Breaking Taboos in Fair Machine Learning: An Experimental Study

JULIAN NYARKO, Stanford University, USA
SHARAD GOEL, Harvard University, USA
ROSEANNA SOMMERS, University of Michigan, USA

Many scholars, engineers, and policymakers believe that algorithmic fairness requires disregarding information about certain characteristics of individuals, such as their race or gender. Often, the mandate to "blind" algorithms in this way is conveyed as an unconditional ethical imperative—a minimal requirement of fair treatment—and any contrary practice is assumed to be morally and politically untenable. However, in some circumstances, prohibiting algorithms from considering information about race or gender can in fact lead to worse outcomes for racial minorities and women, complicating the rationale for blinding. In this paper, we conduct a series of randomized studies to investigate attitudes toward blinding algorithms, both among the general public as well as among computer scientists and professional lawyers. We find, first, that people are generally averse to the use of race and gender in algorithmic determinations of "pretrial risk"—the risk that criminal defendants pose to the public if released while awaiting trial. We find, however, that this preference for blinding shifts in response to a relatively mild intervention. In particular, we show that support for the use of race and gender in algorithmic decision-making increases substantially after respondents read a short passage about the possibility that blinding could lead to higher detention rates for Black and female defendants, respectively. Similar effect sizes are observed among the general public, computer scientists, and professional lawyers. These findings suggest that, while many respondents attest that they prefer blind algorithms, their preference is not based on an absolute principle. Rather, blinding is perceived as a way to ensure better outcomes for members of marginalized groups. Accordingly, in circumstances where blinding serves to disadvantage marginalized groups, respondents no longer view the exclusion of protected characteristics as a moral imperative, and the use of such information may become politically viable.

ACM Reference Format:
Julian Nyarko, Sharad Goel, and Roseanna Sommers. 2021. Breaking Taboos in Fair Machine Learning: An Experimental Study. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, 2021, –, NY, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3465416.3483291

1 INTRODUCTION

For decades, anti-discrimination law has been marked by a debate over two distinct interpretations of the Equal Protection Clause of the U.S. Constitution [37]. Roughly speaking, one interpretation focuses on process and the other on outcomes [1]. The process-based view interprets the mandate not to discriminate as a colorblind "anticlassification" principle, concerned with treating citizens as individuals and not as members of racial or other groups. Under this anticlassificationist view of discrimination, a government policy or decision that classifies individuals based on protected characteristics, such as race or gender, is discriminatory.

In contrast, the outcome-based view focuses on how members of different groups ultimately fare under a policy. Under the "antisubordination" principle, policies are discriminatory if they exacerbate existing inequalities between groups. In particular, a policy that considers protected characteristics may be justified if it reduces social stratification. Further, according to the antisubordinationist view, even a policy that applies equally to members of all race groups can be deemed discriminatory if it disproportionately burdens members of a protected group.

In practice, the process- and outcome-based views often align. For example, in *Brown v. Board*, the Supreme Court cited both rationales in striking down segregation in education.[1] At times, however, the two principles come into conflict. Consider, for instance, a public university that automatically awards 20 out of 100 points to applicants from an underrepresented minority group.[2] Under a process-centered, anticlassificationist approach, this policy is discriminatory because it requires classifying individuals based on their race, which is prohibited regardless of intent or outcomes. By contrast, under an antisubordinationist approach, the affirmative action policy is not necessarily discriminatory, because its purpose and effect may be to reduce group-based social inequality.[3]

While scholars have debated these two perspectives in a variety of domains affected by anti-discrimination law—including education and employment—equivalent discussions appear to be largely absent in current discourse about algorithmic decision making. Many assume that the inclusion of protected characteristics in predictive models would constitute a discriminatory practice. As Kleinberg et al. [26] write, concerns about algorithmic fairness "have led to a large literature that tries to 'blind' the algorithm to race to avoid exacerbating existing unfairnesses in society." When it comes to algorithmic decision making, the battle is not over whether the removal of data pertaining to protected characteristics is necessary or desirable, but whether it is sufficient. Consistent with this view, the last several years have seen an influx of studies that examine not only the removal of race, gender, and other protected characteristics, but also of the many proxies, like one's home zip code, that may correlate with these characteristics in a substantial way [3, 5, 29, 42]. Even parts of the literature that view the incorporation of protected characteristics as a requirement for fair decision making often treat their use as problematic. For instance, algorithms following a "Disparate Learning Process" [30] suggest that the inclusion of protected characteristics is permissible during training, but not at prediction time. Relatedly, a burgeoning literature on causal notions of fairness [23, 29] allow for protected attributes to be used in machine learning classifiers, but restrict their influence to causal paths deemed *a priori* "legitimate." In this way, rather than examining the impacts of protected attributes on algorithmic decisions, a process-based view of discrimination is taken for granted. Dissenting or cautionary voices, although they exist, are not numerous [6, 20, 21, 31, 41].

Yet it remains unclear what rationale underlies the resistance to using protected characteristics in algorithmic decision making. In many cases, commentators have not been explicit in their reasoning. Where reasons are given, the justifications often reflect one of two modes of ethical reasoning: Under a *deontological* account, decision rules can be morally permissible or impermissible, irrespective of their consequences. A deontological justification for blind algorithms holds that it would be "fundamentally unethical or immoral" to condition the allocation of costs and benefits on an individual's group membership [12, 17, 39, 40]. Accordingly, under this account, there is independent value to a blind algorithm, regardless of how blinding affects the algorithmic predictions, because the use of protected characteristics to guide decisions is itself unacceptable.

In contrast, a *consequentialist* view judges the morality of a decision rule based solely on its outcomes. The ethical permissibility of blinding is derived not from the conviction that blinding itself is a virtue, but from the assumption

[1] Brown v. Bd. of Educ., 347 U.S. 483, 494 (1954).

[2] The facts are taken from *Gratz v. Bollinger*, 539 U.S. 244 (2003). Earlier affirmative action practices heard by the Supreme Court were upheld based on a diversity rationale, which emphasizes the benefits of creating a diverse student body. The diversity framework has since become central, and the antisubordination rationale marginal, in modern affirmative action jurisprudence.

[3] Outside of affirmative action, courts have often embraced an anticlassificationist rationale interpretation of the Equal Protection Clause. "At the heart of this [process-based, anticlassificationist] interpretation of the Equal Protection Clause lies the principle that the government must treat citizens as individuals, and not as members of racial, ethnic, or religious groups." Missouri v. Jenkins, 515 U.S. 70, 120–21 (1995) (Thomas, J., concurring).

that blinding is a helpful means to achieving more equitable outcomes. In this way, a consequentialist justification for blinding is based on an underlying assumption that the inclusion of protected characteristics will likely disadvantage members of already-disadvantaged groups, such as racial minorities. Articulating this view in the context of pretrial risk assessments, former U.S. Attorney General Eric Holder suggests that "basing sentencing decisions on static factors and immutable characteristics . . . may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."[4]

While the two viewpoints often lead to the conclusion that algorithms should be blind, recent findings complicate this picture. For example, requiring risk assessment tools to ignore gender has been shown to result in an overestimation of the risk that female defendants will recidivate [6, 38]. If risk assessment tools were permitted to consider gender, not only would the resulting risk scores align better with actual recidivism rates, but they would likely yield lower rates of pre-trial detention for female defendants. Thus, a "gender-specific" tool, unlike a gender-blind tool, may result in better outcomes for members of the marginalized group, even if it violates the deontological rule against considering gender.

In the same vein, blinding risk assessment tools to race may overstate the recidivism risk for Black defendants under specific circumstances. As Mayson [31] suggests, disparate policing practices may render prior arrest and conviction records of Black defendants relatively uninformative about future recidivism risk. As a result, race-blind algorithmic risk assessments can overstate recidivism risk for Black individuals relative to white individuals. A similar phenomenon could, in theory, lead to higher auto insurance rates for Black and Hispanic drivers. Common insurance-pricing algorithms rely in part on the number of speeding tickets a driver has received, and so discriminatory traffic enforcement practices could result in inflated estimates of collision risk for minorities [32].

As these examples illustrate, avoiding the use of protected characteristics through the use of blind algorithms can, in some instances, lead to worse outcomes for members of a historically disadvantaged group. Because the two leading theories point to different prescriptions, these studies highlight the importance of identifying the underlying justification for the preference of blind algorithms. Do people believe that blinding should be preserved even if it imposes costs on the members of disadvantaged groups? Or do they think it is permissible—or even necessary—to consider protected characteristics when doing so avoids imposing additional burdens on marginalized groups?

In this experimental study, we assess the normative basis for the widespread preference for blind algorithms. Surveying a representative sample of the U.S. population, as well as computer scientists and lawyers, we first confirm that, in the absence of additional information, respondents strongly prefer algorithmic tools that avoid the use of race and gender in the context of pretrial risk assessments. In explaining their preference, respondents frequently employ moralizing language, suggesting that their aversion to race- and gender-specific tools stems from process-based concerns about treating individuals differently based on group membership. However, we then show that these anticlassificationist preferences are fragile: a minimal intervention in which respondents read two paragraphs about the abstract possibility that blinding can impose additional burdens on marginalized groups drastically increases the approval for the inclusion of protected characteristics. This intervention yields similar effect sizes among the general public, computer scientists, and lawyers. When prompting participants to justify their preference in an open-ended question, we find that respondents generally do not reference the desirability of the consequences as a basis for their views. However, when forced to choose between a deontological justification (focused on the harm of classifying) and a consequentialist justification (focused on the harm of perpetuating group subordination), a sizeable minority of

[4] Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference, available at https://www.justice.gov/opa/speech/attorney-general-eric-holder- speaks-national-association-criminal-defense-lawyers-57th.

respondents selected consequentialist reasoning as being more reflective of their own motivation. It is this latter group of "closet consequentialists" that drives much of the effect that we observe.

Our results suggest that the preference for blind algorithms, although often framed in ethical terms that suggest an unconditional position, is not grounded in a universal belief that the use of protected characteristics is taboo or morally prohibited. Instead, it appears that stated preferences are driven in part by a consideration of outcomes. Many evidently treat blinding as a useful heuristic that is more likely to avoid burdening marginalized groups, and their beliefs thus depend on the specific context in which algorithmic predictions are made. Our findings call into question the prevailing sentiment that algorithms must eschew race and gender information. They highlight the need for a serious and detailed discussion on the welfare costs and ethical virtues of blind decision making. A balancing of these interests may argue in favor of the inclusion of protected characteristics in specific situations, depending on the context.

2 BACKGROUND & RELATED WORK

Protected class features in algorithmic decision making are often treated as if their exclusion was legally required. However, in reality, it is far from settled whether the inclusion of such features could give rise to successful legal challenges. While the Supreme Court has never ruled on the legality of the use of protected characteristics in algorithmic decision making, we can turn to general principles of anti-discrimination law to observe that current doctrine recognizes some important exceptions to the general rule that the government may not draw distinctions between individuals based on suspect classifications. In Appendix A, we survey some of the relevant case law and legal scholarship.

Our study considers and experimentally manipulates respondents' views on algorithmic decision making. In doing so, it contributes not only to the legal and ethical discourse surrounding blind decision making, but also to an emerging literature on attitudes toward algorithms. Previous work has demonstrated an aversion among the general public towards algorithmic decision making that appears to be fueled by a tendency to assign more weight to errors made by algorithms than by human forecasters [8]. The aversion is difficult to overcome and seems to persist, regardless of whether the decision making process is transparent or not [9, 33].

Among the studies most relevant to ours, Grgic-Hlaca et al. [15] surveyed 100 Amazon Mechanical Turk workers for their perception of the fairness of the use of different features in the context of pre-trial risk assessments. They found that 21% of participants viewed the inclusion of race as fair. However, when subsequently asked if the inclusion of race is fair if it makes the prediction more accurate, 42% of respondents responded positively. For gender, these numbers were 26% and 55% respectively. Similarly, Harrison et al. [19] asked respondents to choose between: (i) an algorithm that uses race as a feature but leads to equality in either accuracy, false positive rate or outcomes; and (ii) an algorithm that does not use race as a feature but leads to differences in accuracy, false positive rates or outcomes. Outcomes were defined as the likelihood of a defendant being granted bail. The authors find a general preference for blind algorithms, even when they lead to disparities in accuracy or outcomes. However, when asked to choose between a blind algorithm and one that equalizes the false positive rate, respondents were split. Although neither of the two studies directly tested the interaction of process- and outcome-based justifications for the exclusion of protected features, they do offer preliminary support for the hypothesis that outcomes and effects implicate attitudes toward algorithmic decision-making procedures.

3 STUDY DESIGN

We designed and conducted a pre-registered 3-part experiment in which participants were randomly assigned to receive information about the outcomes of an algorithmic decision-making process in the context of pretrial risk assessments.

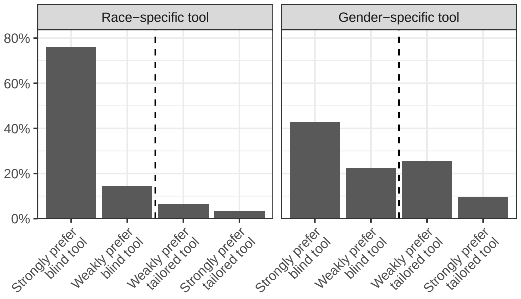

Fig. 1. *The distribution of preferences for blind risk assessment tools, among respondents in the "no context" condition. The figure contains two histograms side-by-side, one for the race-specific tool and one for the gender-specific tool. Each histogram has four bars indicating the share of survey respondents who answered that they [strongly/weakly] prefer blind [tailored] tool. A race-blind tool is almost universally preferred. Responses for the gender-specific tool are less homogenous, although most respondents prefer the gender-blind tool either strongly or weakly.*

Participants were first introduced to the proposal of using algorithmic risk assessments in decisions about pretrial release. Half of participants were asked to consider whether the risk assessment tools should take into account the race of the defendant; half were asked to consider the same question as it pertained to the defendant's gender. Subsequently, participants were randomly assigned to one of three information conditions, described below. Finally, participants were asked to report whether they preferred the blind or non-blind tool (on a five-point scale), as well as their reasoning.

In the "no context" (control) condition, participants were given no further information before reporting their preference regarding the inclusion of race (or gender) as a feature in algorithmic risk assessments. In the general context condition, participants were told that some experts favor allowing risk assessment tools to take account of race [gender] because they fear that blinding can sometimes lead to adverse outcomes for protected groups. The purpose of this condition was to mimic the type of general information respondents may receive if they follow political discourse or public debates in the media that invokes expert opinions on the design of risk assessment tools. Finally, in the specific context condition, participants were assigned to one of five sub-conditions. In each, they were presented with the results from a "hypothetical study" and given illustrations showing how many out of 100 Black [female] defendants and how many out of every 100 white [male] defendants would be released and detained, under blind and tailored algorithms (Figure 5). In all five sub-conditions, the blind algorithm was projected to release 50% of individuals in the protected and unprotected groups. The projections for the tailored algorithm, however, varied systematically in each sub-condition, portraying the tailored algorithm as [weakly/strongly] advantaging, [weakly/strongly] disadvantaging, or not affecting, protected groups as compared to the blind algorithm. We provide further details of our study design, including a discussion of its limitations, in Appendix B.

4 RESULTS

We begin by examining the baseline preference for group-specific (or "tailored") algorithms in our representative sample of U.S adults. Figure 1 depicts the approval rate for race- and gender-specific algorithms for respondents in the "no context" condition (i.e., those who received no additional information). As can be seen, the vast majority of respondents either strongly or weakly prefer the race-blind over the race-specific tool (90% vs 10%), with 76% of respondents saying they "strongly prefer" the blind tool. These findings establish a striking aversion against the use of race as a predictive

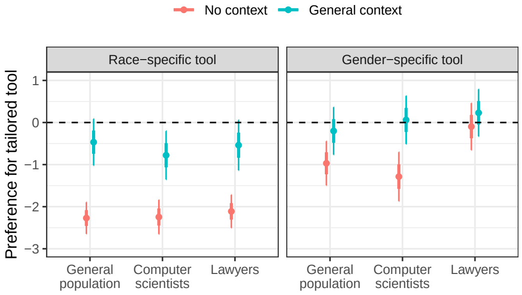

Fig. 2. *Comparison of preferences between the general population, computer scientists, and lawyers. Two panels side-by-side, one for the race-specific tool and one for the gender-specific tool. Each panel has three columns. The left column depicts survey responses for a representative sample of the population. The middle column depicts responses of surveyed computer scientists. The right column depicts results for lawyer respondents. The y-axis ranges from -3 to 1. Two dots in each column depict point estimates for the average response in the "no context" and in the "general context" condition. Vertical lines depict 95% confidence intervals. The plot shows that the preferences across all three groups are similar: Point estimates are negative and significant, suggesting that opposition to the tailored tool is pronounced in the "no context" condition. In the "general context" condition, opposition is small and usually not statistically significant or borderline significant. One exception to this pattern are the lawyer responses with respect to the gender-specific tool, where the point estimate for the "no context" condition indicates no preference and for the "general context" condition it indicates a mild preference for the tailored tool. However, both estimates are not statistically significant.*

feature that is consistent with prior literature [15]. When asked to state their reasons, 46 out of 58 respondents highlight that the use of race would be highly immoral and unethical. For instance, one respondent suggests that "[i]n the U.S., practices, where race is used as a factor to determine whether someone is deemed dangerous or not, is morally wrong, and in some cases, illegal to do. Racial profiling isn't the best way to determine one's livelihood and our country is better than that." Another indicates that "[t]here is already way too much racial bias against minorities, especially blacks and First Nations. Deliberately making a tool that's racially based is biased and wrong."

Respondents similarly have a baseline preference for gender-blind tools, although that preference is somewhat muted (65% vs 35%). Those that do prefer a gender-specific tool often express their belief that women have lower rates of recidivism and should thus benefit from the inclusion of gender as a feature. Indeed, 14 out of 22 respondents who prefer the tailored tool specifically reference a lower propensity for female defendants to commit crimes. For instance, one respondent states: "While I'm well aware that men and women can be equally horrible I think that testosterone should be taken into account and that the reality is that men tend to be much more violent than women." And another says: "I'm under the general understanding that men are statistically more violent than women, therefore a gender bias may sometimes be necessary in risk assessment."

Next, we translate responses into numerical values of equal distance ranging from -3 for "strongly prefer blind tool" to 3 for "strongly prefer tailored tool." In the left column of Figure 2, we then contrast the average rates of approval for the tailored tools between those in the "no context" condition and those in the "general context," where respondents receive information about the abstract possibility that the inclusion of protected features could increase release rates among Black or female defendants. As can be seen, providing abstract information in this way significantly increases the approval for both race- and gender-specific algorithms . The increase is statistically significant by conventional measures (t-test: $p < 0.001$ for race and $p = 0.048$) for gender. Indeed, under the general context condition, we can no longer reject the hypothesis that respondents have, on average, no preference for the blind tools.

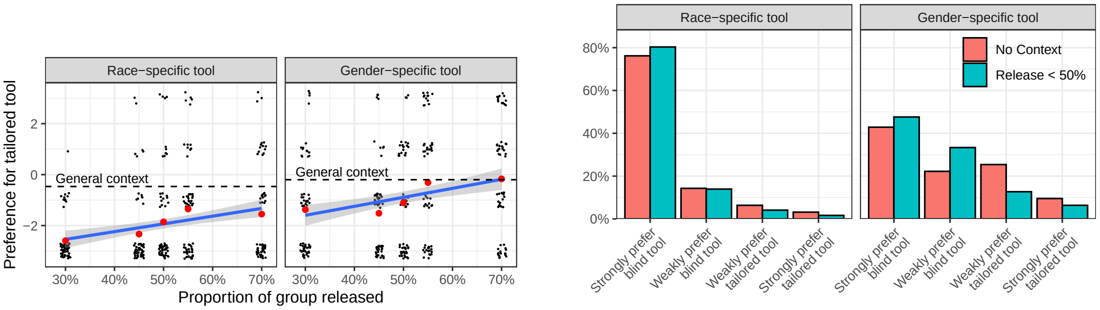

Fig. 3. *Preferences in the "specific context" condition. Left: Two panels side-by-side, one for the race-specific tool and one for the gender-specific tool. The x-axis ranges from 30% to 70% and represents the share Black [female] defendants released under the tailored tool. The y-axis ranges from -3 to 3 and represents support for the blind [tailored] tool numerically. Small, jittered dots represent individual responses. Large, red dots represent average responses. A blue line depicts the results of a linear regression of support on the share of Black [female] defendants released, with a grey-shaded area around the line representing 95% confidence intervals on the regression coefficient. A dashed, horizontal line depicts average support in the "general context" condition for reference. Both panels show that support for the tailored tool increases with the share of Black [female] defendants released under the tailored tool. However, for the race-specific tool, support remains below the average support in the "general context condition." Right: Two histograms side-by-side, one for the race-specific tool and one for the gender-specific tool. Each histogram has four sets of bars, one red and one green, where each set represents a survey response. The y-axis depicts the share of respondents for each response. Red bars represent the proportion of survey responses in the "no context" condition and green the proportion among all responses in the "specific context" condition where the release rate of Black [female] defendants is less than 50%. For the race-specific tool, green and red bars are very similar. A Kolmogorov-Smirnov permutation test for differences in distributions yields a p-value of 0.65. This supports the view that, without further information, respondents generally believe race-specific algorithms result in lower release rates for Black defendants. For the gender-specific tool, it appears that support for the tailored tool may be somewhat more pronounced in the "no context" condition. A Kolmogorov-Smirnov permutation test for yields a p-value of 0.05, suggesting a possible difference between distributions.*

When examining the open-text responses of those who prefer the race-specific algorithm, it appears noteworthy that 20 out of 26 respondents justify their decision by invoking the higher quality of the predictions that would result under the use of race. For instance, one respondent states that "[the tailored tool] will compensate for racial inequality and do a better job at predicting those who pose a risk." Another says: "I think includ[ing] the race specific criteria will help the algorithm make more accurate predictions." Save for two exceptions, no respondent clearly states that their preference for the race-specific tool is due to a higher release rate for Black defendants. Notably, respondents shied away from providing overtly consequentialist justifications for their preferences even though the "general context" provided a clearly consequentialist argument. This finding is consistent with other research showing that consequentialist decision rules are viewed as immoral or less legitimate than decision rules grounded in a deontological conception of ethics [10, 11, 35].

Also in Figure 2, we contrast the responses of the general population to that of computer scientists and lawyers. Interestingly, the treatment effect is both substantively and statistically similar across all groups. To us, this finding is surprising, as lawyers have been among the most vocal advocates for a process-based account for the ethical necessity of blinding in algorithmic decision making. Our results suggest that these tendencies may exist only in their stated preferences, but do not survive an experimental manipulation.

We have shown that manipulating respondent's beliefs about the possible outcome under a race- or gender-specific algorithm alters their approval. We now consider the effect size of this manipulation across varying levels of intensity of the treatment. In Figure 3 (left panel), we plot the average approval rate across each of our "specific context" conditions.

Recall that respondents had to choose between: (i) a blind algorithm releasing 50% of minority and 50% of non-minority defendants; and (ii) and a tailored algorithm releasing λ% of minority and 100-λ% of non-minority defendants. The x-axis depicts λ, the proportion of minority defendants released under each treatment condition. The y-axis depicts the rate of approval for the tailored tool. The blue line depicts a linear regression of the approval on the numeric proportion of released minority defendants.

Respondents' approval of the tailored tool increases with the share of minority defendants that are released under it. Interestingly, under no "specific context" condition is the approval for the race-specific tool as large as under the "general context" condition. We do not know the reason for this result, although we conjecture that the examples we included in our general context condition may have been more relatable than our specific context condition. This view seems to be supported by some of the open-text responses under the general context condition, in which respondents pick up on our specific examples to justify their decision (11 out of 26).

Having found that respondents' aversion to the inclusion of protected characteristics can be significantly reduced if they learn about the potential for adverse consequences, we now consider further evidence that their motivations are driven by a consideration of outcomes. First, we hypothesize that respondents' baseline aversion to race-specific algorithms are grounded in the belief that the inclusion of race would lead to higher detention rates among Black defendants. If true, it should be the case that respondents do not update their beliefs and/or reported aversion in response to the information that the tailored algorithm will detain more Black defendants. In Figure 3 (right panel), we contrast responses of participants in the "no context" condition to those in the "specific context" condition in which release rates of the race-specific algorithm are lower than in the blind algorithm. As can be seen, responses are virtually identical,[5] suggesting that respondents' baseline belief is that the inclusion of race increases incarceration rates for Black defendants. For gender-specific algorithms, we see some differences between the two conditions,[6] which is consistent with the view of many respondents—expressed in the open-text responses when ascertaining baseline preferences—that the inclusion of gender would lead to lower detention rates for female defendants.

In a last step, we contrast responses by those who identify as focusing on process with those who state that outcome-based reasons are more consistent with their own motivation. As explained in Appendix B, we identify respondents' preferred philosophy by examining whether they prefer a consequentialist or deontologist rationale for blind algorithms. As can be seen in Figure 4, responses under the "no context" condition are roughly similar between the two groups. However, the treatment effect under the "general context" condition is greater for self-identified consequentialists. Interestingly, even self-identified deontologists are more receptive to the tailored tool if it implies higher release rates for Black defendants.

5 DISCUSSION

Our findings suggest that the aversion to using protected characteristics in algorithmic predictions of pretrial risk can be drastically decreased by making salient to respondents that blinding could have adverse outcomes for members of protected groups. This result is observed across a representative sample of the general population in the U.S., as well as among computer scientists and professional lawyers. The effect size is largest for our general context condition, which seeks to mirror the abstract way in which information is portrayed in the media and public discourse. In addition, we show that much of the aversion to tailored algorithms is consistent with the belief that tailoring would increase the burden on protected groups. In sum, our results present strong evidence in support of the hypothesis that a large share

[5] The p-value from a Kolmogorov-Smirnov permutation test with 2,000 permutations is 0.65.
[6] $p = 0.05$.

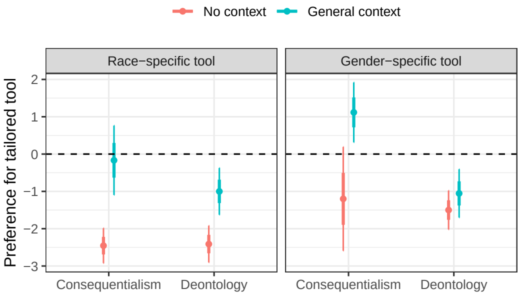

Fig. 4. *Support for group-specific tools among self-identified consequentialists and deontologists. Two panels side-by-side, one for the race-specific tool and one for the gender-specific tool. Each panel has two columns, one indicating a stated preference for consequentialist reasoning one for deontological reasoning. The y-axis ranges from -3 to 2. Two dots in each column depict point estimates for the average response in the "no context" (red) and the "general context" (green) condition. Vertical lines depict 95% confidence intervals. Panels include a horizontal, dashed line at 0. Both for the race- and gender-specific tool, point estimates in the "no context" condition are similar for deontologists and consequentialists. In addition, in both panels, support for the tailored tool is greater in the "general context" condition. The increase is more pronounced for self-identified consequentialists than self-identified deontologists. For deontologists judging the gender-specific tool, the increase is small and not significant.*

of the population views blinding not as a moral imperative, but as an instrument to achieve equitable outcomes. Indeed, under the general context condition, the average approval for the use of race- and gender-specific algorithms is close to 0 (on a scale ranging from -3 to 3), suggesting respondents are largely ambivalent in their preference for blind versus group-specific algorithms. This finding suggests that the use of race- or gender-specific algorithms in pretrial decision making may be politically viable if it can be ensured that they lead to improvements for members of the protected group.

Perhaps the most obvious policy suggestion flowing from our results is an auditing solution. Policy makers could begin by comparing two models, one including the protected features and one excluding them. If the tailored algorithm leads to greater benefits or lower costs for members of the protected group, it would be implemented. Otherwise, the blind algorithm would be used. We note, however, that the legality of auditing and selecting algorithms in this way is unclear. Although the Supreme Court has never ruled on the exact issue, the most relevant opinion is found in *Ricci v. DeStefano* from 2009.[7] In this case, the New Haven Fire Department issued a standardized test to determine promotions of their firefighters. After the firefighters had taken the test, the department found that none of the Black candidates achieved test scores that were high enough for a promotion. Subsequently, officials decided to invalidate the test. The firefighters who would have been promoted under the test sued, alleging racial discrimination because the decision to invalidate the test scores was done under an explicit consideration of the applicants' race. New Haven argued that they needed to invalidate the test because it led to disparities in outcomes across racial groups, which would be unlawful under the disparate impact doctrine of Title VII of the Civil Rights Act of 1964. The Supreme Court sided with the plaintiffs, holding that New Haven's claim that their test would have positioned them in conflict with Title VII lacked the required "strong basis in evidence."

Experts debate whether the decision in *Ricci* implies that the practice of auditing and revising algorithms with an eye for their disparate consequences is unlawful [2, 24, 27]. It is difficult to predict how the Supreme Court would rule. Yet we point out a few aspects that distinguish the New Haven test scores from the described policy recommendation. First,

[7] Ricci v. DeStefano, 557 U.S. 557 (2009).

Ricci was decided under Title VII, which is a statutory anti-discrimination law that regulates employment practices. The Supreme Court has developed an extensive body of case law under Title VII. The resulting doctrine does not fully mirror doctrinal developments under the Equal Protection Clause. For instance, in addition to race-conscious hiring practices, Title VII expressly recognizes as discriminatory policies that are facially neutral, but impose an unequal burden on minorities.[8] There are thus limited inferences one may draw for discrimination under the Equal Protection Clause, which would govern most governmental policies outside of the employment context, including the use of risk assessment tools. In fact, the Supreme Court in *Ricci* expressly refused to assess how the case would be decided under the Equal Protection Clause. Second, and more importantly, the majority rests its opinion in *Ricci* to a large extent on the fact that the New Haven Fire Department, in administering the test, created an expectation in the firefighters not to be judged on the basis of their race. It is this expectation that they then frustrated when they discarded the scores.[9] Policy makers could try to avoid the creation of similar expectations by clearly communicating that the reduction of inequalities in outcomes is one of their stated goals in designing the algorithm. Third, and closely related, in *Ricci*, the New Haven Fire Department chose to discard the test scores *after* the test had been taken. In that way, their practice barred the promotion of firefighters after they had already been entitled to a promotion. An auditing process as we have described it above could easily avoid this outcome. For instance, in assessing the likely effects of including race for algorithmic risk predictions, the models could be fit on historical data. In this way, the decision whether or not to use the race-specific algorithm would be made without using information on the individual defendants who are affected by the decision. All these differences could provide grounds for distinguishing *Ricci* from algorithmic auditing practices in a legally relevant way. However, absent a clear decision, relevant practices would remain in a constitutional grey area.

Beyond the need for more clarity on the legal framework, our results point to the importance of a normative discussion surrounding the ethics and virtues of blind decision making. Contrary to a key assumption underlying the current debate, at least in the context of pretrial risk assessments, there appears to be no widely shared assumption that the inclusion of protected features is immoral or taboo. Indeed, it appears plausible that a more intensive treatment (e.g., through additional examples or more elaborate explanations) could have led to an even greater shift in attitudes towards tailored algorithms. Given a lack of consensus about the motivation for blinding and an increased tendency to administer important societal goods through algorithmic decision making, it may be appropriate to reevaluate the presumed imperative to blind algorithms.

In our general-context condition, we focus on the potential benefits of race- and gender-specific algorithms, because we sought to tease apart the various reasons why people might support blinding. However, there are of course many important situations in which blinding protects historically marginalized groups from discrimination. For example, existing income and wealth disparities can lead to higher loan default rates for Black and Hispanic borrowers—even after adjusting for credit score—and so algorithms that are blind to race (and its close correlates) can offer at least some protection against further entrenching existing disparities[13, 18]. And indeed, deviation from existing principles bear additional dangers. For instance, if the dogma of blind algorithms is successfully challenged, it could create potential for abusive and intentionally discriminatory practices which can be difficult to detect *ex post* [6]. To mitigate the potential for such misuse, blind decision making could be prescribed as a default rule, while setting a high evidentiary standard for policy makers seeking to deviate from the default. Looking forward, we hope our results help researchers and policymakers chart an equitable path for designing and deploying algorithmic tools.

[8] Griggs v. Duke Power, 401 U.S. 424, 426 (1971)

[9] "The injury arises in part from the high, and justified, expectations of the candidates who had participated in the testing process on the terms the City had established for the promotional process. Many of the candidates had studied for months, at considerable personal and financial expense, and thus the injury caused by the City's reliance on raw racial statistics at the end of the process was all the more severe." Ricci v. DeStefano, 557 U.S. 557, 593 (2009).

ACKNOWLEDGMENTS

We thank Robert Bartlett, Ryan Copus, Mark Kelman, and Shaun Ossei-Owusu for helpful comments and suggestions. We are also grateful to Giuliana Carozza Cipollone for invaluable research assistance.

REFERENCES

[1] Jack M. Balkin and Reva B. Siegel. 2002. The American Civil Rights Tradition: Anticlassification or Antisubordination The Origins and Fate of Antisubordination Theory. *Issues in Legal Scholarship* 2, 1 (2002), [i]–17.

[2] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact Essay. *California Law Review* 104, 3 (2016), 671–732.

[3] Robert Bartlett, Adair Morse, Nancy Wallace, and Richard Stanton. 2022. Algorithmic Discrimination and Input Accountability under the Civil Rights Acts. *Berkeley Technology Law Journal* (2022). forthcoming.

[4] Jason R. Bent. 2020. Is Algorithmic Affirmative Action Legal? *Georgetown Law Journal* 108, 4 (2020), 803–853.

[5] Alex Chohlas-Wood, Joe Nudell, Keniel Yao, Zhiyuan Lin, Julian Nyarko, and Sharad Goel. 2021. Blind justice: Algorithmically masking race in charging decisions. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 35–45.

[6] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[7] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), 797–806.

[8] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General* 144, 1 (Feb. 2015), 114–126.

[9] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2016. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64, 3 (Nov. 2016), 1155–1170. Publisher: INFORMS.

[10] Jim A.C. Everett, Nadira S. Faber, Julian Savulescu, and Molly J. Crockett. 2018. The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology* 79 (Nov. 2018), 200–216.

[11] Jim A. C. Everett, David A. Pizarro, and M. J. Crockett. 2016. Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology. General* 145, 6 (June 2016), 772–787.

[12] Richard S. Frase. 2013. Recurring Policy Issues of Guidelines (and Non-Guidelines) Sentencing: Risk Assessments, Criminal History Enhancements, and the Enforcement of Release Conditions Guest Editor's Observations. *Federal Sentencing Reporter* 26, 3 (2013), 145–157.

[13] Talia B Gillis and Jann L Spiess. 2019. Big data and discrimination. *The University of Chicago Law Review* 86, 2 (2019), 459–488.

[14] Sharad Goel, Ravi Shroff, Jennifer L Skeem, and Christopher Slobogin. 2020. The accuracy, equity, and jurisprudence of criminal risk assessment. In *Research Handbook on Big Data Law*.

[15] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems*. 11.

[16] Ian Haney-Lopez. 2012. Intentional Blindness. *New York University Law Review* 87, 6 (2012), 1779–1877.

[17] Bernard E. Harcourt. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. University of Chicago Press.

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[19] 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 (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 392–402.

[20] Daniel E Ho and Alice Xiang. 2020. Affirmative Algorithms: The Legal Grounds for Fairness as Awareness. *University of Chicago Law Review Online* (2020), 134–154.

[21] Aziz Z. Huq. 2018. Racial Equity in Algorithmic Criminal Justice. *Duke Law Journal* 68, 6 (2018), 1043–1134.

[22] Aziz Z. Huq. 2020. Constitutional Rights in the Machine Learning State. *Cornell law Review* 105, 7 (2020), 1875–1954.

[23] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744* (2017).

[24] Pauline Kim. 2017. Auditing Algorithms for Discrimination. *University of Pennsylvania Law Review Online* 166, 1 (Jan. 2017), 189–203.

[25] Pauline T Kim. 2016. Data-driven discrimination at work. *William & Mary Law Review* 58 (2016), 857–934.

[26] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. 22–27.

[27] Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. Accountable Algorithms. *University of Pennsylvania Law Review* 165, 3 (2016), 633–706.

[28] James A. Kushner. 2019. *Government Discrimination: Equal Protection Law and Litigation* (2019-2020 ed.). Clark Boardman Callaghan, Deerfield, IL.

[29] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in neural information processing systems*. 4066–4076.

[30] Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 8136–8146.

[31] Sandra G. Mayson. 2018. Bias in, Bias out. *Yale Law Journal* 128, 8 (2018), 2218–2301.

[32] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour* (2020), 1–10.

[33] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]* (Nov. 2019). arXiv: 1802.07810.

[34] Russell K. Robinson. 2016. Unequal Protection. *Stanford Law Review* 68, 1 (2016), 151–234.

[35] Sarah C. Rom, Alexa Weiss, and Paul Conway. 2017. Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology* 69 (March 2017), 44–58.

[36] Reva B. Siegel. 2003. Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over Brown Symposium: Brown at Fifty. *Harvard Law Review* 117, 5 (2003), 1470–1547.

[37] Reva B Siegel. 2010. From colorblindness to antibalkanization: An emerging ground of decision in race equality cases. *Yale Law Journal* 120 (2010), 1278–1367.

[38] Jennifer Skeem, John Monahan, and Christopher Lowenkamp. 2016. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior* 40, 5 (2016), 580.

[39] Sonja B. Starr. 2014. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review* 66, 4 (2014), 803–872.

[40] Michael Tonry. 2013. Legal and Ethical Issues in the Prediction of Recidivism. *Federal Sentencing Reporter* 26, 3 (2013), 167–176.

[41] Alice Xiang. 2021. Reconciling legal and technical approaches to algorithmic bias. *Tennessee Law Review* 88, 3 (2021). forthcoming.

[42] Crystal Yang and Will Dobbie. 2020. Equal Protection Under Algorithms: A New Statistical and Legal Framework. *Michigan Law Review* 119, 2 (2020), 291–396.

A LEGAL BACKGROUND

Constitutional prohibitions on discrimination are encoded in the Equal Protection Clause of the Fourteenth Amendment. Equal Protection doctrine is primarily committed to the anticlassification principle, which prohibits policies that, overtly or surreptitiously, condition on a forbidden category [36]. However, this does not mean that any consideration of a protected characteristic is prohibited. Instead, under current doctrine, Equal Protection claims are assessed using different tiers of scrutiny, which aim to reflect the presumptive legitimacy or illegitimacy of particular justifications for considering protected characteristics. Higher levels of scrutiny require an increasingly important state interest at stake and a closer nexus between that end and the means used to accomplish it [28].

If a law or policy considers an individual's race, it is subjected to *strict scrutiny*. Strict scrutiny requires that the law or policy advance a compelling state interest and that it is narrowly tailored to that interest.[10]

Today, the most relevant example of a race-conscious practice surviving judicial review can be found in affirmative action in higher education. In this context, the Court has determined that "educational benefits that flow from a diverse student body" serve a compelling state interest.[11] Under this rationale, public universities may consider an applicant's race, so long as the assessment of each candidate is individualized and race is not used as a penalty or for quotas. Although largely untested, an additional exception may potentially be found in the context of government contracting, where the Supreme Court has indicated that programs prescribing a more favorable treatment of minorities could, in principle, survive constitutional challenges if their purpose is to remedy historic discriminatory practices [20].[12]

In contrast to race-based classifications, gender-based discrimination is subject to a lower standard of review, called *intermediate scrutiny*. Under this standard, discrimination requires an "important" or "exceedingly persuasive" interest and a "substantial relationship" between the discriminatory law or policy and the interest in order to survive a constitutional challenge.[13] As with race, the case law is not exceedingly instructive to paint a complete picture of the possible exceptions to the rule that policies should eschew the use of gender.[14]

Few cases have assessed discrimination in the context of algorithmic decision making. Perhaps the most relevant case to date is *State v. Loomis*, decided in 2016 by the Wisconsin Supreme Court.[15] In that case, the male petitioner, Eric Loomis, challenged the use of COMPAS, a risk assessment tool used to inform sentencing decisions. Loomis claimed that the use of the proprietary algorithm violated his right to due process, and that it impermissibly took into account information about the defendant's gender. The court declined to rule on whether the use of gender in algorithmic risk assessment violates the Equal Protection Clause, and limited its analysis to the question of whether such a practice is consistent with due process. On this question, the court determined that the use of gender was not prohibited, because it was deployed for the nondiscriminatory purpose of promoting accuracy in predictions [14].

Further, the court noted that Loomis had not provided sufficient evidence that gender had actually factored into his sentence. Although the decision was appealed, the Supreme Court denied hearing the case. *Loomis* can be considered

[10] Loving v. Virginia, 388 U.S. 1, 11 (1967). Historically, the strict scrutiny standard has been deployed in cases justifying government practices which, today, appear profoundly unethical. In *Hirabayashi v. United States*, 320 U.S. 81, 95 (1943), and *Korematsu v. United States*, 323 U.S. 214 (1944), the Supreme Court held that curfews and forced relocations specifically targeting Japanese Americans were constitutional because they would be "necessary to meet the threat of sabotage and espionage which would substantially affect the war effort and which might reasonably be expected to aid a threatened enemy invasion." *Korematsu* has since been explicitly repudiated by the court. Trump v. Hawaii, 138 S. Ct. 2392 (2018).

[11] Grutter v. Bollinger, 539 U.S. 306 (2003).

[12] Although preferential treatment of minorities was held unconstitutional in the specific cases, the legal test is laid out in Wygant v. Jackson Bd. of Educ., 476 U.S. 267, 277-78 (1986); City of Richmond v. J.A. Croson Co., 109 S. Ct. 706 (1989).

[13] Craig v. Boren, 429 U.S. 190, 197 (1976); United States v. Virginia, 518 U.S. 515, 530 (1996).

[14] Prior cases in which the Supreme Court assumed a sufficiently important state interest often appear outdated. For instance, it assumed that the state's interest in avoiding "illegitimate pregnancies" would justify a California statute that punished rape only if the sexual intercourse involved a woman. Michael M. v. Superior Court, 450 U.S. 464, 471 (1981). That law has since been revised.

[15] State v. Loomis,881 N.W.2d 749 (2016).

a first indication that the use of gender in algorithmic decision-making is not categorically illegal. That said, future developments in the case law will help sketch out the precise contours of the relevant exceptions.

In sum, while anti-discrimination doctrine generally disfavors race- and gender-conscious policies, courts recognize exceptions where such policies serve important ends that could not be achieved without attention to protected characteristics. How this framework applies to algorithmic decision making remains to be seen. The scholarly literature, for its part, is converging on the view that the use of protected features in algorithmic decision making is or should be strictly prohibited. This sentiment is often justified on normative grounds: it is asserted that it would be immoral to condition algorithmic decisions on an individual's group membership [12, 17, 39, 40].

However, given that the use of protected characteristics can sometimes lead to significantly improved outcomes for members of the disadvantaged groups, some commentators have recently suggested a more nuanced approach [25]. For instance, Huq [21, 22] argues that Equal Protection law can accommodate the use of race-conscious algorithms in the criminal justice system and that many of the justifications underlying the anticlassification principle do not apply when the decision is made by an algorithm. Bent [4] and Ho and Xiang [20] suggest that a race- or gender-specific algorithm may be permissible if it seeks to remedy past discrimination. And on a more fundamental level, Yang and Dobbie [42] join previous critiques of the Supreme Court's embrace of the anticlassification principle [16, 34]. They seek a revival of the antisubordination principle under Equal Protection doctrine, which would shift focus from process to outcomes and would then allow the use of protected characteristics for remedial purposes. It thus appears at least possible that the inclusion of protected characteristics, if used for the benefit of members in the protected group, would be held to be consistent with current anti-discrimination laws.

B STUDY DESIGN

We designed and conducted a 3-part experiment in which participants were randomly assigned to receive information about the outcomes of an algorithmic decision-making process in the context of pretrial risk assessments. For this study, we partnered with Prolific Academic to recruit a representative sample of the U.S. population (n=1,009).[16] The study was approved by the Institutional Review Board at our university,[17] and was pre-registered with AsPredicted.[18] Respondents were paid $3 for their participation in the study, which took approximately 10 minutes to complete. The main survey was administered via Qualtrics.com in July of 2020.

After completing a CAPTCHA and attention check, participants were asked to read a short scenario and make judgments about it. The scenario began by explaining the practice of pretrial detention:

> When a person is accused of a crime, a judge typically determines whether the defendant should remain in jail until his or her trial, or whether he or she can be safely released under specific conditions while waiting for the trial.

B.1 Domain manipulation

Participants were then introduced to the proposal of using algorithmic risk assessments in decisions about pretrial release. Half of participants were asked to consider whether the risk assessment tools should take into account the race of the defendant; half were asked to consider the same question as it pertained to defendant gender. Subsequently,

[16] The sample is representative with respect to sex, age and ethnicity.
[17] IRB Protocol #56325
[18] Preregistration with AsPredicted (www.aspredicted.org). AsPredicted#45125.

participants were randomly assigned to one of three information conditions, described below, before being asked to report their preferences.

B.2 Information manipulation

No Context (Control Condition): Participants assigned to the control condition (n = 126) were given no further information before answering the outcome measures, which (as described below) measured their preference regarding the inclusion of race (or gender) as a feature in algorithmic risk assessments, and probed the reasoning behind their preference.

General Context: Participants (n = 120) assigned to the "general context" condition were told that some experts favor allowing risk assessment tools to take account of race [gender] because they fear that blinding can sometimes lead to discrimination against protected groups. The purpose of this condition was to mimic the type of general information respondents may receive if they follow political discourse or public debates in the media that invokes expert opinions on the design of risk assessment tools. Our description of expert opinions for the race-specific algorithm follows the description of Mayson [31]. Respondents in the race domain were told:

> Some experts have argued that not considering a defendant's race could ultimately lead to worse outcomes for black defendants. This is a summary of their argument: "Black men are frequently arrested for minor violations of the law, whereas white men are let go. So a black man with three prior arrests does not indicate a particularly high risk to the public. In contrast, a white man with three prior arrests indicates that the defendant had serious conflicts with the law. A race-blind algorithm would not take into account this difference. It would treat a black defendant and a white defendant with prior arrests as posing an identical risk to the public. In this way, a race-blind algorithm perpetuates racial inequality by overestimating the risk posed by black defendants and underestimating the risk posed by white defendants. It discriminates on the basis of race." Based on this rationale, some experts have suggested that race should be included in a risk assessment.

Our description of expert opinions for gender-specific algorithms was designed to be a content-adjusted, close analogue to the description on race. Thus, the "general context" intervention for gender read:

> Some experts have argued that not considering a defendant's gender could ultimately lead to worse outcomes for female defendants. This is a summary of their argument: "Women are much less likely to be involved in a violent crime than men. Similarly, upon release, a female defendant is less likely to commit a violent crime than a male defendant. A gender-blind algorithm would not take into account this difference. It would treat the male and female defendant as posing an identical risk to the public. In this way, a gender-blind algorithm perpetuates gender inequality by overestimating the risk posed by women and underestimating the risk posed by men. It discriminates on the basis of gender." Based on this rationale, some experts have suggested that gender should be included in a risk assessment.

After reading the relevant passage, participants were asked to imagine two different tools for risk assessment: a race-blind tool, which does not take into account the defendant's race when it assess a defendant's risk to the public, and a race-specific tool which does take account of race. Those assigned to the gender domain were asked to consider a gender-blind vs. gender-specific algorithm.

Specific Context: The remaining participants (n = 613) were assigned to one of the five "specific context" conditions. Instead of receiving a general statement from an expert about the likely effects of blind versus tailored algorithms,

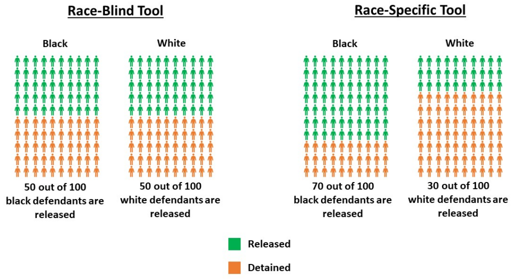

Fig. 5. An illustration of the information presented to respondents in one of the "specific context" conditions. In this case, while the race-blind tool releases an equal proportion of Black and white defendants, the race-specific tool releases a greater proportion of Black defendants than white defendants.

participants in the "specific context" conditions were shown specific information comparing the racial [gender] composition of the pool of defendants who are released under a blind versus tailored algorithm. They were presented with the results from a "hypothetical study" and given illustrations showing how many out of 100 Black [female] defendants and how many out of every 100 white [male] defendants would be released and detained, under blind and tailored algorithms (Figure 5). In all five "specific context" conditions, they observed that the blind algorithm was projected to release 50% of individuals in the protected and unprotected groups. The projections for the tailored algorithm, however, varied systematically, sometimes portraying the tailored algorithm as advantaging, sometimes as disadvantaging, and sometimes as not affecting, protected groups as compared to the blind algorithm. In particular, respondents in the "specific context" condition were randomly assigned to receive one of the following five projections for the tailored tool:

- Release 30% of Black [female] defendants and 70% of white [male] defendants
- Release 45% of Black [female] defendants and 55% of white [male] defendants
- Release 50% of Black [female] defendants and 50% of white [male] defendants
- Release 55% of Black [female] defendants and 45% of white [male] defendants
- Release 70% of Black [female] defendants and 30% of white [male] defendants

An illustration is presented in Figure 5. The purpose of varying the projections for the tailored tool was to investigate how participants' support for tailored risk assessments is affected by different possible outcomes, ranging from "burdens protected groups a lot" to "burdens a little" to "no effect as compared to blind algorithms" to "helps protected groups a lot." Testing this range of outcomes also enables us, through comparison with the control condition, to infer what sorts of projections people naturally make about the likely effects of blind versus tailored algorithms, when given no information at all. For instance, do "no context" participants look like "specific context" participants who've learned that tailoring disadvantages protected groups?

A comprehension-check question asked participants to report which tool releases more Black [female] defendants, offering three choices: race-blind tool releases more defendants, race-specific tool releases more defendants, and both tools release the same number. Before measuring their preferences, participants were asked to presume that the study results they saw would hold true in their state.

B.3 Outcome measures

The primary outcome measure asked participants to choose between a race-blind risk assessment tool and a race-conscious tool (or, for those assigned to the gender domain, to choose between a gender-blind and gender-conscious tool). Participants were offered five answer choices:

- Strongly prefer race-blind tool
- Slightly prefer race-blind tool
- Slightly prefer race-specific tool
- Strongly prefer race specific tool
- Other

Next, participants were presented with an empty text box and asked to explain their response. Following the open response, participants were presented with two rationales for why one might believe that risk assessment tools should not take into account defendants' race (or gender). The order of the presentation of the two rationales was randomized. One rationale reflected a deontological justification by suggesting that blinding is unethical *per se*:

> I wouldn't want to let a risk assessment tool take into account people's race [gender]. It's wrong to categorize people based on their race [gender], because this means that the decision whether or not to release a black [female] defendant is based on what other black [female] defendants have done in the past. It is important to treat people as individuals without paying attention to their race [gender].

The other rationale reflected consequentialist justifications, voicing instrumental reasons for avoiding blind algorithms:

> I wouldn't want to let a risk assessment tool take into account people's race [gender]. Paying attention to race [gender] could lead to more black defendants [women] being detained and more white defendants [men] being released. It is important to make sure that black [female] defendants are not disadvantaged by algorithmic decision-making, which is best achieved by removing race [gender] from the calculation altogether.

Participants were asked to indicate which of the above two views most closely aligns with their own. Following these primary measures, participants provided additional information regarding their general views on incarceration, as well as several demographic factors, such as their age, gender, political affiliation, and education, including their familiarity with law and computer science.

In a second pre-registered study in August of 2020,[19] we partnered with Prolific Academic to recruit 209 U.S.-based computer scientists to complete the same survey measures.[20] Due to sample size constraints, we randomly assigned participants to only two information conditions: the control condition ("no context") and the "general context" condition. This manipulation was crossed with the race-vs.-gender manipulation, yielding a 2x2 factorial design. In a third pre-registered study in September of 2020,[21] we partnered with Centiment[22] to recruit 249 practicing lawyers in the United States. They completed the same survey as the computer scientists. We collected responses from these two additional groups under the assumption that they represent two of the most relevant actors in the process of algorithmic design:

[19] AsPredicted#46771

[20] We collected 242 responses that, according to Prolific's prescreening, were computer scientists or computer science students. However, 33 of them did not respond affirmatively when we asked them whether they had or were working on a computer science degree, so we excluded them from our analysis.

[21] AsPredicted#47852

[22] Centiment.co

The computer scientists as those who design the algorithms and the lawyers as those who impose regulatory constraints on their design.

B.4 Limitations

There are several potential limitations with our study design that we intend to address. First, under the "specific context" condition, we only inform respondents about the proportion of released defendants within each group. We omit other information that may be important in making an informed fairness decision, such as the effect on public safety. In opting for this design, our goal was to make the consequences of the group-specific algorithm as transparent as possible while minimizing the cognitive load on respondents. For instance, providing partial information about the algorithm's performance in order to inform public safety considerations would have raised the question of what performance measures to share. Given the breadth of potential measures as well as their differing interactions with considerations of fairness [7], we determined that providing respondents with only some information that is relevant to public safety would be infeasible and would hurt, rather than help comprehension. At the same time, providing respondents with all relevant information, including various performance metrics, would have increased the cognitive demand significantly. We thus chose to rely on respondents' priors about public safety implications and fairness. We note that none of the free text responses implied that respondents felt they needed additional information about public safety before being able to make a decision.

Next, and related, one may be concerned that respondents in the "specific context" condition knew about the hypothetical nature of the results and thus did not take it seriously. In order to rule out that this assumption affects our responses, one "specific context" condition presented participants with real data from Broward County, FL. Half of participants who saw the Broward County data were then told that the study results are hypothetical, while the other half were told that the results are from a real study. We found similar stated preferences across the two groups, suggesting that the hypothetical nature of the information does not appreciably affect our results.[23]

In addition, one may be concerned that we measured respondents' preferences for outcome-focused and process-focused reasoning only after informing them about the potential effects of the blind tool. In order to rule out any priming effects, we compared preferences for deontological vs. consequentialist reasoning between those in the "no context" condition (where no priming took place) and those in other conditions, finding similar results across groups.[24]

We also note that our study is geographically limited to the U.S., and contextually limited to the use of algorithms in pretrial risk assessments. Without additional assumptions, any findings can not be extrapolated to different countries or contexts.

[23] The average response among those who were told the study was real is -1, corresponding to a weak preference for the blind tool. Among those who were simply asked to imagine the results hold true, the average response was -0.86. The difference is statistically insignificant.

[24] The share of respondents that preferred outcome-centered reasoning is 0.25 among those in the "no context" condition and 0.32 among all others.