Opportunities for a More Interdisciplinary Approach to Measuring Perceptions of Fairness in Machine Learning

C. Malik Boykin
c_boykin@brown.edu
Brown University
Providence, Rhode Island, USA

Sophia T. Dasch
Humboldt University of Berlin
Berlin, Germany

Vincent B. Rice, Jr.
University of Buffalo
Buffalo, New York, USA

Taiwo A. Togun
Venkat R. Lakshminarayanan
SeqHub Analytics LLC
New Haven, Connecticut, USA

Sarah M Brown
University of Rhode Island
Kingston, Rhode Island, USA
brownsarahm@uri.edu

ABSTRACT

As machine learning (ML) is deployed in high-stakes domains, such as disease diagnosis or prison sentencing, questions of fairness have become an area of concern in its development. This interest has produced a variety of statistical fairness definitions derived from classical performance metrics which further expand the decisions that ML practitioners must make in building a system. The need to choose between these definitions raises questions about what conditions influence people to perceive an algorithm as fair or not. Recent results highlight the heavily contextual nature of fairness perceptions, and the specific conditions under which psychological principles such as framing can reliably sway these perceptions. Additional interdisciplinary insights include lessons from the replication crisis within psychology, from which we can glean best-practices for reproducible empirical research. We survey key research at the intersection of ML and psychology, focusing on psychological mechanisms underlying fairness preferences. We conclude by stating the continued need for interdisciplinary research, and underscore best-practices that can inform the state-of-the-art practice. We consider this research to be of a descriptive nature, enabling a deeper understanding and a substantiated discussion.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Human-centered computing** → **HCI theory, concepts and models**.

KEYWORDS

machine learning, experiment design, fairness

ACM Reference Format:
C. Malik Boykin, Sophia T. Dasch, Vincent B. Rice, Jr., Taiwo A. Togun, Venkat R. Lakshminarayanan, and Sarah M Brown. 2021. Opportunities for a More Interdisciplinary Approach to Measuring Perceptions of Fairness in Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, 2021, –, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3465416.3483302

EAAMO '21, October 5–9, 2021, –, NY, USA

ACM ISBN 978-1-4503-8553-4/21/10.
https://doi.org/10.1145/3465416.3483302

1 INTRODUCTION

The struggles of Artificial Intelligence (AI) practitioners to adhere to social values such as fairness are now a matter of public life. To date, AI has failed at meeting nondiscrimination expectations in facial recognition [13], healthcare allocation [43], language processing [7], and recidivism prediction[2]. Technologists began to mitigate biases with purely technical approaches: formalizing fairness by writing it as a constraint that can be added to a learning algorithm [3, 20], but quickly found that many definitions hold different social values and mathematically, are mutually exclusive and cannot co-exist [16, 28]. Practitioners and social scientists alike note that individual learning algorithms are embedded in larger systems that also require attention [25, 58]. Choosing one, therefore has been viewed as beyond the scope of technical skills alone and more human centered computing approached are entering the conversation.

Fairness is not simply translated into an equation; it is socially constructed and generally context-specific, and computing's aim to modularize and abstract everything may hurt the overall endeavor [58]. Fairness constraints, when used carefully, can produce a desirable outcome. For example, under the assumptions that the training labels are incorrect at different rates for different groups, an equal opportunity constrained classifier can recover the true labels at a higher rate than an unconstrained classifier [6]. These definitions may also support computing's broader roles in social change, diagnosing both algorithmic and social processes or as a light on how individuals and groups *express their biases*[1]. This understanding can guide policy discussions and public education to improve technological literacy.

Our main contribution is to demonstrate the relevance of a thoroughly considered research design that allows measuring participants' true fairness preferences, understanding underlying mechanisms, and ensuring generalizable, replicable results. We do not propose that practitioners treat majority perceptions of ML fairness as specific advice beyond descriptive. This research should improve practitioners' understanding of fairness perceptions – enabling subsequent discussions on implementing fairness criteria in practice with a more substantiated scientific foundation. Finally, we highlight the important opportunities of an interdisciplinary approach

by emphasizing the novel dimensions provided by a psychological lens.

2 PERCEPTIONS OF FAIR MACHINE LEARNING

Several recent studies have evaluated *perceptions of fairness*: preferences across fairness definitions, lay person understanding of fairness definitions, and understanding of the design trade-offs when considering fairness in an ML system [24, 57, 64, 75]. We begin by summarizing the key questions, design choices, and findings of these studies, reserving critique or evaluation of the studies for the subsequent sections.

Harrison et al. [24] aim to directly evaluate perceptions of fairness. Their study includes a sample of 502 Mechanical Turk workers, and their design presents each participant with several scenarios based on the COMPAS dataset. This study investigated participants' preference between four competing notions of fairness: equal accuracy, equal false positive rate (FPR), equalized odds, and consideration of race (Black and white people specifically). Participants indicated their preferences by making a series of choices between two models for bail eligibility: Model X, which, for instance, held accuracy constant across racial groups while varying FPR, or competing Model Y, which held FPR constant across racial groups while varying accuracy. Researchers concluded that, when given a choice between equalizing accuracy and equalizing FPR (the chance of being mistakenly denied bail), subjects prefer equalizing FPR.

Saha et al. [57] assessed non-experts' comprehension of fairness metrics and factors that influence comprehension. A preliminary study with 147 participants validated the method and revealed that context (hiring, giving employees awards, or judging a student art project) did not influence understanding. In the main study, 349 participants recruited via Cint to roughly match US demographics saw illustrations of each fairness definition as a rule that a decision maker must follow in order to be fair to those impacted. Participants indicated their agreement with and how much they liked each rule on Likert scales. Participant demonstrated comprehension by applying the rule to determine how many offers to send and evaluating true/false statements about how the rule relates to merit and other factors. Comprehension scores varied across fairness definitions: participants scored low for "equal opportunity" and "false negative rate (FNR)," as compared to other fairness definitions such as demographic parity. Additionally, FNR had the highest variability in comprehension scores and comprehension was negatively correlated with sentiment toward a rule.

Srivastava et al. [64] applied a "descriptive ethics" approach to identify the mathematical notion of fairness that most closely matches lay people's perception of fairness across different contexts. They hypothesized that context would influence preferences, but found that demographic parity was preferred in both recidivism risk assessments and skin cancer risk assessment. In a preliminary study, 20 MTurk volunteers indicated the preferences for a free response text box was over a structured text box to offer explanations for fairness preferences. In the main studies, 100 paid MTurk workers judged the fairness of algorithms in each of two contexts (recidivism risk and skin cancer risk) through an adaptive experimental design. In each test set, participants were shown predictions under two hypothetical algorithms and true outcomes disaggregated by race and gender for the same set of 10 impacted individuals (3 white men, 2 white women, 2 Black men, 3 Black women) and asked to choose which algorithm was more discriminatory. An adaptive algorithm administered each pairwise test, varying the parameters of the participant's choice set (i.e., degree of discrimination, definition of fairness). A simulation was then used to demonstrate 9262 possible tests, wherein random presentation order would require 600 tests to have a high confidence prediction of the participants preference – but only 20 were required for the adaptive presentation procedure. The testing stopped when the algorithm determined a participant's preferences along each dimension of fairness. Finally, participants selected one of three algorithms: high accuracy, large gender disparity; moderate accuracy and gender disparity; or low accuracy, no disparity.

Yu et al. [75] propose a tool to help algorithm designers understand design trade-offs, which they validated with 301 MTurk workers. They reexamined the classical trade-offs between types of errors and accuracy and fairness. To illustrate the trade-offs, they produced a set of Pareto optimal predictors on racial and gender balanced subsets of the nonviolent offenders from the COMPAS dataset [2]. Participants were assigned to one of three conditions: an interactive visual (a confusion matrix view), an interactive text-based tutorial (text view), or a baseline (in which participants received no experience with an instructional tool). Participants then completed a multiple choice test to evaluate comprehension, self-assessed their understanding on a Likert scale, rated their trust on a Likert scale, and indicate their preferred model. Participants in both interactive view conditions demonstrated statistically significant increases in comprehension and half changed their level of trust in the algorithm's predictions after experience with an interactive tool, but were approximately as likely to increase their trust (22.3%) as reduce their trust (25.1%). Yu et al. [75] propose that these results suggest improving participants' comprehension of these inherent trade-offs when evaluating an algorithm's fairness to avoid biasing participants' decisions by influencing their understanding with a two step process: an interactive demo followed by formally illustrated tradeoffs.

3 MEASUREMENT

Human attitudes can rarely be measured directly; instead they must be captured indirectly using a measuring instrument. An often under-considered assumption in measuring attitudes is that the quality of interest is the sole cause of an individual's responses [27]. However, there is often a part of the variance in the measurement of attitude attributable to the measuring instrument or research methodology [50]. This highlights that researchers should consider the signal-to-noise ratio known as method variance. Decisions about which measuring instruments are chosen, which manifest variables are selected to describe the latent constructs, and which potential mechanisms are assessed, are crucial to whether the study results are sufficient signal to noise ratio to provide meaningful insights. Generalizability with respect to experimental setting is also a concern so that results can be applied to the real-world efficiently, as guidance for practitioners. In these sections, we will indicate where studies on ML fairness conducted so far could benefit from

the extensive expertise psychology has developed in measuring human attitudes.

3.1 Reliability

To assess comprehension of fairness definitions, [57] uses a questionnaire and reports Chronbach's Alpha reliability coefficients between .38 and .64, all below the recommended .70 threshold for reliable measurement [29]. This is further complicated by the fact that, in several analyses, items were deleted to increase the reported alpha. Deleting scale items to help increase reliability is risky because it is unlikely that this change in reliability locally generalizes from the sample to the population. This results in artificially inflated reliability of the measure, therefore deflating the interpretability of the results [54]. Tinkering with scale measurements during analyses is an inherently exploratory method that needs to be reproduced in subsequent samples or in subsets of the same sample to increase interpretability[12]. Analyzing the covariance of remaining items with a latent variable model is the recommended practice [12, 54].

3.2 Framing Effects

While answering a survey, participants must continuously make decisions between different options such as choosing between Models X and Y, each representing satisfying a different fairness criterion. The representation of the various options strongly influences the decision-making process. Participants' perception of the choices is referred to as the "decision frame" [68]. The framing of a choice is determined by internal and external factors, such as individual characteristics of the decision-maker, and the presentation of a particular choice [68]. Since different options can be presented in multiple ways, the researchers' choice of the decision frame is crucial, as it most likely will influence the participants' responses. Psychological research has addressed this challenge for several decades, from classical studies about framing effects in risk communication to recent efforts of finding an integrated explanation for different underlying mechanisms of framing effects [15, 37, 46, 63, 68]. Considering framing effects provides a valuable perspective for the discussion of research on people's understanding of ML. The following critiques rely on primarily classical literature, because more recent studies investigate finer nuances of these effects, such as individual differences and context-dependency[8, 44].

3.2.1 Loss Aversion. Participants in Harrison et al. [24] chose between "Accuracy" and "Mistakenly Denied Bail" based on the well-established effect that losses and disadvantages are more influential in decision making for individuals than gains and advantages [40, 70]. We argue that this decision frame weights the criteria asymmetrically and biases participants to choose "Mistakenly denied bail". Individuals' tendency to weigh losses more strongly than equally-sized gains is called loss aversion[69]. McNeil et al. demonstrated how framing an outcome as a loss can substantially change participants' responses. Subjects were presented with two alternative therapies for lung cancer, surgery and radiation therapy, and were asked to indicate their preference. However, one group of participants was given the mortality rate of the surgery (e.g. 10 percent), while the other group was shown the survival rate (e.g. 90 percent). The fundamental information is identical, but the treatment option is presented differently – in one case, in terms of gains (survival), and in the second case, in terms of losses (mortality). Compared to radiation therapy, surgery was chosen significantly more often in the experimental condition when the survival rate of surgery was reported (e.g. 90 percent), instead of the mortality rate. This finding replicates across contexts, for example, motivation to perform a breast self-examination depends on the framing of cancer outcomes[38]. Emphasizing the negative consequences of omitting the examination was more effective in motivating people. The described studies demonstrate the extent to which differently framed messages (loss vs. gain frame) may result in deviating attitudes, even if content wise the same information is presented.

These framing principles (gains versus losses) come into effect in [24] when assessing participants' preferences for certain fairness criteria over others. In particular, the labeling of the graphs shown to participants framed certain fairness criteria (for example, "Accuracy") in a positive / gain frame, and other fairness criteria (such as "Mistakenly Denied Bail") in a negative / loss frame. The different frames result in a weighted asymmetry between the fairness criteria. We argue that the chosen presentation gives more weight to "Mistakenly Denied Bail" in the decision about which model is more fair, thus subjects are more likely to be guided by this fairness criterion instead of accuracy. This assumption is in line with the results: participants considered the model more fair, when FPR, "Mistakenly Denied Bail", was equalized between the groups, whereas an unequal accuracy rate remained the necessary trade-off [24]. Unfortunately, the design of the study makes it impossible to identify whether these results reflect a true preference for equal FPR over equal accuracy, or whether these results are simply due to the specific framing choices made by researchers.

3.2.2 Level of Abstraction. Asking participants to compare "Mistakenly Denied Bail" to "Accuracy" in Harrison et al. [24] also makes a comparison across levels of abstraction. "Mistakenly Denied Bail" is a concrete event, while "Accuracy" is an abstract metric. In addition to the effect of loss aversion, this framing aspect might further cause participants to overvalue the fairness criteria of equalized FPR, "Mistakenly Denied Bail". We argue that the participants may have relied on the availability heuristic, thus overestimating the probability of the occurrence that a person is mistakenly denied bail. The availability heuristic is a mental shortcut that people apply to estimate the probability or frequency of events. In order to do so, they use their perception of availability, i.e. how easy it is to recall the relevant information [67]. This relationship between the ease of retrieval and the estimation of frequency has been demonstrated in many studies [10, 33, 41, 67]. For example, participants overestimated the frequency of lethal events that were more concrete and easier to imagine [33]. Tversky and Kahneman [67] asked participants to estimate whether words with the letter 'R' in the first or third position were more frequent. It is much easier to retrieve words that have the particular letter in the first position than in the third so a majority of participants incorrectly expected that more words had R in the first position. Moreover, Trope and Liberman [66] proposed that an object can be expressed at different levels of abstraction, or construal. Take "cellular phone" as an example of a low level abstraction, since the same object could also be referred to as a "communication device", a description that represents a higher level of construal [66].Moreover, it is even sufficient to manipulate

Table 1: A summary of studies investigating peoples' perceptions of machine learning fairness. The number beside each attribute indicates a section of this paper that provides more details.

	Target Study (2)	Recruiting (2)	Num Participants (2)	Reliability (3.1)	Framing (3.2)	Mechanisms (3.3)	Generalizibility (3.4)	Power and Replication (4)
Harrison et al. [24]	Preferred fairness standard	MTurk	502	Sampling bias	Loss aversion & level of abstraction	Sampling bias (via mechanisms like social identity)	Sampling bias & uncorrected significance levels	Uncorrected significance levels, Underpowered convenience sample
Yu et al. [75]		MTurk	87-100 / condition		Loss aversion & level of abstraction	Sampling bias (via mechanisms like social identity)	"Demographically balanced" (unrepresentative) scenarios	Demographically balanced" (unrepresentative) scenarios
Srivastava et al. [64]	High versus low stakes	MTurk	100 / condition	Qualitative methods can't show statistical cause		Quantitative follow-ups (like mediation analysis) needed to understand preference shifts	Quantitative follow-ups (like mediation analysis) needed to under- stand preference shifts	Small number of multiple comparisons maintains statistical power
Saha et al. [57]	Fairness tradeoff comprehension	Cint	147 (validation), 349 (main)	Cronbach's Alpha of .38 to .64			Scale items deleted to inflate reliability	Scale items deleted to inflate reliability

the mindset regarding the construal level to influence participants' probability estimates [72]. The level at which an event is described, whether abstract or concrete, affects decision-making processes [9]. A low level of construal is more likely to lead to the application of the availability heuristic and overestimation of frequency, which migh bias participants to prefer one fairness criterion over the other.

3.3 Mechanisms

Experiments can also reveal why people make the choices they make. Harrison et al. [24] demonstrated that when people are presented with options of fairness metrics, they make comparative choices and Srivastava et al. [64] showed that when stakes were higher participants were less willing to sacrifice accuracy for group-wise fairness. This highlights that the gravity of the decision and its' impact on people's lives has implications for people's preferences that should be considered in future studies and algorithmic design. Additionally, they give us new questions to test. Do people care most about accuracy in all medical decisions or just life-threatening ones? At what point on the severity continuum do preferences shift? If a particular medical condition is known to disproportionately impact one population does this change people's preferences?

Because a large number of causal factors, both spurious and of interest, influence an expressed preference, it can be complicated to identify mechanisms behind choice. Mediation analysis, the process of identifying variables that help to explain the influence of independent variables on outcomes, aids in this process. For example, in [64], multiple factors could be at play in driving preference for equal FNR over equal accuracy. For instance, Srivatsava's paradigm allowed 20 participants to generate the closest wording from a truncated set of wording options to explain their fairness preference. Triangulating causes for why preference choices were made from qualitative analyses provides clues as to what we may empirically test in the future, but it does not allow us to infer causal reasoning. Additionally, people are complicated and may make decisions for multiple reasons. Structural equation models with simultaneous equations enable testing multiple causal pathways at once and comparing competing causal models[29, 49]. If participants were given the option to endorse multiple explanations, then we could test multiple causal pathways in parallel and compare them. Mediation analysis can also help statistically mitigate framing effects that are not eliminated through experimental design.

How people arrive at their preferences should also be considered. Qualitative approaches begin to help us understand the range of causes that might explain why people prefer fairness metrics within their specific paradigm, but experiments can go further [24, 64]. In situations where there may be multiple causes to consider, mediation analyses and structural causal models can help us to empirically test competing models of how the concepts interact [3]. This approach will facilitate the use of fairness enhancing interventions and what contextual factors should be considered when designing and deploying consequential ML models.

Several researchers highlight the potential influence of demographic context in algorithm application as well as selection of groups of people chosen to evaluate ML fairness in terms of fairness judgements [11, 42]. Social psychology has shown that perceptions about fairness and inequality in society vary across racial groups[61]. Many decisions that algorithms are deployed to make –from bail, to creditworthiness, to health screenings– affect ethnic and racial minority populations differently than majority populations [18, 45, 71]. Consistent with the idea that people who identify with a particular social group want their group to have resource advantages over others [65], a participant in [24] revealed in an open ended response that as a white person they wanted to choose the algorithm that is most fair to white people, but it is possible

that the methods obscured the possibility to see the effect of this at scale. Conversely, minority groups' preferences about algorithmic fairness and the contexts wherein they may prefer one fairness metric or another could be different from the majority of white M-Turk samples used by many of these studies [47]. Mturk workers are particularly unrepresentative of the larger U.S. population, tending to have lower average incomes, higher education levels, lower average ages, and smaller percentages of most non-white groups, particularly Black and African Americans [32, 59]. Additionally, experimental findings from several disciplines suggest that people from western, educated, industrialized, rich and democratic (WEIRD) societies vary considerably in most fundamental cognitive, social, and affective processes. Specifically, "WEIRD" societies rely more on analytical reasoning strategies than rules to explain or predict behavior and to make fairness decisions that distribute resources more equally. Alarmingly, top psychology journals found that 96% of subjects were from Western industrialized countries that represent only 12% of the world population. Despite this fact, findings in these publications are often taken to be broadly representative. Future studies could consider whether preferences generalize across many groups, variation in group disadvantage, and deployment contexts (eg a predominantly Black city as opposed to a predominantly white one). For example, the number of people needlessly jailed shifts radically when the bias affects 20% of the population versus 80% as would the number of fatalities due to inaccuracies diagnosing a variably prevalent disease. It is possible that these demographic realities have implications for people's algorithmic fairness preferences so understanding ML fairness will be improved by psychological expertise on inter-group relations.

3.4 Generalizability

All of the criminal justice scenarios in these studies rely on the COMPAS dataset collected by Propbulica [2] and evaluate perceptions of recidivism prediction, without acknowledging the challenges in that context beyond the algorithms [19]. Yu et al. [75] base their ML tutorial interface on a race- and gender- balanced datasets (1500 white, 1500 African American, 800 male, and 800 female). This might reflect an over-representation of groups compared to real-world examples, thereby causing bias in subjects' evaluations of fairness trade-offs separate from the real-world constructs of interest. Second, Yu et al. [75] focus on comprehension and evaluation metrics that might be more informative of ML problems in general, rather than those specific to the context of sentencing, racial bias, or criminal justice. The design assumptions regarding subjects making informed decisions through an interactive interface therefore might not adequately address differences in user-experience across different subsets of participants [53] or focus attention to the most relevant features of the decision context.

4 POWER, REPLICATION

In addition to the conceptual challenges of measuring human attitudes, we must assure that the statistical inferences are well powered for reliable results. In particular, psychology is facing a replication crisis: recent efforts to replicate key findings have not always succeeded[36]. Replication challenges are often attributed to strong pressure to publish in high impact outlets with reputations for publishing primarily novel positive findings [30] and discouraging negative findings and replications, leaving an incomplete and biased literature that is centered around positive findings [74]. In order to avoid the distrust that follows failed replication, an important interdisciplinary shift is necesary within psychology to detect irreproducible research, de-incentivize its publication, and positively promote reproducibility. Statistically underpowered studies and publication bias are deeply correlated. It is possible to calculate the likelihood of finding a positive result given a theorized effect and sample size, so low powered studies and those with a p-value of or near .05 paired with an unexpected result should be considered a red flag, indicating that this effect is unlikely to replicate at the time of peer review in order to filter-out irreproducible studies[74]. Without procedural changes, researchers are incentivized to employe practices that make their findings quickly publishable, even if they are not entirely representative of a true effect [14]. Statistically underpowered studies have increased odds of detecting a false effect and exaggerating the magnitude of true effects [14]. Studies often contain so many statistical tests that an acceptable number would be statistically significant even if the power of any single test was inadequate [36]. To correctly evaluate participants' preference for a human judge versus a model based on 12 conditions through six pairwise comparisons as [24] would require, for example, significance testing at a Bonferroni-corrected significance threshold – lowering the significance threshold from $p = 0.05$ to $p = 0.0041667$. By reporting trends with significance thresholds as high as p=0.08, the authors risk interpreting trends that could emerge due to statistical chance. These challenges are amplified for subgroups of participants; Harrison et al. [24] uses a convenience sample in which participants were included in the study simply because of ease of recruitment, leading to to an over-representation of white participants and reduced statistical power and bias against non-white groups, reducing the likelihood of generalization or replication. Practices like this could lead to interpreting spurious, non-replicating effects as significant [31]. Other studies, despite having smaller sample sizes, mitigate power issues by limiting the number of comparisons [64, 75] or employing statistical techniques that are appropriate for selecting between a large number of models (see use of Akaike information criterion in [64]).

5 DESCRIPTIVE WORK

An empirical approach to understanding perceptions of fairness is distinct from establishing a normatively correct standard of fairness. We argue that, augmenting the ML fairness literature with insights from psychology is a descriptive, not prescriptive, endeavor – one designed to better understand people's attitudes about algorithmic fairness, not to dictate which fairness notion is socially optimal. For example, the association between greater comprehension and decreased likelihood of agreeing with a fairness rule is not advice to make rules hard to understand or prevent the public from learning about them [57]. Better understanding the perceptions, and crucially, how they vary across social strata, over time, across contexts, will help better position technology. This work is of interest for computer scientists in serving to help make broader changes to AI design and can guide the development of tools that document how

algorithms work and what metrics they've been designed to accommodate, such as extending the concept of model cards [39]. This line of work can drive a positive social change by serving as a microcosm of society that is tractable for study [1]. Better descriptive work can then fuel future modeling, formalization, and ultimately new types of interventions that may prove exciting. Descriptive work affects understanding and therefore fundamentally impacts the problem formulation which has crucial implications for fairness of the resultant systems [48].

6 FURTHER CONTRIBUTIONS

We believe that interdisciplinary teams and diverse perspectives are necessary to address the dangers of discrimination by algorithms. We propose three ways in which psychological expertise can advance this work: (1) measure and compare the perspectives and needs of diverse stakeholders; (2) enable all stakeholders to participate in a conversation, and thereby help researchers understand the societal context in which algorithms are applied; (3) study how algorithms might change the very nature of discrimination in our society more broadly.

(1) Psychological insights highlight the importance of our goal, as stated in Section 5, to gain a better understanding of fairness perceptions more broadly from measurements, not to establish norms for application. Lowery et al. [35] revealed that the endorsement of restorative justice policy, using affirmative action as one example, is driven by how the policy affects the white ingroup; impact on the Black outgroup (positive or negative) is not considered. Hence, insight is to be gained by highlighting whether opinions shift by demographic group, or the perceptions of impact on the ingroup versus the outgroup. This could be reverse engineered to help elucidate latent biases in people's preferences for one type of fairness over another. Simply adopting the majority opinion without this insight would risk disregarding the needs of marginalized groups who are often disproportionately affected by biased algorithms and left without recourse [4, 5]. Social psychology can contextualize ethics: identity and the perception of fairness are deeply intertwined in determining when, why and how individuals think about matters of justice [17]. Considering different perspectives is especially important in contexts of competing models of justice, as is often the case with ML models. Extending from their logic, there is power in directly deciding which ML model is most fair. After all, it is often the case that the group who decides what is most objectively fair benefits most from the fairness decision. In the context of job candidate screening using ML, different stakeholders, such as job applicants, hiring managers, and researchers in psychology and ML, require different degrees for explainability of the model's decisions [34]. However, in order to ensure the acceptance of the ML decision tool by all stakeholders, especially the people who are negatively affected by the decision of the ML model (the rejected candidates), all must receive a sufficient and satisfying explanation for the decision. While rejected applicants might prioritize explainability, hiring managers probably prioritize other criteria, such as a high accuracy rate, for an ML model. As demonstrated, different stakeholders are likely to arrive at different definitions of fairness, considering trade-offs between explainabilty and accuracy [34]. It is important to measure all these varying preferences and perspectives reliably in order to understand and balance them.

(2) Understanding preferences and societal context in which the ML models would be deployed through stakeholder dialogue is essential. A narrow predictive approach may be overall insufficient in applications such as risk of appearing in court for a bail hearing [23, 58]. Contextual forms of discrimination that against certain populations would persist even with perfect accuracy. For example, underlying circumstances may interfere with court date attendance. Disadvantaged groups may face less resources to secure child care, transportation barriers, less flexible working hours, or other complications increasing the likelihood of failure to report on time for their appointment. Beyond exploring and discussing different fairness definitions, it is also necessary to explore, with all involved stakeholders, deployment contexts and consider nondeployment[23, 52, 58]. Such dialogues require an interdisciplinary perspective on AI and diverse set research methods, including qualitative methods [62]. Quantitative approaches run the risk of neglecting the context in which ML models are applied, given that the context under which the training data was collected is rarely recorded and must often be simplified. Individual behavior can only be fully understood in the context in which it occurs, so qualitative methods that complement the findings of quantitative research enhance the study of complex social environments[62]. Beyond understanding preferences, practitioners and communities should be engaged as equal partners throughout the development of ML systems to reduce discrimination.

(3) A psychological lens to understanding and mitigating algorithmic discrimination is a more practical approach than focusing solely on abstract ethical ideals. Social psychology aims to understand how people interact and can help explain and predict future social failures of AI before they get deployed. For example, Waytz and Schroeder [73] describe a passive process wherein people ignore or fail to identify the human mental capacities, perspectives, or lived realities of other individuals, called dehumanization by omission. The common choice to report average accuracy across all samples of a ML model is a prevalent example of dehumanization by omission in machine learning. Another relevant finding from social psychology is that people systematically disregard others who are irrelevant to their own goals [56]. This predicts both how participants in a study of how AI is perceived will respond, and how the individuals in the complex social systems that build AI work. Indeed, authors at Google and the Partnership on AI propose a framework for auditing algorithmic systems as they are developed that largely calls for reengineering the social processes from problem formulation to deployment [52]. Individuals from differing academic and professional disciplines tend to demonstrate systematic variation in hierarchy orientation[22]. Hierarchy orientation predicts and influences empathy and could have implications for whether people are motivated to ignore or seek to resolve issues of algorithmic biases and inequalities [60]. Social dominance orientation (SDO) is defined as the support of inequality between social groups and predicts racist and sexist ideologies and policies [51]. Since computing as a field is dominated by men, who tend to score higher on SDO, it is likely (though yet unknown) that norms of computing as a discipline would mirror these policies. Within

computing, many of those most critical of purely algorithmic interventions have been Black women and others from marginalized backgrounds. Timnit Gebru is known for revealing biases in joint work with Joy Buolamwini and Deb Raji [13] and proposing transparency and documentation focused interventions [21, 26, 39, 52]. In contrast, those providing purely theoretical analyses and provable algorithmic interventions have been predominantly male and white. With the context of human behavior, the impact of the ignored dimensions of society is not surprising in its harm to groups who have been historically, currently, and systematically excluded from it. Another factor that has the potential for harm in the future is the fact that algorithms seems to decrease culpability [55]. Telling participants that discrimination was the result of implicit attitudes as opposed to explicit ones, reduced their sense that anyone should be blamed or punished. This effect was more pronounced for discrimination that was attributable to algorithms [55].

However, if decisions about whether to make algorithms more equitable across groups are enacted by people who lack empathetic concern for individuals they find irrelevant to their own goals, then algorithms become a causal mechanism between discriminatory individuals who are in-turn held less accountable for the discrimination resulting from the mechanism through which they caused it. Thus, the link between hierarchy orientation, empathic concern, and preferences for algorithmic fairness metrics needs to be explored. Future research should not focus solely on identifying and correcting algorithmic bias, but also better understanding the implications of these disproportionate outcomes for society as a whole [55].

7 CONCLUSION

Under realistic conditions, statistical definitions of group-wise fairness are mutually-exclusive and therefore enforcing one type of fairness requires allowing other types of bias. To help narrow the scope of AI fairness, researchers have investigated which notions of fairness reflect real-world perceptions and understanding of fairness [24, 57, 64, 75]. However, social constructs such as fairness preferences are heavily dependent on context, framing, and social factors such as demographics [9, 66–70, 72]. Additionally, researchers must take proper statistical precautions in order to measure these preferences reliably and accurately: under-powered designs, sampling problems, publication pressure, and demographic assumptions can all lead to studies that fail to replicate [14, 36, 74]. Therefore, it is crucial to be sensitive to existing psychological literature (i.e., framing) as well as appropriate statistical methods when applying psychological insights to AI fairness. Insights from methods used in Judgment and Decision Making, Social Psychology, and Cognitive Psychology should be considered in order to prevent ML researchers from unnecessarily repeating the same methodological mistakes that have already been problematized in those fields. We leverage well-established psychological literature on framing and statistical replication to critically evaluate current work in AI fairness. While the current state-of-the-art provides researchers an excellent starting-point for better understanding AI fairness, there is still significant work ahead in creating generalizable, reliable assessment tools. Adopting an interdisciplinary approach - one that incorporates not only insights from psychology, statistics, philosophy, and other fields but also perspectives of practitioners such as nurses and social workers - is necessary in order to create fairer AI. Such an approach will appropriately address this societal challenge, and thus ensure fulfilling our obligation to future generations in implementing a responsible and fair use of algorithmic techniques.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 252–260. https://doi.org/10.1145/3351095.3372871

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[4] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[5] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social Forces* (2019).

[6] Avrim Blum and Kevin Stangl. 2020. Recovering from biased data: Can fairness constraints improve accuracy?. In *1st symposium on foundations of responsible computing (FORC 2020)*. tex.organization: Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[8] Christopher J Boyce, Alex M Wood, and Eamonn Ferguson. 2016. Individual differences in loss aversion: Conscientiousness predicts how life satisfaction responds to losses versus gains in income. *Personality and Social Psychology Bulletin* 42, 4 (2016), 471–484.

[9] João N Braga, Mário B Ferreira, and Steven J Sherman. 2015. The effects of construal level on heuristic reasoning: The case of representativeness and availability. *Decision* 2, 3 (2015), 216.

[10] João N Braga, Mário B Ferreira, Steven J Sherman, André Mata, Sofia Jacinto, and Marina Ferreira. 2018. What's next? Disentangling availability from representativeness using binary decision tasks. *Journal of Experimental Social Psychology* 76 (2018), 307–319.

[11] Meredith Broussard. 2018. *Artificial unintelligence: How computers misunderstand the world.* mit Press.

[12] Timothy A Brown. 2015. *Confirmatory factor analysis for applied research.* Guilford publications.

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[14] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience* 14, 5 (2013), 365–376. Publisher: Nature Publishing Group.

[15] Nick Chater, Jian-Qiao Zhu, Jake Spicer, Joakim Sundh, Pablo León-Villagrá, and Adam Sanborn. 2020. Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science* 29, 5 (2020), 506–512.

[16] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163. Publisher: Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.

[17] Susan Clayton and Susan Opotow. 2003. Justice and identity: Changing perspectives on what is fair. *Personality and social psychology review* 7, 4 (2003), 298–310.

[18] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc* (2016).

[19] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209.

[20] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589 event-place: Atlanta, GA, USA.

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé, and Kate Crawford. 2018. Datasheets for Datasets.

arXiv (2018). http://arxiv.org/abs/1803.09010

[22] Hillary Haley and Jim Sidanius. 2005. Person-organization congruence and the maintenance of group-based social hierarchy: A social dominance perspective. *Group Processes & Intergroup Relations* 8, 2 (2005), 187–203.

[23] Moritz Hardt. 2020. Fairness and Machine Learning: Limitations and Opportunities [Conference Presentation]. SPSP Presidential Plenary: Bias in the Age of AI and Big Data; Chair: Rodolfo Mendoza-Denton. https://www.youtube.com/watch?v=S2knhRMZRuI

[24] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.

[25] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830 Place: Glasgow, Scotland Uk tex.numpages: 16.

[26] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 306–316.

[27] Karl G Jöreskog, Ulf H Olsson, and Fan Y Wallentin. 2016. *Multivariate analysis with LISREL*. Springer.

[28] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[29] Rex B Kline. 2015. *Principles and practice of structural equation modeling*. Guilford publications.

[30] Sander L. Koole and Daniël Lakens. 2012. Rewarding Replications: A Sure and Simple Way to Improve Psychological Science. *Perspectives on Psychological Science* 7, 6 (2012), 608–614.

[31] Jeffrey T Leek and Roger D Peng. 2015. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences* 112, 6 (2015), 1645–1646. Publisher: National Acad Sciences.

[32] Kevin E Levay, Jeremy Freese, and James N Druckman. 2016. The demographic and political composition of Mechanical Turk samples. *Sage Open* 6, 1 (2016), 2158244016636433.

[33] Sarah Lichtenstein, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs. 1978. Judged frequency of lethal events. *Journal of experimental psychology: Human learning and memory* 4, 6 (1978), 551.

[34] Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. 2018. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning*. Springer, 197–253.

[35] Brian S Lowery, Miguel M Unzueta, Eric D Knowles, and Phillip Atiba Goff. 2006. Concern for the in-group and opposition to affirmative action. *Journal of personality and social psychology* 90, 6 (2006), 961.

[36] Scott E Maxwell. 2004. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological methods* 9, 2 (2004), 147.

[37] Barbara J McNeil, Stephen G Pauker, Harold C Sox Jr, and Amos Tversky. 1982. On the elicitation of preferences for alternative therapies. *New England journal of medicine* 306, 21 (1982), 1259–1262. Publisher: Mass Medical Soc.

[38] Beth E Meyerowitz and Shelly Chaiken. 1987. The effect of message framing on breast self-examination attitudes, intentions, and behavior. *Journal of personality and social psychology* 52, 3 (1987), 500. Publisher: American Psychological Association.

[39] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[40] Kellen Mrkva, Eric J Johnson, Simon Gächter, and Andreas Herrmann. 2020. Moderating loss aversion: loss aversion has moderators, but reports of its death are greatly exaggerated. *Journal of Consumer Psychology* 30, 3 (2020), 407–428.

[41] Nadia Hanin Nazlan, Sarah Tanford, and Rhonda Montgomery. 2018. The effect of availability heuristics in online consumer reviews. *Journal of Consumer Behaviour* 17, 5 (2018), 449–460.

[42] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

[43] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. tex.publisher: American Association for the Advancement of Science.

[44] Mathias Osmundsen and Michael Bang Petersen. 2020. Framing Political Risks: Individual Differences and Loss Aversion in Personal and Political Situations. *Political Psychology* 41, 1 (2020), 53–70.

[45] Kellie Owens and Alexis Walker. 2020. Those designing healthcare algorithms must become actively anti-racist. *Nature medicine* 26, 9 (2020), 1327–1328.

[46] Zoe Oxley. 2020. Framing and Political Decision Making: An Overview. In *Oxford Research Encyclopedia of Politics*. Oxford University Press.

[47] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science* 23, 3 (2014), 184–188.

[48] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 39–48.

[49] Judea Pearl. 2012. *The causal foundations of structural equation modeling*. Technical Report. CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.

[50] Philip M Podsakoff, Scott B MacKenzie, Jeong-Yeon Lee, and Nathan P Podsakoff. 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology* 88, 5 (2003), 879.

[51] Felicia Pratto, Jim Sidanius, Lisa M Stallworth, and Bertram F Malle. 1994. Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of personality and social psychology* 67, 4 (1994), 741.

[52] 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873 event-place: Barcelona, Spain.

[53] Richard Rau, Erika N Carlson, Mitja D Back, Maxwell Barranti, Jochen E Gebauer, Lauren J Human, Daniel Leising, and Steffen Nestler. 2019. What is the structure of perceiver effects? On the importance of global positivity and trait-specificity across personality domains and judgment contexts. *Journal of Personality and Social Psychology* (2019).

[54] Tenko Raykov. 2007. Reliability if deleted, not 'alpha if deleted': Evaluation of scale reliability following component deletion. *Brit. J. Math. Statist. Psych.* 60, 2 (2007), 201–216.

[55] Jennifer A Richeson. 2020. The Mythology of Racial Progress [Conference Presentation]. SPSP Presidential Plenary: Bias in the Age of AI and Big Data; Chair: Rodolfo Mendoza-Denton. https://www.youtube.com/watch?v=S2knhRMZRuI

[56] Miriam J Rodin. 1987. Who is memorable to whom: A study of cognitive disregard. *Social Cognition* 5, 2 (1987), 144–165.

[57] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*. PMLR, 8377–8387.

[58] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 59–68.

[59] Kim Bartel Sheehan. 2018. Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Communication Monographs* 85, 1 (2018), 140–156.

[60] Jim Sidanius, Nour Kteily, Jennifer Sheehy-Skeffington, Arnold K Ho, Chris Sibley, and Bart Duriez. 2013. You're inferior and not worth our concern: The interface between empathy and social dominance orientation. *Journal of personality* 81, 3 (2013), 313–323.

[61] Jim Sidanius and Felicia Pratto. 2001. *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

[62] Mona Sloane and Emanuel Moss. 2019. AI's social sciences deficit. *Nature Machine Intelligence* 1, 8 (2019), 330–331.

[63] Paul Slovic, John Monahan, and Donald G MacGregor. 2000. Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and human behavior* 24, 3 (2000), 271–296.

[64] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2459–2468.

[65] Henri Tajfel, Michael G Billig, Robert P Bundy, and Claude Flament. 1971. Social categorization and intergroup behaviour. *European journal of social psychology* 1, 2 (1971), 149–178.

[66] Yaacov Trope and Nira Liberman. 2010. Construal-level theory of psychological distance. *Psychological review* 117, 2 (2010), 440.

[67] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology* 5, 2 (1973), 207–232.

[68] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science* 211, 4481 (1981), 453–458. Publisher: American Association for the Advancement of Science.

[69] Amos Tversky and Daniel Kahneman. 1986. uRational choice and the framing of decisions. Journal of business, 59 (4), part 2. *S251) S275* (1986).

[70] Amos Tversky and Daniel Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics* 106, 4 (1991), 1039–1061. Publisher: MIT Press.

[71] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, explainable, and accountable AI for robotics. *Science Robotics* 2, 6 (2017). https://doi.org/10.1126/scirobotics.aan6080 arXiv:https://robotics.sciencemag.org/content/2/6/eaan6080.full.pdf

[72] Cheryl Wakslak and Yaacov Trope. 2009. The effect of construal level on subjective probability estimates. *Psychological Science* 20, 1 (2009), 52–58.

[73] Adam Waytz and Juliana Schroeder. 2014. Overlooking others: Dehumanization by comission and omission. *TPM: Testing, Psychometrics, Methodology in Applied Psychology* 21, 3 (2014).

[74] Bradford J Wiggins and Cody D Chrisopherson. 2019. The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology* 39, 4 (2019), 202.

[75] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM designing interactive systems conference*. 1245–1257.