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

**Sophia T. Dasch**[1,2], **Vincent B. Rice, Jr**[1], **Venkat R. Lakshminarayanan**[3], **Taiwo A. Togun** [3], **C. Malik Boykin**[1], **Sarah M. Brown** [4]

[1] Brown University
Providence, RI, USA

[2] Humboldt University of Berlin
Berlin, Germany

[3] SeqHub Analytics LLC
New Haven, CT, USA

[4] University of Rhode Island
Kingston, RI, USA

## 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 the development of ML. 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. We consider this research to be of a descriptive nature, enabling a deeper understanding and a substantiated discussion.

## 1 Introduction

Advances in Artificial Intelligence (AI) are now a matter of public life. In the course of daily life, a person may now interact with AI multiple times in both high and low stakes aspects of their life. As decisions are outsourced to machine learning (ML) powered AI systems, the need for facing the challenge of creating fair decision-making processes is becoming ever more pressing. Unfair decision making by AI systems occurs in many domains: Gender recognition systems fail systematically on darker-skinned faces, healthcare algorithms codify existing biases and amplify health disparities, popular natural language processing models perpetuate stereotypes, and algorithms used in sentencing and bail decisions disproportionately label Black people as high risk incorrectly and white people as low risk incorrectly (Buolamwini and Gebru, 2018; Obermeyer et al., 2019; Angwin et al., 2016). This is especially important because there is a risk that marginalized groups in particular will be more severely impacted by these algorithmic techniques (Barocas and Selbst, 2016). In response, machine learning researchers have formalized fairness by writing it as a constraint that can be

NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA).

added to a learning algorithm (Friedler et al., 2019; Barocas et al., 2019). Many definitions exist, each with different social and political values, but they are mutually exclusive and cannot co-exist (Chouldechova, 2017; Kleinberg et al., 2016). In order to deploy any of these fair machine learning interventions, a practitioner must choose a type of fairness to employ, a decision that technical expertise alone does not prepare one to make, some additional information is needed to make that decision.

The approach to constrain learning with a second objective is unlikely to be sufficient; we will need a broader understanding to prevent ML-powered AI from perpetuating discrimination. Practitioners and social scientists alike note that individual learning algorithms are embedded in larger systems that also require attention (Holstein et al., 2019; Selbst et al., 2019). Practitioners have seen that more data can improve fairness, but how much is good enough remains an open question (Holstein et al., 2019). 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 (Selbst et al., 2019). For example, the protected attributes are generally modeled as exchangeable and independent even though they are intended to capture and represent aspects of individuals' identities that are instead hierarchical and interacting (Hanna et al., 2020).

However, under some assumptions of how the bias occurs, fairness constraints can provide the desired outcome. For example, assuming that the training labels are incorrect at different rates for different groups of people, adding a specific fairness constraint (equal opportunity) can improve performance on the base task and overcome the errors (Blum and Stangl, 2020). This demonstrates that we need to understand the bias and the way it occurs as well. Further, computing can have broader roles in society than building new technologies; computing can, for example, serve as a diagnostic, formalizer, for rebuttal, or synecdoche as well (Abebe et al., 2020). For example, fairness definitions may carry a greater utility as a diagnostic than as an intervention, for metric for quantifying and monitoring how real-world systems are serving. In this light, understanding how these definitions are perceived can serve as a formalizer and synecdoche, helping make explicit and timely how individuals and groups *express their biases*. This understanding can guide policy discussions and public education to improve literacy of the technologies that impact people's daily lives.

Our main contribution is to demonstrate the relevance of a thoroughly considered research design that allows measuring participants' true fairness preferences, understanding underlying mechanisms, ensuring real-world applicability and replicable results. Moreover, we also want to emphasize that this research is of a descriptive nature. We do not propose that practitioners' direct recommendations be based on majority perceptions of ML fairness. First and foremost, this research should lead to a clearer understanding of the situation to enable subsequent discussions on implementing fairness criteria in practice with a more substantiated scientific foundation. In the following, we will present the initial papers and their used approaches to explore this research challenge. Afterwards, we will highlight the important opportunities of an interdisciplinary approach by emphasizing the benefits that a psychological perspective could bring to this research.

## 2 Perceptions of Fair Machine Learning

Largely inspired by the need to understand the impact of ML beyond the traditional performance metrics, several recent studies have evaluated preferences across fairness definitions, the degree to which people understand fairness definitions, and the inherent design trade-offs when considering fairness in an ML system (Harrison et al., 2020; Saha et al., 2020; Yu et al., 2020; Srivastava et al., 2019). In this section we provide summaries of the key questions, design choices, and findings of these studies, reserving critique or evaluation of the studies for the subsequent sections.

Harrison et al. (2020) aims to directly evaluate perceptions of fairness. Their study includes 502 Mechanical Turk workers. They present participants with scenarios based on the COMPAS dataset. Harrison et al. (2020)'s study investigated participants' preference between multiple competing notions of fairness: accuracy, false positive rate (FPR), outcomes, and consideration of race in decision making. They assessed these notions of trade-off by giving participants choices between competing models of bail eligibility. In each choice, participants could choose 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. By providing similar choices pairing each of their four notions of fairness, and observing consistency in choices across participants, Harrison et al. (2020) investigated which fairness notion was most-preferred by subjects. They concluded that, when given a choice between equalizing accuracy and equalizing FPR (the chance of being mistakenly denied bail), subjects prefer equalizing FPR. However, as discussed in Section 4, their conclusions are based on a weak preference that may fail to emerge in other studies.

Saha et al. (2020) aim to assess non-experts' comprehension of fairness metrics, examining the degree of comprehension and factors that influence comprehension. They conduct a preliminary study with three scenarios: hiring, giving employees awards, and judging a student art project to validate the method. They observed no variation in comprehension based on the context scenario, so they restrict the larger studies of the main research questions to only one scenario. They recruit participants roughly matching the US population on Clint, 147 for the validation study and 349 for the main study. In their main study they provide visual representations that illustrate the fairness definitions as a decision rule that a decision maker must follow in order to be fair for participants. Participants were asked if they agreed with and if they liked each rule on two separate Likert scales. Then comprehension is evaluated through a series of comprehension questions presented in several forms. Some ask the participant to apply the rule to determine how many offers to sent, other questions ask participants to consider true/false statements about how the fairness rule relates to hiring based on merit and other factors.

Saha et al. (2020) found that comprehension scores are lower for rules representing certain fairness definitions rather than others. Based on their comprehension scores, participants had a harder time understanding fairness rules representing the definitions “equal opportunity” and “false negative rate (FNR),” compared to other fairness definitions (for example, demographic parity). Additionally, they found that FNR generated the greatest disagreement between participants in their comprehension scores. Finally, they found that those with the greatest comprehension scores tended to show the greatest negative sentiment towards a given fairness rule.

Srivastava et al. (2019) apply 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 hypothesize that distinct notions will be preferred across different contexts, but find that demographic parity is simply preferred in both recidivism risk assessments and health skin cancer risk assessment.

In addition to eliciting fairness preferences, the authors also gather explanations from participants regarding their choices. They recruited 20 MTurk volunteers for a pilot study in which they establish participants' preferences between two explanation interfaces (a free response text box was preferred over a structured text box), validating the interface they selected for the final study.

Then, in their main studies, the authors assess participants' perceptions of fairness using an adaptive experimental design. Paid MTurk workers are asked to judge the fairness of algorithms designed to predict a defendant's recidivism risk (Context 1) and skin cancer assessment (Context 2), with 100 participants evaluating each context. In each context, the algorithms vary in how discriminatory or fair they are along several dimensions of fairness: demographic parity, error parity, false discovery rate, and false negative rate.

In each test set participants are provided pairwise choices between hypothetical predictions, and then shown the racial and gender breakdown of the two sets of predictions. Each test includes the same set of 10 individuals (3 white men, 2 white women, 2 Black men, 3 Black women) with true outcomes and the predictions for each choice. Participants are then asked to to choose which algorithm is more discriminatory.

Each pairwise test is administered by an adaptive algorithm that varies the parameters of the participant's choice set (i.e., degree of discrimination, dimension of fairness). Simulation is used to demonstrate that of the 9262 possible tests, random presentation order would require 600 tests to have a high confidence prediction of the participants preference – but only 20 are required for the adaptive presentation procedure. Therefore, each participant receives up to 20 tests. The testing stops when the algorithm determines a participant's preferences along each dimension of fairness. The authors then evaluate fairness preferences with a survey, where participants are presented 3 algorithms that represent different positions in an overall accuracy/fairness space. One is highly accurate with a large gender disparity, the second is moderately accurate with a moderate gender disparity, and the third is the least accurate but equal across genders.

Yu et al. (2020) propose a tool to help algorithm designers understand the trade-offs, which they validated based on the responses of 301 MTurk workers. They reexamine the classical trade-offs between types of errors and between accuracy and fairness. To illustrate the trade-offs, they produced a set of Pareto optimal predictors on on racially and gender balanced subsets of the nonviolent offenders COMPAS dataset (Angwin et al., 2016). After participants interact with an instructive tool to aid in their understanding, the authors assess their comprehension of inherent fairness and accuracy trade-offs. They compared subjects' evaluations and comprehension of the model's results across three conditions: a 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). They then measured participants along to three separate metrics: objective comprehension (their score on a multiple-choice test), subjective evaluation (their ranking of their understanding on a Likert scale), trust (their perceived trust score on a Likert scale) and additional information through expert interviews. They also asked participants to choose which model that they preferred (i.e., which was most consistent with their values) while allowing them to interactively view the statistical consequences of their chosen model (i.e., the FPR versus FNR Pareto frontier, as well as the fairness versus accuracy Pareto frontier). Yu et al. (2020) found that both interactive views offered a statistically significant increase in comprehension. Additionally, they found that while half of the subjects changed their level of trust in the algorithms predictions after experience with an interactive tool, that subjects were approximately as likely to increase their trust (22.3%) as reduce their trust (25.1%).

Based on their findings, Yu et al. (2020) propose a two-step design process in improving participants' comprehension of these inherent trade-offs when making decisions of an algorithm's fairness. First, they propose that they explicitly make participants aware of these trade-offs in an interpretable, accessible fashion – for example, through interactive demos, visualizations, and gamification. Second, they propose that instrument designers formally identify the trade-offs between competing models of fairness, for example, by comparing models at Pareto frontiers. By neglecting these principles, Yu et al. (2020) argue that researchers investigating algorithmic fairness might bias their participants understanding of fairness trade-offs, biasing their decisions.

## 3 Measurement

It is a complicated endeavor to measure human attitudes, since they can rarely be measured directly, but must instead be captured indirectly using a measuring instrument. One often under-considered assumption in measuring attitudes and behaviors is that some quality within the respondents is causing their answers (Jöreskog et al., 2016). However, there is often a part of that variance that can be attributed to the measuring instrument or method that researchers choose to use (Podsakoff et al., 2003). This highlights the need that researchers should consider the signal-to-noise ratio, wherein a portion of the noise is attributable to the actual method itself – this is known as method variance. The decisions which measuring instrument is chosen, which manifest variables are selected to describe the latent constructs, and which potential mechanisms are assessed, are crucial to whether the study results will be reliable and reflect the true preferences of the participants or not.

Moreover, it is of particular importance that the generated results are generalizable, thus can be applied to the real-world, as these studies should provide guidance for practitioners as soon as possible. Psychological research has developed extensive expertise in measuring human attitudes. And in the following we will point out some weaknesses of the studies on ML fairness conducted so far, to demonstrate why the discipline of psychology brings a valuable perspective to this discussion.

### 3.1 Reliability

To assess comprehension of ML fairness, Saha et al. (2020) uses a questionnaire and reports Chronbach's Alpha reliability coefficients between .38 and .64, none of which meet the recommended .70 threshold for reliable measurement (Kline, 2015). 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 measurement practice because it is unlikely that this change in reliability locally generalizes from the sample to the population. This artificially inflates the reliability of the measure for reporting as it deflates the interpretability of the results (Raykov, 2007). Tinkering with scale measurements during analyses is an inherently exploratory method that needs to be reproduced to increase interpretability(Brown, 2015). In either case, analyzing the covariance of remaining items with a latent variable model is the recommended practice (Brown, 2015; Raykov, 2007). The

<table>
<tr><th></th><th>Harrison et al. (2020)</th><th>Yu et al. (2020)</th><th>Srivastava et al. (2019)</th><th>Saha et al. (2020)</th></tr>
<tr><td>Target Study (2)</td><td>Preferred fairness standard</td><td></td><td>High versus low stakes</td><td>Comprehesnion of fairness tradeoffs</td></tr>
<tr><td>Recruiting (2)</td><td>MTurk</td><td>MTurk</td><td>MTurk</td><td>Cint</td></tr>
<tr><td>Num Participants (2)</td><td>502</td><td>87-100 / conditon</td><td>100 / condition</td><td>147 (validation), 349 (main)</td></tr>
<tr><td>Reliability (3.1)</td><td>Sampling bias reduces reliability</td><td></td><td>Qualitative methods can't establish statistical cause</td><td>Cronbach's Alpha of .38 to .64</td></tr>
<tr><td>Framing(3.2)</td><td colspan="2">Confound: framing effects (loss aversion& level of abstraction)</td><td></td><td></td></tr>
<tr><td>Mechanisms (3.3)</td><td colspan="2">Confound: sampling bias (via mechanisms like social identity)</td><td rowspan="2">Quantative follow-ups (like mediation analysis) needed to understand why context changes preferences</td><td></td></tr>
<tr><td>Generalizability (3.4)</td><td>Sampling bias &amp; uncorrected significance levels reduce generalizability</td><td rowspan="2">"Demographically balanced" instead of representative scenarios</td><td rowspan="2">Scale items deleted to inflate reliability</td></tr>
<tr><td>Power and Replication (4)</td><td>High uncorrected significance cut-offs - findings could be artifactual<br>Convenience sample - selected for ease of recruitment, rather than likelihood of replication<br>Reduced statistical power for under-represented sub-groups</td><td>Small number of multiple comparisons maintains statistical power</td></tr>
</table>

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.

reliability limitation of this study clearly shows that this research could benefit from a psychological perspective on measurement methods and, in particular, an interdisciplinary pool of reviewers can provide the necessary expertise to evaluate human subject experimental work.

### 3.2 Framing Effects

One way researchers make use of measurement instruments is observing participants' choices in a questionnaire or survey, and from patterns in those choices, assessing preferences. While answering a survey, participants must continuously make decisions between different options. For example, in the paper by Harrison et al. (2020), participants must decide between a Model X and a Model Y, which both offer different trade-offs between fairness criteria. Both fairness criteria are presented in a particular way, but insufficient consideration is given to the fact that this representation of the various choice options strongly influences the decision-making process. The manner in which particular choices are perceived by the agent involved in the decision-making process is referred to as the "decision frame" (Tversky and Kahneman, 1981). The framing of a choice is determined by many different factors, such as individual characteristics of the decision-maker, but also by external factors, such as the presentation of a particular choice, and its anticipated outcomes and eventualities (Tversky and Kahneman, 1981). Since different choice options can nearly always be presented in various ways, the choice of the decision frame is crucial, as it most likely will influence the participants' responses. This challenge has been addressed by psychological research for several decades, producing a vast literature ranging from classical studies about framing effects in risk communication to recent efforts of finding an integrated explanation for different underlying mechanisms of framing effects (Tversky and Kahneman, 1981; McNeil et al., 1982; Slovic et al., 2000; Chater et al., 2020; Oxley, 2020). 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(Osmundsen and Petersen, 2020; Boyce et al., 2016). This expertise can contribute to avoiding confounding variables from the onset or to identify and tackle the contamination in constructive ways. In the following, we will discuss potential framing effects in the paper by Harrison et al. (2020) to illustrate our argument.

### 3.2.1 Loss Aversion

In Harrison et al. (2020)'s paper there are two fairness definitions which are framed in a particularly stark contrast to each other: "Accuracy" versus "Mistakenly Denied Bail". We argue that this particular decision frame results in an asymmetrical weighting of the fairness criteria presented to participants – giving one fairness criterion indirectly more weight than the other one. We base our assumptions on the well-established effect that losses and disadvantages seem to play a more important role in decision making for individuals than gains and advantages (Tversky and Kahneman, 1991; Mrkva et al., 2020). Individuals weigh losses more strongly than equally-sized gains (Tversky and Kahneman, 1986). This tendency to overvalue losses is called loss aversion (Tversky and Kahneman, 1986). Often the potential outcomes of a particular decision can be expressed both as gains and losses. Since the tendency to avoid losses is asymmetrically stronger than the desire to seek equally-sized gains, individuals will give greater weight and respond more strongly when the same option is framed as a loss rather than a gain. The following findings will illustrate this effect further.

McNeil et al. demonstrate 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 which one they prefer. However, the treatment alternatives were framed differently in the experimental conditions. 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 of the surgery (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. When the treatment mortality rate was given (e.g. 10 percent), the tendency to avoid loss is more strongly triggered. This tendency to react more strongly to losses is also evident in other studies. Meyerowitz and Chaiken showed peoples' motivation to perform a breast self-examination depends on the framing of cancer outcomes. Authors presented participants with a pamphlet that either emphasized the increased cancer risk associated with omitting the self-exam (loss frame), or the equivalent information framed in terms of the reduced risk associated with conducting the self-exam (gain frame). 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 frames versus losses frames) come into effect in Harrison et al. (2020) 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 manner in which fairness criteria are presented is important because it could potentially influence the participants' responses. "Accuracy" is a positive attribute and should be increased, whereas "Mistakenly Denied Bail" is a negative attribute of the model, a potential mistake, and should be avoided. Equal accuracy could also be framed as a negative attribute, as equal error rates instead. 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 (Harrison et al., 2020). 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.

### 3.2.2 Level of Abstraction

Asking participants to compare "Mistakenly Denied Bail" to "Accuracy" also makes a comparison across levels of abstraction, since "Mistakenly Denied Bail" is framed as a very concrete event, while accuracy is an abstract metric whose concrete meaning can hardly be deciphered without additional context. Besides the aforementioned effect of loss aversion, this framing aspect might further cause that the fairness criteria of equalized FPR, "Mistakenly Denied Bail", is overvalued compared to an equalized accuracy rate. In the following, we argue that the participants might have intuitively relied on the availability heuristic, thus overestimating the probability of the occurrence that a person is mistakenly denied bail. The availability heuristic describes 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 (Tversky and Kahneman, 1973). This relationship between the ease of retrieval and the estimation of frequency has been demonstrated in many studies (Tversky and Kahneman, 1973; Lichtenstein et al., 1978; Nazlan et al., 2018; Braga et al., 2018). In one experiment by Tversky and Kahneman (1973), participants were asked to estimate whether words with the letter 'R' in the first or third position were more frequent. To answer this question, they relied on the availability heuristic. Naturally, it is much easier to retrieve words that have the particular letter in the first position than in the third position. This experiment was performed with 5 consonants (K,L,N,R,V), and the majority of the subjects expected that words in which the respective consonant is at the beginning are more frequent. In fact, the reverse is true: words in which the particular 5 consonants are in the third position are more frequent. The increased ease of recalling words with the consonants in the first place falsely led participants to believe that those words were more commonly used. Another study demonstrated that the frequency of lethal events was overestimated by participants if events were more concrete and easier to imagine (Lichtenstein et al., 1978). Presumably, the availability heuristic caused participants to give more weight to these occurrences.

Trope and Liberman (2010) proposed that an object can be expressed on different levels of abstraction, levels of construal. High levels of construal refer to abstract representations, whereas low levels of construal refer to more concrete representations. In order to illustrate these levels, the authors use the term "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. According to a study by Braga et al. (2015), the level at which an event is described, whether abstract or concrete, affects decision-making processes. A low level of construal is more likely to lead to the application of the availability heuristic. Moreover, experiments by Wakslak and Trope (2009) demonstrated that it is sufficient to manipulate the mindset regarding the construal level to influence participants' probability estimates. Subjects who were induced into thinking more concretely, using a low-level-construal mind-set, estimated events as more likely than subjects, who were led to think more abstractly. Again, this study shows that there seems to be a connection between the level of abstraction in which an event is conceptualized and probability estimates of its occurrence.

When the two fairness criteria presented in Harrison et al. (2020)'s study are compared, it becomes apparent that the accuracy rate is described at a higher level of abstraction than FPR, "Mistakenly Denied Bail". The mere term "Accuracy" would hardly be enough to identify the concrete consequences of this fairness criterion, rather background information is needed to know that "Accuracy" is the measure for correctly categorized people who were either correctly granted bail or correctly denied bail. This decoding process reveals the higher level of abstraction, whereas the term "Mistakenly Denied Bail" immediately evokes a very concrete scenario. Due to the increased ease of retrieving relevant examples, participants may rely on the availability heuristic and thus intuitively overestimate the probability of occurrence of this particular event. Such an overestimation may lead to an inappropriate evaluation of the significance of this mistake. As a result, participants would again be tempted to give more weight to this fairness criterion when deciding between Model X or Y. As mentioned above, participants actually tended to perceive the model as fairer, in which the criterion "Mistakenly Denied Bail" was equalized. However, the used study design does not provide sufficient information to determine whether this tendency is due to a true preference or to the particular representation of the two models. This dilemma again highlights the importance of a deliberate reflection on potential framing effects of the measurement instruments and it demonstrates the value of a psychological perspective on these challenges. The paper by Harrison et al. (2020) illustrates that an experimental design has the potential to produce misleading results if framing effects are not carefully considered. This is a cautionary tale: given the high societal importance of this research and the rapid entry of ML powered AI into high-stake domains, it of utmost importance to quickly generate reliable results to advance the debate on ML Fairness.

### 3.3 Mechanisms

Experiments can also help us to understand why people are making the choices they are making. Harrison et al. and Srivastava et al. help to give us clues about where to look for causal pathways that may lead to people's choices of fairness metrics. Harrison et al. (2020) demonstrates that when people are presented with options of fairness metrics, they make comparative choices and Srivastava et al. (2019) shows us that these choices are not acontextual. In their study, when the stakes were higher, when outcomes were considered, or when serious health concerns were involved, people made consistently different choices between algorithmic accuracy versus groupwise fairness. This highlights that the gravity of the decision and its' impact on people's lives have implications for people's preferences that should be considered in future studies and in 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? All of these questions may matter and should be explored in future experiments.

Because a large number of important causal factors - spurious causes, in addition to causes of interest - can influence a 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, can be useful in understanding the influence of the factors of interest. For example, in Srivastava et al. (2019), multiple factors could be at play in driving preference for FNR-based-equality instead of accuracy-based equality. 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 decision for multiple reasons. Through the use of simultaneous equations, structural equation models give us the flexibility to test multiple causal pathways at once or to compare competing causal models(Kline, 2015; Pearl, 2012). If participants were given the option to endorse multiple explanations, then we could test multiple causal pathways in parallel or comparatively. Aside from a true conviction about a fairness dimension, factors such as framing (for example, the abstractness or construal level) could be a mediator - causally involved in the relationship between the fair/unfair scenario and the participant's response to it.

The question of how people arrive at their preferences should also be considered. Harrison et al. and Srivastava et al. both used qualitative approaches to help us understand the range of causes that might explain why people prefer fairness metrics within their specific paradigm. Importantly, in highlighting that features of choices lead people to choose different metrics, we should be designing experiments that empirically test the causal mechanisms hinted at in these designs. And 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 may fit together (Barocas et al., 2019). This approach will help designers to better know when it may be more important to consider fairness metrics and what kinds of fairness metrics to consider given the context in which a decision model will be implemented. We also need to know if these choices are due to whimsical causes such as method variance and question framing, or due to the true convictions people have about fairness, equality, and the gravity of the decision being made, and the people affected.

Furthermore, the demographic context in which the algorithm is to be applied as well as groups of people who evaluate the ML fairness are also potential mechanisms, which might heavily influence people's fairness judgements. Participants from different racial groups have differing perceptions about fairness and inequality in society (Dominance, 1999). Many kinds of decisions that algorithms are deployed to make decisions about –from bail, to creditworthiness, to health screenings– affect ethnic and racial minority populations differently than majority populations (Dieterich et al., 2016; Owens and Walker, 2020; Wachter et al., 2017). Consistent with the idea that people who identify with a particular social group want their group to have resource advantages over other groups (Tajfel et al., 1971), a participant in Harrison et al. (2020) revealed in an open ended response that as a White person they wanted to choose the algorithm that is most fair to White people. While this concept was engaged in Harrison's study, it is possible that the methods obscured the possibility to see effects. 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 White M-Turk samples that many of these studies have used (Paolacci and Chandler, 2014). Future studies could consider whether preferences generalize across many groups and across variation in group disadvantage. Additionally, unequal FPR between Black and White people may yield differing preferences when the ML model is deployed in a predominately Black city as opposed to a predominately White one. The volume of people who could be needlessly jailed shifts radically when the bias affects 20 percent of the population versus 80 percent. As well, some diseases that algorithms may screen for are more prevalent in some groups than others, and it is possible that these kinds of demographic realities have implications for people's preferences. In this respect, future research on ML fairness can certainly benefit from the extensive psychological expertise on inter-group relations.

### 3.4 Generalizability

All of the criminal justice scenarios across these studies rely on the COMPAS dataset collected by Propbulica Angwin et al. (2016). These evaluate perceptions in the context of recidivism prediction, without acknowledging the challenges in that context beyond the algorithms Eckhouse et al. (2019).

However, several assumptions underlying Yu et al. (2020)'s design recommendations provoke questions about real-world applicability given their constructs of interest. First, Yu et al. (2020) base their ML learning interface on a "balanced dataset" - i.e., equal numbers of men and women, and equal numbers of blacks and whites (1500 White, 1500 African American, 800 male, and 800 female). This might over-represent, for example, white women compared to a real-world example – and cause bias in subject's evaluations of fairness trade-offs in ways that separate from the real-world constructs that Yu et al. (2020) seek to measure. Second, Yu et al. (2020) 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 they make about helping subjects make informed decisions through an interactive interface therefore might not adequately address differences in user-experience across different subsets of participants Rau et al. (2019). Statistical issues like these - for example, replication (the likelihood of reliably and validly measuring the same construct across repetitions of the same experiment) - are further discussed in Section 4.

## 4 Power, Replication

In addition to the conceptual challenges of measuring human attitudes, we must assure that the statistical inferences are well powered for the results to be reliable. In particular, psychology has been viewed as facing a replication crisis: recent efforts to replicate key findings have not yielded similar results (Maxwell, 2004). Replicating results remain a critical factor in comprehensively developing our understanding of perceptions of ML fairness. The inability to replicate these studies in psychology is often attributed to strong pressure to publish in high impact journal and outlets. High-profile scientific journals have earned a reputation for publishing primarily novel positive findings (Koole and Lakens, 2012). These publications often discourage the negative findings and replications, leaving incomplete and biased literature that is centered around positive findings (Wiggins and Chrisopherson, 2019). To the extent that concepts like AI fairness can benefit from generalizable, trustworthy research in psychology, irreproducible discoveries ultimately hinder the progression of the state of science. Therefore an important interdisciplinary insight from psychology - possibly even more important than positive findings - is how to detect irreproducible research, and how to dis-incentivize its publication.

Statistical power and publication bias are deeply correlated; given an effect and sample size, it is possible to calculate the likelihood of finding a positive result. As a result, low powered studies and studies with a p value of or near .05 paired with an unexpected result should be a red flag that this effect is unlikely to replicate. In Wiggins and Chrisopherson (2019), the authors argue that red flags such as these should be incorporated into publication criteria in order to filter-out irreproducible studies. If not, researchers are incentivized to perform practices that make their finding quickly publishable, even if these findings are not entirely representative of a true effect (Button et al., 2013). Statistically underpowered studies have reduced odds of detecting a true effect – but even if these underpowered studies discover a true effect, the magnitude of that these effects are potentially exaggerated are extremely high Button et al. (2013).

Another common research practice that results in easily-misinterpreted, low powered studies is uncorrected multiple comparisons - that is, simultaneously testing multiple hypothesis. Often, studies contain so many statistical tests that an acceptable number would be statically significant even if the power of any single test was inadequate Maxwell (2004). Critiquing these types of practices is critical in today's research environment, because there are high social and economic costs associated with assuming the truth of irreproducible effects that are nevertheless assumed to be true. (Wiggins and Chrisopherson, 2019).

Harrison et al. (2020) evaluates participant's preference for a human judge versus a model based on 12 comparisons through six pairwise comparisons. To correctly interpret these results, would require, for example, significance testing at a Bonferroni-corrected significance threshold – lowering the 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. Other studies included in this review, despite having smaller sample sizes, mitigate power issues by limiting the number of comparisons (Yu et al., 2020; Srivastava et al., 2019) or employing statistical techniques that are appropriate for selecting between a large number of models (see Srivastava et al. (2019)'s use of Akaike information criterion).

Furthermore, Harrison et al. (2020) themselves acknowledge that their 502 MTurk participants represent a convenience sample - meaning that they were included in the study simply because of ease of recruitment, rather than a more representative sample. Mturk participants in particular have been identified as not representative to the larger U.S. population Sheehan (2018). Mturk samples generally tend to have lower average incomes, higher education levels, lower average ages, and smaller percentages of most non-White groups, particularly Black and African Americans Levay et al. (2016). In the case of Harrison et al. (2020), this led to an under-representation of non-White participants – leading to reduced statistical power for these sub-groups, as well as statistical bias, which reduces the likelihood that Harrison et al. (2020)'s findings generalize or replicate. Practices like this could lead to interpreting spurious, non-replicating effects as significant (Leek and Peng, 2015). The replication crisis in psychology has lead to improved research practices, these should be followed when evaluating perceptions of machine learning.

## 5 Descriptive Work

It is also important to emphasize that an empirical approach to understanding perceptions of fairness is distinct from establishing a normatively correct standard of fairness. In other words, augmenting the ML fairness literature with insights from psychology is a descriptive project – one designed to better understand people's attitudes about algorithmic fairness, rather than one designed to decide which fairness notion is socially optimal. For example, Saha et al. (2020) note that their comprehension score negatively predicted particpants' likelihood of endorsing fairness rules. Greater comprehension was associated with decreased likelihood of agreeing with a fairness rule; however, describing this psychological tendency is different from endorsing it as a socially-optimal norm. Even though empirical work suggests that a fairness rule is more likely to be endorsed if it is less comprehensible, there is no normative implication (i.e., it does not imply we ought to make fairness rules less comprehensible so people find them fairer). Therefore, this research should lead to a deeper understanding of ML fairness perceptions, enabling a substantiated discussion. We are not proposing to use past and future results as direct recommendations for the practical implementation.

# 6 Conclusion

Notions of AI fairness are mutually-exclusive and therefore enforcing one type of fairness requires allowing other types of bias- this makes it impossible to create AI that is fair by all standards. To help narrow the scope of AI fairness, researchers have investigated which notions of fairness reflect real-world perceptions and understanding of fairness (Harrison et al., 2020; Saha et al., 2020; Yu et al., 2020; Srivastava et al., 2019). However, social constructs such as fairness preferences are heavily depend on context, framing, and social factors such as demographics (Tversky and Kahneman, 1973, 1981, 1986, 1991; Trope and Liberman, 2010; Braga et al., 2015; Wakslak and Trope, 2009). 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 (Maxwell, 2004; Wiggins and Chrisopherson, 2019; Button et al., 2013). 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 reinventing the wheel or making 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. For example, conclusions based on under-powered designs (Harrison et al., 2020) or demographic unrealities (Harrison et al., 2020; Yu et al., 2008) may not actually help make it any easier for researchers to successfully understand perceptions of fairness. Therefore, adopting an interdisciplinary approach - one that incorporates insights from psychology, statistics, philosophy, and other fields - is necessary in order to create fairer AI. These results of human experiments about perceptions of fair ML should appear in ML venues and be understood by ML researchers, but there might be another area of growth in our reviewing practices. Most ML researchers will not be aware of all of these factors, so we need a more interdisciplinary approach to both conducting and peer reviewing human subject experiments. We believe that only an interdisciplinary approach will appropriately address this highly relevant, societal challenge, and thus ensure that we will fulfill our obligation to future generations in implementing a responsible and fair use of algorithmic techniques.

# References

# References

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 252–260, 2020.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016. Publisher: HeinOnline.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning*. fairmlbook.org, 2019.

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

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

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

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

Timothy A Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

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

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

Nick Chater, Jian-Qiao Zhu, Jake Spicer, Joakim Sundh, Pablo León-Villagrá, and Adam Sanborn. Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science*, 29(5):506–512, 2020. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

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

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

Social Dominance. An intergroup theory of social hierarchy and oppression. 1999. Publisher: New York: Cambridge University Press.

Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, 46(2):185–209, 2019. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A Comparative Study of Fairness-enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 329–338, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287589. URL `http://doi.acm.org/10.1145/3287560.3287589`. event-place: Atlanta, GA, USA.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512, 2020.

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

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 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, pages 1–16, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300830. URL `https://doi.org/10.1145/3290605.3300830`. Place: Glasgow, Scotland Uk tex.numpages: 16.

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

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

Rex B Kline. *Principles and practice of structural equation modeling*. Guilford publications, 2015.

Sander L. Koole and Daniël Lakens. Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, 7(6):608–614, 2012.

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

Kevin E Levay, Jeremy Freese, and James N Druckman. The demographic and political composition of Mechanical Turk samples. *Sage Open*, 6(1):2158244016636433, 2016. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

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

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

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

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

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

Nadia Hanin Nazlan, Sarah Tanford, and Rhonda Montgomery. The effect of availability heuristics in online consumer reviews. *Journal of Consumer Behaviour*, 17(5):449–460, 2018. Publisher: Wiley Online Library.

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

Mathias Osmundsen and Michael Bang Petersen. Framing political risks: Individual differences and loss aversion in personal and political situations. *Political Psychology*, 41(1):53–70, 2020. Publisher: Wiley Online Library.

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

Zoe Oxley. Framing and political decision making: An overview. In *Oxford research encyclopedia of politics*. 2020.

Gabriele Paolacci and Jesse Chandler. Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Judea Pearl. The causal foundations of structural equation modeling. Technical report, CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE, 2012.

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

Richard Rau, Erika N Carlson, Mitja D Back, Maxwell Barranti, Jochen E Gebauer, Lauren J Human, Daniel Leising, and Steffen Nestler. 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. Publisher: American Psychological Association.

Tenko Raykov. Reliability if deleted, not 'alpha if deleted': Evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology*, 60(2):201–216, 2007. Publisher: Wiley Online Library.

Debjani Saha, Candice Schumann, Duncan C McElfresh, John P Dickerson, Michelle L Mazurek, and Michael Carl Tschantz. Measuring non-expert comprehension of machine learning fairness metrics. volume 119, Vienna, Austria, 2020. PMLR.

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

Kim Bartel Sheehan. Crowdsourcing research: data collection with amazon's mechanical turk. *Communication Monographs*, 85(1):140–156, 2018. Publisher: Taylor & Francis.

Paul Slovic, John Monahan, and Donald G MacGregor. 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):271–296, 2000. Publisher: Springer.

Megha Srivastava, Hoda Heidari, and Andreas Krause. 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*, pages 2459–2468, 2019.

Henri Tajfel, Michael G Billig, Robert P Bundy, and Claude Flament. Social categorization and intergroup behaviour. *European journal of social psychology*, 1(2):149–178, 1971. Publisher: Wiley Online Library.

Yaacov Trope and Nira Liberman. Construal-level theory of psychological distance. *Psychological review*, 117 (2):440, 2010. Publisher: American Psychological Association.

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

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

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

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

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 2(6), 2017. doi: 10.1126/scirobotics.aan6080. URL `https://robotics.sciencemag.org/content/2/6/eaan6080`. Publisher: Science Robotics tex.elocation-id: eaan6080 tex.eprint: https://robotics.sciencemag.org/content/2/6/eaan6080.full.pdf.

Cheryl Wakslak and Yaacov Trope. The effect of construal level on subjective probability estimates. *Psychological Science*, 20(1):52–58, 2009. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

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

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

Lei Yu, Chris Ding, Steven Loscalzo, Lei Yu, Chris Ding, Chris Ding, Steven Loscalzo, and Steven Loscalzo. Stable feature selection via dense feature groups. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 803, New York, New York, USA, 2008. ACM Press. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401986. URL `http://portal.acm.org/citation.cfm?doid=1401890.1401986http://dl.acm.org/citation.cfm?doid=1401890.1401986`.