normal' male behavior.

To Fricker's two types of epistemic injustice, Kristie Dotson adds a third type: oppressive epistemic systems. These occur when the dominant epistemic systems cannot account for the experience of certain groups because of their systemic inadequacy [27]. While hermeneutical injustices can often be identified as gaps within the dominant epistemic systems, epistemic oppression is particularly resilient because it shows the inadequacy of the epistemic system itself [27]. For example, white supremacy is not only a social and political system, but also an epistemic system that is remarkably resilient because it alters the way people understand the world and thus enables white supremacist beliefs to persist in spite of significant social efforts to reform and eradicate racist institutions [65]. Understanding white supremacy as an oppressive epistemic system can thus show why political reform continually fails to eradicate it and also point to a way forward. For the system to change, it is necessary to target the white supremicist belief systems that both have explicit defenders and have permeated the implicit biases that affect even well-meaning individuals in society.

The ways in which different social groups are frequently represented contributes to a social imaginary deeply imbued with prejudicial assumptions about, for example, what an expert looks and sounds like or who is assumed to be dangerous and criminal. On a social level, the representations we see have an enormous influence on our default assumptions about people and their capabilities. This, in turn, can have a profound influence on people's lives and opportunities. For example, while we can identify the injustice in law enforcement disproportionately targeting people of color, the default assumption of Black criminality is deeply rooted in our systems of belief, leading to people being more ready to assume that ordinary objects in Black people's hands are weapons [28]. Likewise, Black people are more likely to face a credibility deficit when bringing to light the racism to which they are daily subject, a form of testimonial injustice [32]. When prejudice permeates our system of belief and the default assumptions we make about the world, it can manifest in a myriad of ways both within and beyond algorithmic systems.

Addressing oppressive epistemic systems requires looking beyond current interpretations of antidiscrimination law and algorithmic fairness. Kimberlé Crenshaw calls attention to the ways racism operates as an epistemic and not only political system of discrimination. She argues that legal reforms will miss their target if scholars lose sight of the deeper roots of racist belief systems and, indeed, that previous efforts to reform laws and institutions have "merely repackaged racism" in a new form that maintains and legitimizes the "perpetuation of material subordination of Black" citizens [25], leaving unaddressed the ways racialized oppression works through "popular consciousness" [25]. Crenshaw argues that "racism is a central ideological underpinning of American society" and that "critical scholars who focus on legal consciousness alone thus fail to address one of the most crucial ideological components of the dominant order" [25]. We draw inspiration from Crenshaw's analysis of antidiscrimination law and call on scholars to "transcend oppressive belief systems" [25]. We argue that her critiques can inform the use of an epistemic lens for analyzing algorithmic harms in ways that identify the roots of algorithmic harms rather than repackage them in new ways.

2.3 Distinguishing the Epistemic Goals of an Inquiry: Categories of Intervention and Analysis

Just as one individual may have different epistemic goals at different times given the purpose of their quest for knowledge, so too may researchers have different epistemic goals when they are studying algorithmic fairness. Are they seeking to discover whether an algorithm is fair? Do they know there have been harms done to particular individuals and seek to quantify and prove the extent,

nature, or causes of these harms? Are they trying to correct the algorithmic system to redress known harms or design new measures for ensuring algorithmic fairness? In this section, we break these goals down to highlight distinct epistemic contexts within which algorithmic fairness researchers operate.

Researchers may analyze algorithms with the aim of investigating the application of an algorithm on a given population. When researchers adopt an investigative goal they are seeking to discover new knowledge about how an algorithm works on a given population and whether the algorithm treats people fairly. To meet this goal, researchers seek clarity around the design choices in an algorithmic system and seek to analyze the ways in which it has an impact on the populations under analysis. They are seeking new knowledge about the world as it is. Once researchers have identified distinct harms, they may shift their inquiry from a more open-ended mode of discovery towards more closed measurements that quantify the harms done to specific populations. Whereas the investigative aim inquires into whether algorithms treat people fairly, the aim of proof starts with the claim that there has been an algorithmic harm and seeks various ways of measuring and quantifying disparate treatment between different populations. For both goals (investigation and proof), the researcher seeks to describe, measure, and understand aspects of our existing social order.

But not all inquiries into algorithmic fairness are seeking to better understand the existing social order. Sometimes, researchers are seeking to design better systems: the aim of amelioration. Rather than describing what is true of the existing world, the ameliorative epistemic aim is explicitly normative and seeks to envision a more just, fair, and equitable future. This requires us to direct our attention beyond the world as it is to instead envision how it should be. We will return to these different epistemic goals in § 4 after illustrating how the epistemic lens works in the context of specific algorithmic systems.

3 THE EPISTEMIC LENS IN CONTEXT

To illustrate the efficacy of the epistemic lens in identifying specific types of algorithmic harms, we ground our analysis in a collection of real-world use cases involving algorithmic systems across a variety of domains including advertising, translation, image captioning, and search. In such systems, algorithms mediate the flow of information from platforms to individuals in ways that shape individuals' options, opportunities, and perceptions [35, 36, 48–50].

The epistemic lens shows how existing legal and technical definitions of fairness fail to capture key examples of algorithmic harm and, accordingly, fall short of addressing them. Looking beyond law and computer science to notions from philosophy, we show how the epistemic lens can be applied to help understand the gaps in existing frameworks and point to potential solutions to begin to fill these gaps by further identifying the roots of specific instances of harm. We first look at allocative harms in online ad targeting that are captured by antidiscrimination law (§ 3.1). We then show injustices that remain in online ad targeting and delivery that are better identified through the lens of epistemic injustice (§ 3.2). Lastly, we turn to harms often thought of as representational harms to show how an epistemic analysis can help better capture the root causes of the harms and suggest different ways to address them (§ 3.3).

3.1 Allocative Harms in Online Ad Targeting Captured by Antidiscrimination Law

Discrimination in online advertising is a growing concern in light of its increasingly outsized influence on employment, credit, and housing markets. Algorithms now play a role in identifying candidates for recruitment, discerning the most promising candidates within a pool of applicants, and predicting which employees might be open to leaving their current roles [44]. The ways in which algorithms analyze personal information, identify trends across groups, and make predictions and recommendations based on patterns in the data increase the likelihood that the opportunities they present to individuals will reflect patterns of disadvantage in the underlying systems from which these data are drawn [43]. Through these and related mechanisms, algorithms shape individuals' access to opportunities [43].

Algorithmic use cases may implicate federal statutes such as Title VII of the Civil Rights Act of 1964 [2], the Age Discrimination in Employment Act of 1967 (ADEA) [3], or the Fair Housing Act [5], which prohibit the publication of advertising for employment or housing that indicates a preference, limitation, or discrimination based on a protected classification [1, 4, 6]. Such laws aim to prevent employers from using their advertising to target or exclude potential candidates in ways that reflect historical disparities and exacerbate disadvantage along the lines of sex, race, and age.

In certain cases, online targeted advertising may be found to violate antidiscrimination law by indicating a preference, limitation, or selection criteria based on protected classifications. In the clearest examples, there may be evidence of an explicit decision to target ads based on sex, age, or race. For example, an investigation by ProPublica in 2016 found that Facebook's advertising platform enabled advertisers to target housing and employment ads to users based on their interests or background and to exclude groups of users based on criteria such as their "ethnic affinities" [11]. The U.S. Equal Employment Opportunity Commission found reasonable cause to believe several employers had violated Title VII and the ADEA by placing targeted ads for employment on Facebook that excluded women or older users from seeing the ads [66]. In a settlement with civil rights organizations, Facebook agreed to stop allowing advertisers to target ads for housing, employment, and credit based on users' age, gender and ZIP code [66]. Facebook also agreed to restrict location-based targeting using a metric (a minimum radius of 15 miles) that can be adjusted "based on facts showing that it is either irrelevant or ineffective to address concerns about perpetuating racial segregation" [66].

Concerns have also been raised with respect to Facebook's "special ad audience" tool, which uses a machine learning algorithm to identify users who are similar to a class of users selected by an advertiser such as its existing workforce [58]. Discrimination has been shown with respect to similar systems trained on the data of past job candidates (e.g., Amazon's automated system for reviewing job applications which was trained on past hiring decisions learned to "penalize resumes that included the word 'women's,"' as in "women's chess club captain" or graduates of women's colleges [26]). In 2022, the US Department of Justice filed a lawsuit against Meta, claiming that Facebook permitted the special ad audience algorithm to consider protected characteristics in selecting

an audience to receive housing ads [58]. In a settlement, Facebook agreed to stop using the special ad audience tool for housing ads [58].

3.2 Epistemic Injustices in Online Ad Targeting and Delivery

Despite efforts to limit advertisers' ability to use protected characteristics as ad targeting criteria, the potential for discriminatory advertising that perpetuates patterns of disadvantage persists. Seemingly neutral targeting criteria may be closely tied to protected characteristics, such as when targeting users with an interest in media directed at a particular demographic group [45], as well as in ways that are more difficult to anticipate ex ante. Additionally, platforms can introduce discriminatory effects through the use of ad delivery algorithms that show ads only to the subset of the advertiser's selected target audience that is most likely to engage with them [43].

Ad delivery algorithms can produce discriminatory effects without advertisers' knowledge, even in cases where advertisers intend to use neutral targeting criteria and aim to reach diverse audiences [43]. For instance, delivery can be skewed based solely on the demographics of a person depicted in an ad image [41]. In addition, skews often reflect disparities in employee demographics at the companies represented in the ads, which delivery algorithms learn and replicate [39]. One study found that ads on Facebook for supermarket cashier jobs reached an audience of 85% women, while ads for taxi driver roles reached an audience of 75% Black users, and ads for lumberjack positions reached an audience of over 90% men and 70% white users [9]. Another demonstrated that the gender skew in Facebook's delivery of job ads is not explained by differences in qualification [39]. For example, when targeting the same audience, a greater proportion of women were delivered ads for software engineering jobs at Netflix (where 35% of employees in tech-related positions are women) than for those at Nvidia (where 19% of all employees are women) [39].

Disparities in presentation of employment and housing opportunities can amplify occupational and residential segregation as well as socioeconomic stratification, along the lines of race, ethnicity, gender, and age. Yet many cases where such disparities arise with the use of neutral targeting criteria may fall outside the scope of existing legal protections despite the harms they engender. We argue that the epistemic lens can serve as a framework for characterizing significant harms perpetuated by algorithmic systems that are left unaddressed by existing doctrine. It can also provide a framework for expanding antidiscrimination law, similar to how it has historically been expanded to address other classes of injustices tied to subordinating people on the basis of their group identity [12]. For instance, Catharine MacKinnon successfully advocated in 1986 that sexual harassment is a form of employment discrimination under Title VII of the Civil Rights Act of 1964, leading the Supreme Court to recognize the theory of sexual harassment [8]. Scholars who draw connections between equal protection and antisubordination point to a deeper level of harm underlying the legal case for antidiscrimination [12, 16, 31]. At its root, antisubordination is about the persistent denigration of certain classes as subordinate to the dominant groups in society. The epistemic injustice lies in the oppressive epistemic systems that embed these hierarchical assumptions into relations between groups. Getting to the epistemic root of the problem can help us remain alert to the ways ad delivery is likely to continue to follow discriminatory patterns so that the harms can be identified and addressed.

For these cases, evaluating ad targeting and delivery systems through the lens of epistemic injustice may help track potential injustices that remain. Ad delivery algorithms can be understood as perpetuating gender-based and racial stereotypes operating in the background of existing social structures. The epistemic lens can help explain why this type of harm is both resistant and difficult to capture with legal definitions of discrimination. The persistence of discriminatory effects from Facebook's advertising platform, despite the removal of protected characteristics from the set of permitted targeting criteria for housing and employment ads, is just what we would expect if the root of the problem lies in deeper epistemic systems that have structured the ways people interpret the world. Fricker argues that "prejudice will tend to go most unchecked when it operates by way of stereotypical images held in the collective social imagination, since images can operate beneath the radar of our ordinary doxastic self-scrutiny" [32].[1]

Stereotypical images tend to reinforce patterns of racism and sexism in ways that are typically beyond our attention or control, showing up in the implicit biases that operate in even well-meaning individuals. Where social media algorithms learn and perpetuate stereotypes, they create digital images mirroring the collective social imagination, but as black boxes that reinforce residual prejudice and existing inequalities. Further, the fact that the algorithm perpetuates these disparities without the employer's knowledge—and, in some cases, undermines efforts the employer may have made to choose ad targeting criteria that would result in a gender-balanced target audience for their job ads—contributes to the epistemic injustice in this case. By influencing our social imagination and access to opportunity without our knowledge, this prejudice escapes the human processes of detection and correction of prejudice, by which "hearers' beliefs may at some point serve as a corrective force" when confronted with the realization that prejudice has had an "impact directly on hearers' perceptions of speakers" [32]. The use of ad delivery algorithms hinders the detection and correction of discriminatory presentation of ads, resulting in a lack of action to address the underlying biases and to create more equitable opportunities for disadvantaged groups.

An epistemic lens can also be used to help understand how an algorithm perpetuates discrimination and how best to address it. With roots in deep-seated stereotypes and lack of representation, these discriminatory patterns are likely to persist, despite efforts to address them. In its 2022 lawsuit, the US Department of Justice claimed that Facebook's ad delivery system relied on protected characteristics such as race, national origin, and sex [58]. In the settlement, Meta agreed to develop a new system for delivering housing ads in order to address disparities in race, ethnicity, and sex between the advertisers' target audiences and the audiences to whom the system actually delivers the ads [58]. On January 9, 2023, the US Department of Justice announced that it had reached a key agreement with Meta regarding compliance targets for its new

[1] Doxastic typically refers to belief systems.

ad delivery system to reduce variances between the targeted and actual audiences to "less than or equal to 10% for 91.7% of those advertisements for sex and less than or equal to 10% for 81.0% of those advertisements for estimated race/ethnicity" [59].

However, Facebook's changes to reduce the variances between targeted and actual audiences may have the effect of reducing the utility of the Facebook ad delivery algorithm without making meaningful improvements with respect to addressing the algorithm's discriminatory effects. The ad delivery algorithm is optimized to deliver ads to the subset of the audience predicted to be most likely to engage with the ad. Delivering the ad to a wider audience to include people unlikely to engage with the ad arguably will not have much of an effect on discriminatory outcomes. However, the ad platform could be designed, instead, to help advertisers understand the delivery algorithm, inform them of likely skew in the presentation and conversion rates, and identify the underlying causes, such as the replication of existing workforce disparities. The platform could provide tools to help advertisers design ad campaigns that will increase engagement across a diverse audience (as opposed to simply selecting targeting criteria that reflect a diverse audience). For example, such tools could help advertisers select a piece of text or an image—such as an image of a person from a specially disadvantaged group—to accompany the ad that the algorithm predicts will increase delivery and engagement with respect to disadvantaged groups [9]. In this way, algorithms have a role to play in both helping to understand ways in which algorithmic personalization works to filter employment and housing opportunities and in developing solutions that address the ways in which algorithms contribute to discriminatory effects.

3.3 Harmful Representations in Algorithmic Systems

Scholars have pointed to the difficulties of addressing representational harms through existing technical measures of fairness or antidiscrimination law. Representational harms are classes of harms that fall well beyond the scope of existing antidiscrimination doctrine, but they contribute to social stratification along the lines of race, ethnicity, age, gender, religion, sexuality, or disability. We argue that evaluating these algorithmic systems using an epistemic lens can help us better understand why certain representations are harmful due to their embodiment of oppressive epistemic systems. We also suggest that clarity with respect to which assumptions guide specific algorithmic choices can help point to better ways to address persistent algorithmic harms. The subsequent section will build on these suggestions to highlight how distinguishing between the epistemic aims of our discussions of algorithmic harms can help guide choices amongst different interventions.

Consider, for example, how the algorithms underlying translation software that import gendered stereotypes into translations are not designed to deliberately exclude or discriminate against women. However, researchers have demonstrated that the machine learning-aided translation of sentences like 'she is a doctor' and 'he is a nurse' into a gender neutral language and then back to English yielded the result 'he is a doctor' and 'she is a nurse' [20]. Clearly something went wrong from the perspective of justice, but it likely cannot be traced back to any particular programmer or decision that was made with the intent to discriminate. One likely cause is that the algorithm was trained on historical data in which healthcare professionals were more clearly sorted along gendered lines than is currently the case. However, one cannot point only to 'biased' training data to dismiss concerns about these results, as they reflect and reinforce deeply-embedded cultural assumptions that endure today. Researchers adopting an investigative goal of finding instances of gender bias were able to highlight and prove instances of algorithmic gender bias in the world as it is. Yet adopting a different, ameliorative epistemic goal can help guide corrective action. System architects can design potential interventions to better reflect the world as we would like it to be. For example, Google has since implemented a feature that flags a translation as being gender-specific and prompts the user to make a gender selection rather than inserting its own choice into the translation [47]. Google has also reported that it plans to add an option for non-binary gender in translation and to implement similar capabilities for handling gender-specific language in other algorithmic systems, such as auto-complete systems [47].

Examples of representational harms have also been found in the context of image captioning systems, demonstrating the potential for such algorithmic systems to reinforce the subordination of marginalized groups. Katzman et al. (2021) propose a taxonomy of representational harms in image captioning systems, encompassing the following six types of harms in this context: denying people the opportunity to self-identify, reifying social groups, stereotyping, erasing, demeaning, and alienating [42]. Wang et al. (2022) carried out an experiment using an image captioning system applied to real-world datasets and showed that such systems can generate captions that include words known to be demeaning or incorrect words that are likely explained by stereotyping [67]. Classification systems can reinforce gendered stereotypes when, for example, images of women are captioned based on their appearance and images of men are captioned based on their role [42, 67]. These systems can also undermine people's dignity when, for instance, an algorithm incorrectly classifies an individuals' gender. They also implicate Western secular epistemic frameworks when, as an example, algorithms misclassify religious wedding clothes as costumes [42]. Misclassifications also have the potential to lead to deadly consequences, such as the misclassification of a cellphone in the hand of a Black man as a weapon [28]. In each case, algorithms learn and reflect back to us oppressive epistemic frameworks that often operate under the level of our conscious awareness.

Applying the epistemic lens to cases where something has gone wrong within an algorithmic system can help clarify the social impact of these classifications, as well as the root causes of the harms often described as representational. While in a sense these are harms of misrepresentation and misidentification, they also highlight and reinforce problematic stereotypes embedded in the epistemic systems that structure how we find meaning in the world. Algorithms can reinforce harmful cultural assumptions and simplify complex social constructs. For example, a range of problematic assumptions underlie the misclassification of a person's gender identity including the assumption that gender can be determined based on an individual's appearance, can be described in binary terms, or can be determined at all without the subject's participation. These assumptions reinforce a limited social imaginary in ways

that contribute to the subordination of people who identify as outside the gender binary. The epistemic lens can be helpful in building or updating algorithmic systems that are not based on problematic assumptions but instead are based on a more inclusive social imaginary that moves in the direction of recognizing that an individual's gender is as an aspect of their identity and expression not a label to be assigned by another person or by an algorithmic system.

4 ADDRESSING ALGORITHMIC HARMS USING THE EPISTEMIC LENS: DISTINGUISHING THE EPISTEMIC GOALS OF AN INQUIRY

As we have seen, algorithmic systems mediate people's access to information and opportunities, thereby exercising pervasive influence in many different social domains. Assessing and addressing algorithmic fairness includes a range of different types of questions that may be relevant to a particular instance of biased algorithmic design or unfair algorithmic results. When auditing an algorithm for fairness, researchers may have a number of different kinds of questions in mind.

In each case, when we ask whether an algorithm is fair, we not only need to figure out how we are defining fairness, we also must ask what is the purpose of our inquiry. Much of the current literature on algorithmic fairness is consumed with this first question about how to specifically define fairness and the many ways in which different technical specifications of fairness are mutually incompatible [13, 21, 23, 29, 33, 37, 46, 55, 63]. Uncertainty with respect to fairness definitions risks leading us to either be complacent with the status quo or giving up too soon on the prospect of securing fair algorithmic systems.

However, by drawing on the epistemic lens we have introduced herein, we think there is potential for progress by starting with a much simpler task of specifying the purpose of any given inquiry into the fairness of an algorithm. Asking about the aim of our inquiry can help us to determine which fairness measures could be appropriate in that context and why. While there are many potential questions researchers could ask about whether an algorithm is fair, we will group them into two different epistemic contexts: descriptive questions seeking to understand the world as it is and normative questions about what a just future should look like.

In this section, we distinguish two epistemic contexts that are relevant to tackling this question. We believe that a clear understanding of the epistemic aim of any given inquiry or effort to change the system will help clarify which specific tools are appropriate to the task in these different contexts. At this point one might wonder, what role can (or should) technology play in trying to build a more just and equitable social order? Should algorithms be designed to be descriptively accurate—i.e., to describe the world as it is? Or should they be designed to live up to our normative ideals—i.e., to represent what a just world might look like?

4.1 The Epistemic Goal of Understanding the World as It Is

Let's take the first proposal: that algorithms should be designed to be descriptively accurate. There is certainly an important role to be played by tools that can help us better understand and track the actual distributive patterns in the world. Algorithmic systems can play an important part in this research to help us track distributive patterns and potentially highlight instances or forms of bias that had previously gone unnoticed. For example, the push for researchers to be able to access the data underlying the major social platforms that influence so much of society is, in part, a push to have access to research to help us build a more accurate understanding of the world as it is [35, 36].

The first, descriptive, epistemic context contains all those inquiries seeking to describe and understand what is true of the existing world. Here, the epistemic aim of the inquiry is understanding facts about the world, how algorithms were built, how they currently operate, and the impact of these systems on our society. Researchers asking questions in this category seek empirically verifiable facts and aim to present and expose the ways systems operate. For example, the ProPublica investigation into COMPAS exposed the disparity in risk assessment scores by race [10]. This investigation required uncovering the survey questions underlying the algorithm's predictive scoring as well as empirical facts about rearrest rates of those who had been assigned a particular risk score. This case is clearly one of discovering and describing the world (and the algorithm) as it is to make the case that the algorithm was unfair. Likewise, researchers who exposed the targeting of housing and employment ads based on "ethnic affinities" [11] and those who exposed the disparity of results through the "special ad audience" tool [58] sought to reveal existing patterns of discrimination in the operation of Facebook's ad targeting algorithms.

Notably, some of these cases expose clear patterns of unjust distribution of resources and opportunities. But insofar as they expose assumptions about who might be appropriate for a given category of job, they also reveal existing patterns of epistemic injustice. If two people are equally qualified for a given job or equally likely to be re-arrested after parole, why might algorithms offer different recommendations to people of two different social groups? These findings can help us see the historical biases that have been inadvertently encoded in algorithms based on the ways the world operates in patterns of prejudice and inequality. In addition, through their continued operations, these algorithms also influence the future.

4.2 The Epistemic Goal of Building a More Just, Fair, and Equitable Future

Another aim of researchers when they call for access to the underlying platform data is to study the ways in which these algorithms influence people through the filtering of information and opportunities that ultimately have an impact on what comes to be true of our social order. Algorithms increasingly mediate our access to information, opportunities, and each other. In this role, algorithms are far from a neutral purveyor of information. They instead shape options, information, and opportunities in ways that have a profound influence on society. This brings us to the second proposal: because algorithms play a critical role in influencing the creation

of new social patterns, one could argue that we should ensure that they influence people in ways that lead to a more just and equitable future.

The second, normative, epistemic context involves those future-oriented questions involved in trying to envision and build a more just, fair, and equitable future. Rather than describing the world as it is, the inquiry is aimed at determining what it should look like and, in turn, how we can move from our current world towards a more just future. In some cases, historically biased data does not reflect current reality because there has been significant social progress in many areas breaking down gendered and racial sorting of professions. For example, while, historically, doctors were predominantly men and nurses were predominantly women, that is no longer an accurate description of the current world. An algorithm that learned these historically biased gendered patterns would not be an accurate representation of the world as it is. But what if the skewed results by gender reflect the current reality?

We still have a long way to go to rectify the myriad effects the long history of racism, sexism, and other forms of social prejudice have caused. In some cases algorithms produce accurate results, but they are nevertheless morally problematic. When googling 'CEO' the results are likely to show mostly white men. Yet this search result does accurately portray the makeup of CEOs. In recent years, there has been enormous attention paid to gender equity in leadership yielding important gains for women. For example, 2022 was a record-setting year for increasing the number of women serving as CEOs in Fortune 500 companies. In 2002, there were only 7 women in that role. By June of 2021 there were 41 and by March of 2022 the number rose to 74 women [18]. However, this is still a mere 15% of all Fortune 500 companies. Clearly, search results that show mostly white men in this role are accurate; however, that does not mean that they are fair or reflect society's values.

As another example, to properly identify the harms in binary gender labeling, we do must understand the inaccuracy of reductive binary labels of people's gender expressions. Gender identities and expressions are not reducible to two simple categories. Algorithmic systems built on binary gender classifications do not represent the array of people's lived experiences. Results skewed by this simplistic classification system would be inaccurate, failing to capture the world as it is, but it also reinforces the default assumption that gender is a simple binary category. Correcting these algorithms forces us to move into territory that asks us to think about what should be the case. This normative step requires participatory input from those impacted by these classifications to determine which representations would be more empowering as well as accurate to their lived experiences.

4.3 The Interplay between the Two Epistemic Goals

We have presented these epistemic goals as distinct for the purpose of examining which fairness measures might be appropriate. But we do not mean to suggest that these two epistemic contexts are utterly separate. There is clearly significant interplay between understanding the world as it is and attempts to envision and build a more just future. There is immense value in gaining an accurate understanding of the world as it is and the ways in which prejudice, bias, mis- and disinformation are impacting people currently. Understanding how algorithmic personalization works to filter information and opportunities to individuals is an important first step in asking questions about how the current system works and how we can and should address emergent social problems. Without an accurate understanding of what is actually happening in the world, it is more difficult to call out injustices in order to demand change.

The examples discussed herein illustrate that the normative questions are clearly intertwined with an accurate descriptive understanding of the world as it is. Empirical research reveals and describes patterns of injustice in algorithmic system. Normative research takes these injustices and asks us to envision what a just world should look like. Adopting the epistemic lens can help determine which kinds of questions we are asking in a given research project and clarify when accuracy is the desired result and when our metric should be oriented explicitly towards change. In both cases, clarity with respect to who is setting the goals for accuracy or justice matter. But these are distinct aims likely to lead to different potential answers about which among the many ways of understanding the fairness or unfairness of algorithms are best suited to answer the specific question being asked.

These two epistemic contexts can also diverge in significant ways. For example, when searching for CEOs, there may be two different results that could be appropriate given the epistemic context. First, if we are looking for an accurate picture of the existing world, results for a search of Fortune 500 CEOs should show accurate headshots of everyone with that title. This accurate result will show a vast majority of white men in this position. However if all image results are geared towards accuracy, this may reinforce existing stereotypes about the types of people who are qualified for leadership positions. There may be contexts in which it is appropriate to highlight different results to help shift our social imaginary in ways that align with our normative priorities.

What if we could start to build a better world by ensuring that a more equitable array of careers, talents, and capacities is highlighted in ways that can start to undo longstanding prejudicial stereotypes? For example, while it remains true that most CEOs are white men, some of the reasons why this is true can be traced to the assumptions we make about expertise and authority. If our default assumptions expect leadership and authority to have a certain look or voice, we will continue to promote or elect individuals who fit our assumptions about what leadership should be. If, however, we work to change those representations—by highlighting individuals who are knowledgeable experts, competent leaders, and clear authorities but do not fit our usual assumptions about authority and leadership, we can start to challenge the longstanding prejudice that lies in our collective epistemic frameworks and social imaginary.

Of course this suggestion may be quite difficult to implement in practice. A general purpose image search system is ill-equipped to determine the epistemic goals of a user's inquiry. Yet algorithmic oppression is a persistent and important problem. Search results and suggested completion of search terms include racist, sexist, and denigrating results [57]. Existing search systems are not neutral. Results predictably reflect and reinforce harmful representations and perpetuate epistemic injustices against already marginalized groups. Most recently, we have seen these stereotypes creeping

into generative AI systems like image generators that represent Asian women with hypersexualized avatars [38].

Society is rife with conflicts over what justice or fairness requires. But these conflicts are not always as intractable as some people assume. Notwithstanding continued contestations over what would be truly just and fair, there is plenty of progress that can be made by working to address clear cases of prejudicial discrimination tied to longstanding, well-documented histories of racism, sexism, heterosexism, and classism. Waiting for the debates to settle on which understanding of fairness is the best will actually undermine important efforts to address the clear cases of prejudice and discrimination that are causing harm. Furthermore, by emphasizing the epistemic uncertainty and debates over fairness metrics, this has the risk of inadvertently undermining important efforts to address clear injustices in existing systems. To see this, we need only look to the ways epistemic uncertainty has been leveraged to sow public doubt about climate change in order to undermine public support for the collective efforts needed to avoid catastrophic climate change.

Shifting representation to reflect the world as we would like it to be could help to expand our moral imagination, subtly adjusting our expectations and, ideally, thereby working to counteract the implicit biases that so often permeate and undermine even good faith efforts to build a more equitable world. Building more equitable representations is not about fabricating an imaginary world. It is about highlighting aspects of the world as it is that have been buried in our social imaginary due to longstanding epistemic systems of injustice. This better reflects our actual world and also helps to break down prejudicial assumptions in order to help build a world characterized by a true equality of opportunity.

By making explicit the various inequities embedded in decision making (including both explicit and implicit biases learned by machine learning algorithms), algorithms can help us identify areas of interest that call for moral attention. It can prompt new questions about which values are important in different domains and when inequities are justifiable and why. All of this can help us make explicit our default assumptions about how the world works and how we would like it to be structured moving forward.

5 CONCLUSION

This position paper introduces a new epistemic lens for evaluating questions of algorithmic fairness and advancing our understanding of algorithmic harms. The epistemic lens directs our attention to the epistemic frameworks that shape our interpretations of the world as it is and the ways we envision possible futures. Through this epistemic lens we have identified a deeper level of harm not currently captured by existing distinctions between allocative and representational harms: many instances of algorithmic harms are rooted in the ways that algorithms operate to reinforce various forms of epistemic injustice. We argue that the epistemic lens provides a theoretical foundation for expanding existing approaches to capture a wider class of algorithmic harms. Applying this lens to real-world cases, we have demonstrated its potential for understanding the foundational causes of allocative and representational harms and for targeting interventions to the source of the problem. We then highlight that addressing algorithmic injustice often involves two different sets of questions. At times researchers seek to understand and describe the world as it is, revealing patterns of injustice in the operation of algorithmic systems. At other times, researchers seek to address these harms by making normative changes designed to bring about a better future. Clarity with respect to which types of questions are being asked in particular contexts can help researchers make explicit the criteria by which they judge whether or not an algorithm is fair and design effective interventions to address allocative and representational harms.

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 the participants of the 15th and 16th annual Privacy Law Scholars Conferences (PLSC 2022 and PLSC 2023), members of the Data Co-ops Working Group, and the anonymous reviewers for their helpful feedback on earlier 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. *Age Discrimination in Employment Act of 1967.* 29 U.S.C. §§ 621–634.
[4] 1967. *Section 623(e) of Age Discrimination in Employment Act of 1967.* 29 U.S.C. § 623(e).
[5] 1968. *Fair Housing Act.* 42 U.S.C. § 3601 et seq.
[6] 1968. *Section 3604(c) of the Fair Housing Act.* 42 U.S.C. § 3604(c).
[7] 1974. *Equal Credit Opportunity Act.* 15 U.S.C. § 1691 et seq.
[8] 1986. *Meritor Savings Bank v. Vinson.* 477 U.S. 57 (1986).
[9] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Delivery Can Lead to Skewed Outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (November 2019), 1–30.
[10] 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
[11] 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
[12] Jack M. Balkin and Reva B. Siegel. 2003. The American Civil Rights Tradition: Anticlassification or Antisubordination? *U. Miami L. Rev.* 58 (2003), 9–34.
[13] 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
[14] 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).
[15] Alistair Barr. 2015. Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms. *Wall Street Journal* (1 July 2015). https://www.wsj.com/articles/BL-DGB-42522
[16] Derrick Bell. 1987. *And We Are Not Saved: The Elusive Quest for Racial Justice.*
[17] Yochai Benkler, Rob Faris, and Harold Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics.* Oxford University Press, New York, NY.
[18] Katharina Buchholz. 2022. Only 15 Percent of CEOs at Fortune 500 Companies are Female. *Statista* (8 March 2022). https://www.statista.com/chart/13995/female-ceos-in-fortune-500-companies
[19] 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
[20] 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
[21] 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
[22] Danielle Keats Citron and Frank A. Pasquale. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89 (2014), 1–33.

[23] 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

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

[25] Kimberlé Williams Crenshaw. 1988. Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law. *Harvard Law Review* 101, 7 (May 1988), 1331–1387.

[26] Jeffrey Dastin. 2018. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. *Reuters* (9 Oct. 2018).

[27] Kristie Dotson. 2014. Conceptualizing Epistemic Oppression. *Social Epistemology* 28 (2014), 115–138. Issue 2.

[28] Trone Dowd. 2021. The Deadly Consequences of Carrying a Cell Phone While Black. *Vice News* (4 Mar. 2021).

[29] 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

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

[31] Owen M. Fiss. 1976. Groups and the Equal Protection Clause. *Phil. & Pub. Aff.* 5 (1976), 107–177.

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

[33] 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

[34] Alvin I. Goldman. 2010. Epistemic Relativism and Reasonable Disagreement. In *Disagreement*, Richard Feldman and Ted A. Warfield (Eds.). Oxford University Press, Oxford, 187–215.

[35] Ayelet Gordon-Tapiero, Alexandra Wood, and Katrina Ligett. 2022. The Case for Establishing a Collective Perspective to Address the Harms of Platform Personalization. In *Proceedings of the 2022 Symposium on Computer Science and Law* (Washington DC, USA) *(CSLAW '22)*. Association for Computing Machinery, New York, NY, USA, 119–130.

[36] Ayelet Gordon-Tapiero, Alexandra Wood, and Katrina Ligett. 2023. The Case for Establishing a Collective Perspective to Address the Harms of Platform Personalization. *Vanderbilt Journal of Entertainment & Technology Law* 25 (2023), 635–689.

[37] 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

[38] Melissa Heikkila. 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/

[39] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for Discrimination in Algorithms Delivering Job Ads. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3767–3778. https://doi.org/10.1145/3442381.3450077

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

[41] Levi Kaplan, Nicole Gerzon, Alan Mislove, and Piotr Sapiezynski. 2022. Measurement and Analysis of Implied Identity in Ad Delivery Optimization. In *Proceedings of the 22nd ACM Internet Measurement Conference* (Nice, France) *(IMC '22)*. Association for Computing Machinery, New York, NY, USA, 195–209. https://doi.org/10.1145/3517745.3561450

[42] 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.*

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

[44] Pauline T. Kim and Erika Hanson. 2016. People Analytics and the Regulation of Information Under the Fair Credit Reporting Act. *Saint Louis University Law Journal* 16 (2016), 17–34.

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

[46] 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

[47] James Kuczmarski. 2018. Reducing gender bias in Google Translate. *Google Blog* (6 Dec. 2018).

[48] Katrina Ligett and Kobbi Nissim. 2020. Data Co-Ops: Challenges, and How to Get There. *DIMACS Workshop on Co-Development of Computer Science and Law* (11 November 2020). https://youtu.be/ZZugFpAOA64

[49] Katrina Ligett and Kobbi Nissim. 2020. Data Cooperatives in the Real World: Progress and Challenges. *Radical Exchange Conference RxC 2020* (19 June 2020). https://youtu.be/vUbuOiyosjI

[50] Katrina Ligett, Kobbi Nissim, and Matt Prewitt. 2021. Computer users of the world, unite. *The Boston Globe* (15 October 2021). https://www.bostonglobe.com/2021/10/15/opinion/computer-users-world-unite/

[51] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf

[52] Michael P. Lynch. 2010. Epistemic Circularity and Epistemic Incommensurability. In *Social Epistemology*, A. Haddock, A. Millar, and D. Pritchard (Eds.). Oxford University Press, Oxford, 262–277.

[53] José Medina. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations.* Oxford University Press, New York, NY.

[54] 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

[55] 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

[56] 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

[57] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press, New York, NY.

[58] US Department of Justice. 2022. Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising: Lawsuit is the Department's First Case Challenging Algorithmic Discrimination Under the Fair Housing Act; Meta Agrees to Change its Ad Delivery System. (21 June 2022). https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known

[59] US Department of Justice. 2023. Justice Department and Meta Platforms Inc. Reach Key Agreement as They Implement Groundbreaking Resolution to Address Discriminatory Delivery of Housing Advertisements. (9 January 2023). https://www.justice.gov/opa/pr/justice-department-and-meta-platforms-inc-reach-key-agreement-they-implement-groundbreaking

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

[61] Peter Railton. 2006. Normative Guidance. In *Oxford Studies in Metaethics*, Russ Shafer-Landau (Ed.). Vol. 1. Oxford University Press, Oxford, 3–34.

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

[63] 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

[64] 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

[65] Briana Toole. 2021. What Lies Beneath: The Epistemic Roots of White Supremacy. In *Political Epistemology*, Elizabeth Edenberg and Michael Hannon (Eds.). Oxford University Press, Oxford, 76–94.

[66] 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 Job Ads. Press Release. (25 September 2019). https://www.aclu.org/press-releases/historic-decision-digital-bias-eeoc-finds-employers-violated-federal-law-when-they

[67] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring Representational Harms in Image Captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 324–335. https://doi.org/10.1145/3531146.3533099