Procedural Justice and Risk-Assessment Algorithms*

A. J. Wang[†]

Yale Law School

June 21, 2018

Abstract

Statistical algorithms are increasingly used in the criminal justice system. Much of the recent scholarship on the use of these algorithms have focused on their "fairness," typically defined as accuracy across groups like race or gender. This project draws on the procedural justice literature to raise a separate concern: does the use of algorithms damage the perceived fairness and legitimacy of the criminal justice system? Through three original survey experiments on a nationally-representative sample, it shows that the public strongly disfavors algorithms as a matter of fairness, policy, and legitimacy. While respondents generally believe algorithms to be less accurate than either of these methods, accuracy alone does not explain their preferences. Creating "transparent" algorithms helps but is not enough to make algorithms desirable in their own right. Both surprising and troubling, members of the public seem more willing to tolerate disparate outcomes when they stem from an algorithm than a psychologist.

Keywords: algorithms; procedural fairness; bail setting; statistics.

[*] Pre-registration of analysis is available at https://goo.gl/3XCccy. This most up-to-date version of this working paper is available at https://ssrn.com/abstract=3170136.

[†] I am grateful for the wisdom, advice, and support of Tom Tyler and the students of the Spring 2018 Empirical Research Seminar. This project would not have been possible without the efforts of Michael Sadowsky and David Shor at Civis Analytics. Finally, I thank Ian Ayres, Kristen Bell, John Bullock, Alex Coppock, James Michelson, Nathaniel Olin, Olivia Robinson, and Anna Windemuth for their helpful comments and feedback.

An algorithm cannot take into account factors such as human emotion and need. It is stupid to allow a mathematical equation, the value of which is only as useful as the data it utilizes to operate, to determine the fate of a being that possesses free will. That's outright absurd, stupid, and dangerous.

I think using the human element is most fair and humane. Guidelines would be second best, but seem to lack some humanity. An algorithm is the most cold. People are not statistics.

Someone's fate shouldn't be determined by an algorithm.

Survey respondents, assessing the use
of algorithms in bail hearings

Introduction and Theory

Although mathematical formulae and statistical analyses have been applied to criminal justice since the 1920s (Mathiesen, 1998, pg. 458), their use has increased with the development of modern machine learning algorithms. Currently, statistical algorithms are used in a variety of places in the criminal justice system, though most notably in bail and sentencing (Electronic Privacy Information Center, 2018; Huq, 2019). This project focuses on the use of algorithms in the bail risk-assessment context, but the general principles of predictive modeling and algorithm design apply to other contexts (see Kehl et al., 2017, pg. 13-15).

In the bail context, many courts use risk-assessment algorithms to predict whether a defendant is likely to recidivate if released or will fail to appear at a future court date, i.e., to predict future behavior (Workgroup, 2017, pg. 51). These risk-assessment algorithms may incorporate a variety of characteristics such as the defendant's past criminal history, income, gender, peer networks, and other demographic characteristics to generate predictions (Picard-Fritsche et al., 2017, pg. 5-6). Some courts view the use of risk-assessments as an improvement over the common practice of making bail decisions based on the charge alone—without regard for community danger, the likelihood of appearance, or ability to pay (see, e.g., ODonnell v. Harris County, pgs. 1159 and 1161) (describing the adoption of risk-assessments as "laudable" and "salutary"). Given the recent success of bail reformers in reducing wealth-based disparities in pre-trial detention, the interest in risk assessment tools is only likely to increase.

Much of the recent work on the use of algorithms in the criminal justice system falls into two categories. First, a growing number of legal commentators have highlighted the constitutional challenges posed by algorithms. They argue that the use of algorithms may violate the Equal Protection Clause because algorithms often operate differently across protected classes like race or gender (Huq, 2019; Starr, 2014).

The second stream of work, largely based in computer science, attempts to define and formalize these disparities under the broad label of "algorithmic fairness." The exact definition of fairness varies in this context, but it broadly centers on how algorithms generate results across classes such as racial or gender categories (Corbett-Davies et al., 2017; Kleinberg et al., 2017; Berk et al., 2017).

But the term "fairness" is actually somewhat misleading in this context. For example, many authors discuss two conflicting notions of "fairness": false positives (e.g., when a defendant is flagged as "high risk" when they are low risk) and false negatives (e.g., when a defendant is flagged as "low risk" when they are high risk).[1] Under this framework, fairness is defined as accuracy across groups. One major concern of this literature is to document how these two (and other) notions of fairness are in conflict: the more an algorithm minimizes false positives, the more it generates false negatives (Feller et al., 2016). Calibrating an algorithm to favor one form of accuracy over another, in turn, often has disparate impacts across racial groups (Kleinberg et al., 2017).[2] In part stemming from this concern, many policymakers and computer scientists call for organizations that use algorithms to release more information about how algorithms operate and what inputs they use (Goodman and Flaxman, 2017). Many commercial algorithms neither disclose which features are incorporated nor what statistical transformations are applied. According to the Electronic Privacy Information Center, over 40 states use a risk-assessment tool at some point within the criminal justice system (2018).

While it would be overreaching to claim that accuracy across groups is *irrelevant* to fairness, it is equally so to view accuracy as the *only* determinant of fairness. Indeed, the literature on procedural justice suggests that whether individuals perceive a procedure to be fair is the consequence of many different factors, of which accuracy is only one. In particular, the literature on procedural justice highlights four factors: "opportunities for participation, a neutral forum, trustworthy authorities, and treatment with dignity and respect" (Tyler and Sevier, 2013).[3] Procedures that are weak along these dimensions—i.e., processes that lack procedural justice—present not only normative but also empirical problems for the criminal justice system. We might intrinsically care about popular legitimacy because we believe that defendants ought to be judged by procedures they believe are fair. But even outside of subjective beliefs, a lack of procedural justice is associated with lower levels of compliance with the law (Tyler, 1990).

Algorithms have a mixed record along these dimensions (Simmons, 2018). On the one hand, algorithms are "neutral" in the sense that, conditional on the same inputs, algorithms will always produce the same results. The same set of facts put in front of different judges, by contrast, can result in substantial variance in outcomes (Ramji-Nogales et al., 2007). On the other hand, algorithms offer minimal opportunities to be heard. Defendants only provide information and do not influence the workings of the algorithm. The extent to which algorithms are trustworthy is unknown, and the opportunity for treatment with dignity and respect is minimal—insofar as algorithms minimize the need for interaction at all. For an extended discussion, see pages 9 to 16 of (Simmons, 2018).

Despite the risk that algorithms undermine procedural justice, relatively few works discuss the procedural justice implications of algorithms. In contrast to widespread beliefs among

[1] Berk et al. (2017) identifies at least four other forms of fairness.

[2] The "algorithmic" part of this literature is also somewhat misleading. To the extent the literature creates metrics for different definitions of fairness, these definitions apply equally to all procedures that generate some predictions across different groups—algorithmic or otherwise.

[3] See also MacCoun and Tyler (1988).

experts that participants would exhibit "algorithm aversion," recent work has shown that people trust algorithms over experts for tasks like estimating someone's weight based on a photo, recommending romantic matches, and predicting the popularity of songs (Logg et al., 2018).[4]

But many of these commercial contexts are significantly different from the criminal justice system, which—at least historically—has placed great weight on human interaction and human decision-making. In a small-sample qualitative study that examined contexts where algorithmic involvement would be unusual, such as applying for a promotion at work and applying for a personal loan in addition to contexts where algorithmic involvement is more common, such as re-routing passengers on overbooked flights, or pricing insurance premiums, many participants reacted negatively to the "impersonal and dehumanising" features of algorithmic decision-making (Binns et al., 2018). Within the criminal justice context, an ethnographic study of two criminal courtrooms revealed how legal professionals (including judges, prosecutors, and probation officers) use a variety of tactics to minimize the influence of algorithms (Christin, 2017). For example, they would ignore the output of algorithms, sometimes manipulate the inputs to achieve the desired outcome, and openly criticize the use of algorithms.

The only work that addresses the issue of procedural justice in the criminal justice directly is Ric Simmons's article "Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System" (forthcoming in the University of California–Davis Law Review) (2018). Using a 580-person Mechanical Turk sample, Simmons shows that respondents are as likely to accept the outcomes of predictive algorithms as they are the decisions of judges. The Simmons study, however, asks people to choose between two procedures: either a judge makes decisions or a judge makes decisions but is given algorithm results as a piece of information to use. In that situation, people are largely indifferent between the two options.[5]

This project presents evidence to the contrary. Using a larger, nationally-representative sample, I show that the public rejects the use of algorithms as a matter of fairness and of policy. I demonstrate through three original survey experiments that many of the concerns that scholars have raised about algorithms are insufficient to explain why individuals dislike them so much.

Experiment 1 establishes, as a baseline measure, that algorithms are extremely unpopular, and importantly, that even when algorithms are presented as being as accurate as other procedures, they are still unpopular. Experiment 2 asks whether transparency improves opinions of algorithms. The answer is yes, but not to an extent that makes people actu-

[4] Note, however, that earlier studies from the 1950s and 1970s indicate that respondents strongly prefer human decision-making over algorithmic decision-making (Dietvorst et al., 2015).

[5] Simmons's work is methodologically distinct from the existing project in a number of ways. It does not use the term "algorithms," instead referring to them as "computer programs." In addition, Simmons varies the acceptability of a judge making a decision without the help of an algorithm versus a judge making a decision with the help of an algorithm. Notably, these treatments have substantially different lengths. The treatment mentioning "computer programs" is about 250 words whereas the one without is about 150 words. This 70% $^{(250-150)}/_{150}$ difference in length may account for the lack of difference between the two conditions: the treatment may be buried in the length of the text. Finally, with a sample size of 600, Simmons is only powered to detect differences in a binary outcome of about eleven percentage points. With a sample size of 3,000, this project is powered to detect differences of five percentage points. Indeed, Simmons presents no statistical tests at all to measure the magnitude or uncertainty of the effects detected.

ally want to use algorithms. Finally, Experiment 3 addresses the emerging literature on "algorithmic fairness." It shows that while respondents have a negative reaction when psychologists exhibit disparate impact, they are indifferent when algorithms exhibit the same disparate impact.

As technology improves or as regulations are created, many of the concerns of algorithm critics will be met: "fairness" concerns decrease as algorithms increase in accuracy, and transparency can be achieved through regulatory fiat. But this project suggests that, when this day arrives, Americans will still dislike algorithms. As the criminal justice system increasingly embraces these impersonal, statistical and computational tools, courts increasingly risk a decline in legitimacy—a danger that is almost entirely ignored by existing scholarship. By acknowledging and documenting the problem, this project tries to take the first step forward towards solving it.

Experiment 1: Understanding Preferences

This section is guided by two broad questions:

1. What are people's preferences for different procedures that might be used in bail hearings?
2. To what extent are these preferences attributable to evaluations of accuracy?

This experiment measures people's attitudes towards three types of risk assessment procedures: expert psychologists, mandatory guidelines, and statistical algorithms. In practice, none of these procedures are the sole determinant of bail hearings, but what they have in common is that they reflect the portion of bail hearings that rely on *expert* judgment. In addition to establishing procedural preferences, this experiment also assesses whether individuals' preferences change based on the accuracy of the procedure.

Design

The overall structure of the experiment is described in the Figure 1. I first randomly assign individuals to view one of the procedures. Respondents are asked to assess the single procedure on fairness ("In your opinion, would you say this procedure is Very Fair, Somewhat Fair, Unsure, Somewhat Unfair, or Very Unfair?") and policy ("If you were arrested, would you want the city to use this procedure for your case?"). As a measure of legitimacy, respondents are also asked that if this procedure were adopted, how much would they agree with the statement "You should accept the decisions made by courts, even if you think they are wrong." After individuals rate a single procedure, I show them all three procedures and ask respondents to rank them based on fairness ("How would you rank these procedures in terms of fairness?"), their desire to use it in a proceeding if they were arrested ("How would you rank the procedures based on whether you would want the city to use them if you were arrested?"), and their desire to use it in a proceeding for everyone in their city ("How would you rank the procedures based on how much you would want to the city to use them for everyone that is arrested?”).

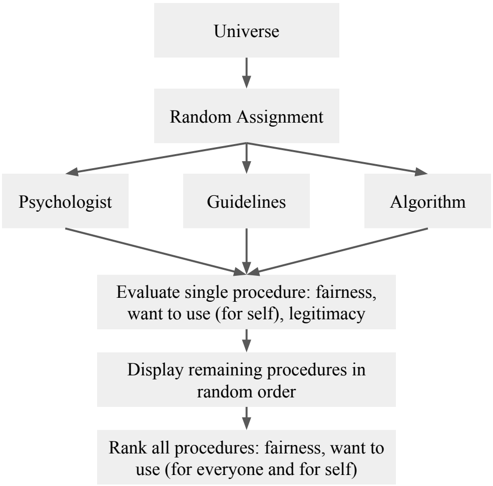

Universe
Random Assignment
Psychologist
Guidelines
Algorithm
Evaluate single procedure: fairness, want to use (for self), legitimacy
Display remaining procedures in random order
Rank all procedures: fairness, want to use (for everyone and for self)

Figure 1: Diagram of Experiment 1

In addition to randomizing the procedure order, I also randomly assign the information that respondents receive about the accuracy of the different procedures. Half the sample receives no information about accuracy. For respondents who are assigned to view accuracy, I also randomize the accuracy level of the different procedures into either “80% Accuracy” or “60% Accuracy,” which I refer to as the “High Accuracy” and “Low Accuracy” conditions based on Table 1.

Accuracy of Psychologists	Accuracy of Guidelines	Accuracy of Algorithm	Probability of Assignment
Low	Low	Low	$^1/_{28}$
High	High	High	$^1/_{28}$
Low	Low	High	$^1/_{14}$
Low	High	Low	$^1/_{14}$
Low	High	High	$^1/_{14}$
High	Low	Low	$^1/_{14}$
High	Low	High	$^1/_{14}$
High	High	Low	$^1/_{14}$
None	None	None	$^1/_2$

Table 1: Treatment Assignment of Accuracy Condition in Experiment 1

In total, this experiment has 3,369 respondents, drawn from a nationally representative sample managed by Civis Analytics, a polling and data science firm. Descriptive statistics are calculated using the survey weights provided by Civis Analytics (weighted to a national general population sample). Following Miratrix et al. (2017), causal analyses estimate the sample average treatment effects (calculated without weights).[6]

Results

We find three main results from this experiment:

1. Respondents strongly disapprove of algorithms as a matter of fairness, a preference of policy, and a source of legitimacy. Depending on the outcome, they are either ambivalent about a choice between guidelines and psychologists, or they prefer psychologists. They never prefer algorithms.
2. This rank order aligns with respondents' priors about accuracy because they view algorithms as the least accurate procedure. However, this ordering cannot be explained by differences in accuracy alone. Even when respondents are told that procedures have the same average levels of accuracy, they still disfavor algorithms.
3. When algorithms are stipulated to be more accurate than other procedures, respondents prefer algorithms. Indeed, respondents on average choose the more accurate procedure, whether it is algorithms, psychologists, or guidelines. However, when the more accurate procedure is an algorithm, respondents prefer it at lower rates than when the more accurate procedure is a psychologist or guideline. Respondents appear to balance a general aversion to algorithms against a desire for high accuracy. Without information about accuracy, these preferences are consistent because respondents view algorithms as inaccurate.

Which Procedures Do People Prefer?

Across different types of preferences (fairness, policy, and legitimacy) and across different question types (evaluating a single procedure and ranking all three procedures), the results are consistent: **respondents strongly disapprove of algorithms** as a matter of fairness, a preference of policy, and a source of legitimacy. Depending on the measure, they are either ambivalent about guidelines and psychologists, or they prefer psychologists.

Below, I show some representative results. Results that are not displayed follow similar patterns. The following analyses are based on respondents who did not receive any information about accuracy (n = 1,713) and thus reflect preferences absent external manipulations.

As shown in Figure 2, respondents viewed algorithms as less fair than other procedures. On average, respondents rated algorithms at 0.4 (on a scale that ranged from -2 to +2) compared to approximately 0.8 for the other two options. This difference in magnitude of 0.4 constitutes approximately 10% of the total scale.

[6] This decision was *not* specified in the pre-analysis plan. I decide not to use weights because a limited number of the causal analyses have somewhat different outcomes with and without weights, suggesting that the weights may be misspecified and the population average treatment effects may be "undetectably biased" (Miratrix et al., 2017, pg. 25).

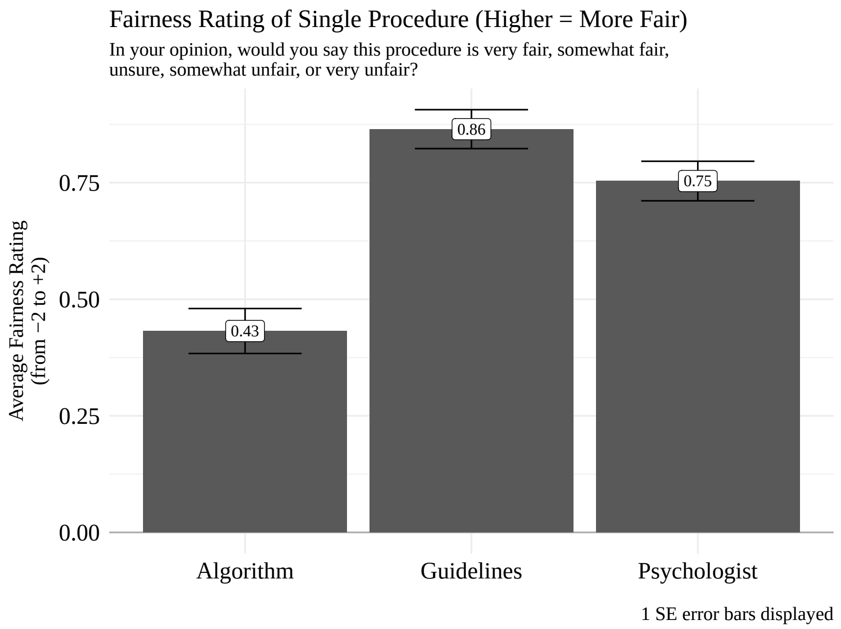


Figure 2: Respondents Strongly Disfavor Algorithms as a Matter of Fairness

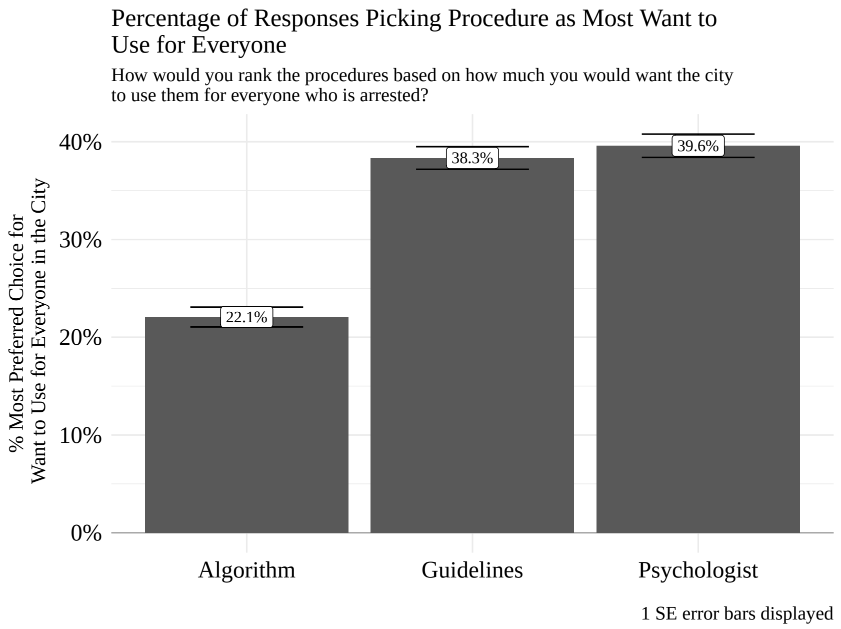


Figure 3: Respondents Strongly Disfavor Algorithms as a Matter of Policy

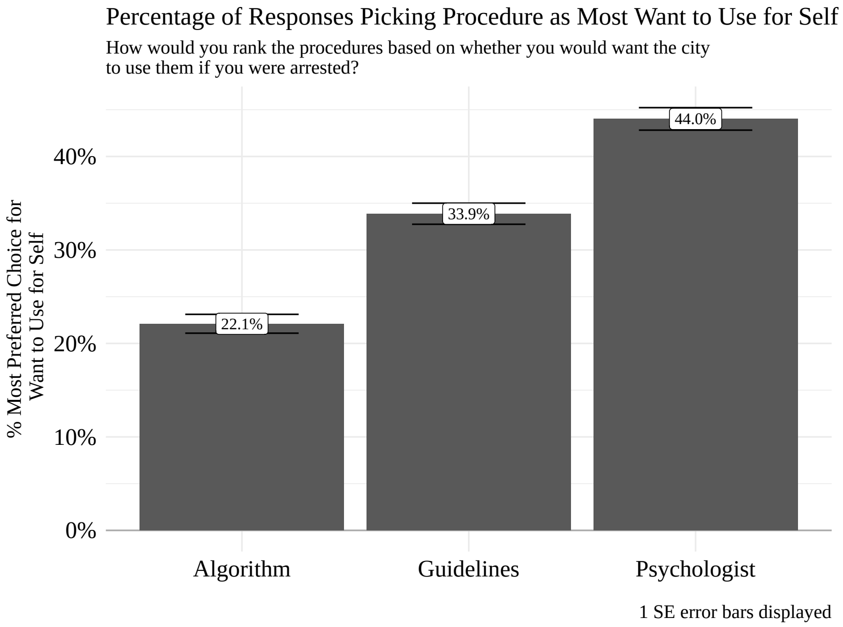


Figure 4: Respondents Most Prefer Psychologists

The results after viewing a single procedure are consistent with asking respondents to rank all three procedures. Figure 3 displays the average rank of each algorithm, recoded so that the most preferred procedure was recorded as +2, the second-most preferred procedure was recorded as +1, and the least-preferred procedure was recorded as 0. Again, algorithms were the least-preferred procedure. A Friedman rank sum test suggests these differences are significant at $p < 0.001$.[7]

In practice, when individuals make policy choices from multiple options, they rarely rank different outcomes. Instead they only indicate their most preferred policy. We can approximate this analysis by asking what percentage of the time a procedure is chosen as the most preferred policy, as shown in Figure 4.

In short, when we examine average rankings, we typically see that guidelines and psychologists are equally preferred. But if we isolate the question to people's most preferred policy, we see psychologists pull out ahead. Why is this? Among respondents who picked psychologists as their first choice, they were far more likely to choose guidelines as their second choice than algorithms as their second choice. Similarly, those who picked guidelines as their first choice strongly favored psychologists over algorithms. However, among respondents who picked algorithms as their first choice, they broke evenly between guidelines and psychologists for their second choice.

[7] For a discussion of the Friedman rank sum test, see Hollander and Wolfe (1999). Though not a causal analysis, these results are reported using the unweighted data.

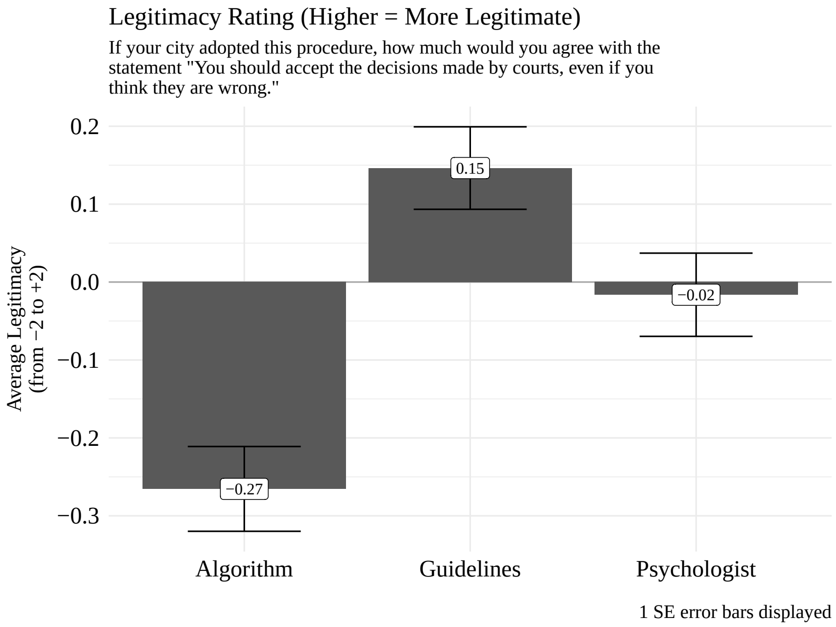


Figure 5: The Use of Algorithms Decreases Legitimacy

These disagreements are not confined to matters of policy or fairness: selecting one procedure over another also impacts the legitimacy of the judicial system, as shown in Figure 5. This impact is approximately 0.3 points on a four-point scale (about 7.5%), a modest but statistically significant result (at $p < 0.001$).

How Much Does Accuracy Drive Preferences?

One explanation for these preferences might be that people have different baseline beliefs about the accuracy of these procedures. As shown in Figure 6, respondents possess meaningful beliefs about the accuracy of different procedures. Moreover, these beliefs conform exactly to the rank-order of procedures discussed earlier.

To answer this question, we randomized the accuracy level of different procedures according to the probabilities described in Table 1 ($n = 1{,}656$). We can measure the extent to which preferences changed based on accuracy by comparing the aggregate responses from the condition in which no information about accuracy was given and from conditions in which information about accuracy was given (since, in the aggregate, the average accuracy level was equal across conditions).[8] If respondents' preferences were entirely driven by accuracy, we would expect that preferences should be approximately similar across different procedures

[8] An alternative to this analysis compares preference levels only among respondents were accuracy levels were randomized to be equal across all three conditions. This analysis yields substantively similar results to what is displayed here, but is less powered.

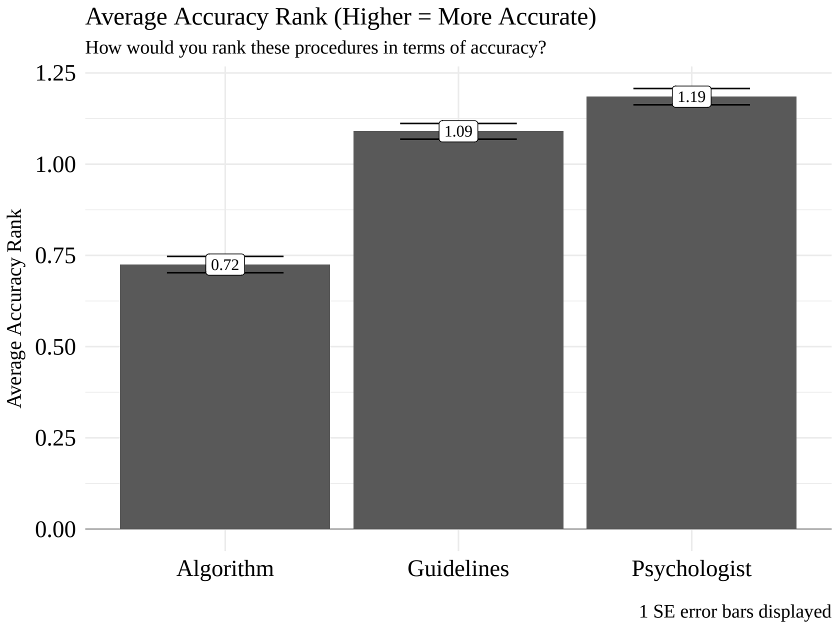


Figure 6: Absent Other Information, Respondents Think Algorithms are Less Accurate than Psychologists or Guidelines

when the procedures have the same average level of accuracy. Instead, the rank-ordering of preferences is identical to the condition in which no accuracy information was displayed.[9]

Does this stability of preferences indicate that respondents are simply insensitive to the accuracy manipulation? As I demonstrate below, this is not the case. Preferences are actually quite sensitive to increases in accuracy. However, the marginal effects of accuracy are similar across all procedures, so they do not overcome baseline preferences for different procedures.

To estimate the effects of accuracy, we first disaggregate the set of accuracy levels reported in Table 1 into a set of pairwise comparisons between (1) algorithms and psychologists and (2) algorithms and guidelines.[10]

We can calculate the marginal effect of having a high accuracy level by comparing the rate at which a procedure is chosen when its accuracy level is the same as versus higher than the accuracy level of other procedure. We estimate the following model using ordinary least squares regression for each pairwise comparison:

$$Y = \text{Relative Accuracy} + \text{Accuracy of Omitted Variable} + \text{Covariates}$$

[9] See Figure A1.

[10] This mapping is reported in Table A1.

where Y is 1 if the respondent preferred the algorithm and 0 otherwise. Relative accuracy takes on the form of their "Same," "Algorithm Higher," or "Other Procedure Higher." Covariates include controls for age, age squared, race, gender, education, household income, marital status, 2012 Presidential vote choice (as a proxy for partisanship), and region.

The results, which are reported in Table 2, show that respondents are quite sensitive to changes in relative accuracy. Moreover, the marginal effects of accuracy appear to be substantively similar across procedures.

	Desire to Use Algorithm for Everyone	
	Compared to Psychologist	Compared to Guidelines
Algorithm More Accurate	0.24^{***}	0.26^{***}
	(0.03)	(0.03)
Psychologist More Accurate	-0.20^{***}	
	(0.03)	
Guidelines More Accurate		-0.24^{***}
		(0.03)
Accuracy: Other	0.03	0.02
	(0.02)	(0.02)
Observations	1,656	1,656

Note: *p<0.1; **p<0.05; ***p<0.01
Covariate controls and intercept omitted.

Table 2: The Effects of Accuracy Are Similar Across Procedures

This result should be puzzling: if the marginal effect of accuracy is similar (and non-zero) across different procedures, then why do respondents disapprove of algorithms even when we experimentally manipulate accuracy? This occurs because people have meaningful preferences even when the accuracy levels are *experimentally manipulated to be identical.* Because of these baseline differences, the marginal effects of accuracy—which are very similar across procedures—do little to change the aggregate preference.

For example, Table 3 shows that when algorithms are the more accurate procedure, respondents prefer algorithms 66% of the time. However, when psychologists are the more accurate procedure, respondents prefer them 79% of the time. And when they are at the same level of accuracy, around 59% of respondents prefer psychologists. The same pattern of preferences appears when comparing guidelines to algorithms.

We find similar results when we analyze the effect of accuracy versus the effect of the procedure from another perspective. Instead of comparing the rank-ordering given different types of relative accuracy, we can leverage random assignment of the first procedure and the accuracy of the first procedure on people's preferences. We estimate the following regression

$$Y = \text{Procedure} + \text{Accuracy} + \text{Procedure} * \text{Accuracy} + \text{Covariates}$$

	Prefer Alg.	Prefer Non-Alg.	Standard Error
Algorithm v. Psychologist			
Psychologist More Accurate	21.2%	78.8%	2.1%
Same Accuracy	41.2%	58.8%	1.8%
Algorithm More Accurate	65.6%	34.4%	2.1%
Algorithm v. Guidelines			
Guidelines More Accurate	20.1%	79.9%	2.2%
Same Accuracy	43.9%	56.1%	1.7%
Algorithm More Accurate	69.7%	30.3%	2.1%

Table 3: Respondents Prefer Algorithms When They Are the More Accurate Procedure, But at Lower Rates than When the Non-Algorithmic Procedure Is More Accurate (Percent Favoring More Accurate Procedure in Grey)

where Y ranges from -2 to +2, Procedure is the algorithm, guideline, or psychologist, and Accuracy is either the high or low condition. We then calculate the marginal effects of different procedures and compare them to different accuracy levels. Because the procedure and accuracy level are randomly assigned, we can directly compare the magnitude of their treatment effects.

The full results are not reported here, but in both assessments of the fairness of a procedure and the respondent's desire to use in a proceeding, the effects of the procedure are roughly comparable to those of accuracy. Looking at the effects on the respondent's desire to use a given procedure, the difference between having a procedure that uses a psychologist versus one that uses an algorithm is 0.46 on a 4-point scale (approximately 11%). This effect is statistically indistinguishable from the effect of having a procedure with a high accuracy level versus a low accuracy level (0.45).

Experiment 2: Transparency

Given that the public generally holds negative views on algorithms, what tactics might make algorithms seem more palatable? A variety of computer scientists and policymakers have called for "algorithmic transparency": a publicly available description of how algorithms work and what inputs they rely on (Goodman and Flaxman, 2017). Such advocates typically argue that transparency is necessary for political and regulatory oversight, so that third parties can recognize and correct algorithms that are incorrect or discriminatory (see, e.g., State v. Loomis).

It seems that the literature on procedural justice would support this view: Tyler and Sevier (2013), for example, notes that transparency is necessary to "facilitate[] the belief that decision making procedures are neutral [by] reveal[ing] that decisions are being made in rule based and unbiased ways." But transparency with respect to algorithms seems somewhat different than with other, human-driven procedures. Algorithms, after all, are statistical models that many people might not understand if they were disclosed. Indeed, even relatively simple machine learning models like random forest models, defy easy interpretation of how a single variable might affect an outcome (Ritter, 2013; Gromping, 2009).

This section is guided by two broad questions:

1. Does transparency (or the lack thereof) impact people's opinions of algorithms?
2. How does this effect change under different levels of accuracy?

Design

The structure of Experiment 2 is virtually identical to Experiment 1 with one main difference: instead of varying the procedure between Algorithms, Psychologists, and Guidelines, respondents are randomly assigned to either evaluate a transparent algorithm ("Transparent" condition) or a non-transparent algorithm ("Non-Transparent" condition).[11] Respondents are also shown short arguments for and against transparency.[12]

This experiment has 3,404 respondents, drawn from the same source as Experiment 1.

Results

We find two main results from this experiment:

1. Respondents prefer transparent to non-transparent algorithms, though they don't have a strong desire for either.
2. Although respondents tend to prefer the more accurate procedure, they do so at higher rates for transparent than non-transparent algorithms.

In general, people tend to prefer transparent to non-transparent algorithms. When asked to rate one procedure, the fact that an algorithm is transparent increases perceptions of fairness by 0.4 (about 10% of the total scale) and desire to use by 0.3 (about 7.5%) of the total, as displayed in Figure 7. When asked to rank the two types jointly, people are more likely to choose the transparent algorithms along all dimensions, as displayed in Table 4. This is true even when the prompt stipulated that the transparent algorithm was $3,000,000 more expensive than the non-transparent algorithm.

But even though people prefer the more transparent algorithm, it's not clear whether they want to use the algorithm at all. When respondents were asked to rate a single procedure

[11] Beyond this, there are two other differences of note. First, the distribution of accuracy ratings in Experiment 2 ("High" and "Low" again) each independently assigned to each treatment condition with 50% probability. Second, when asking respondents to rank the two algorithms types, the survey asks a question about which type the respondent prefers if the transparent one is $3,000,000 more expensive than the non-transparent algorithm.

[12] The full question text is available in the Appendix.

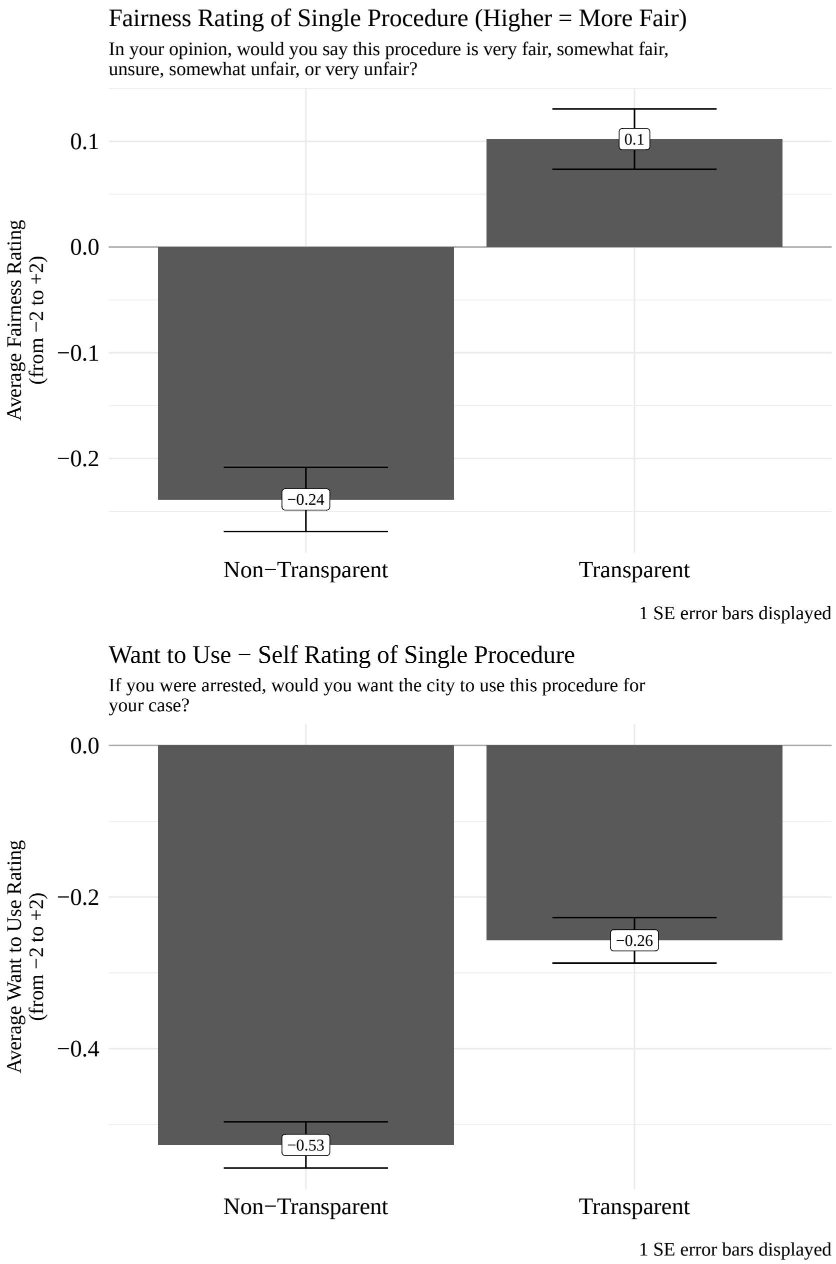


Figure 7: Respondents Prefer Transparent Algorithms (Evaluating a Single Procedure)

Outcome	Prefer Transparent	Prefer Non-Transparent
Fairness	71.4%	28.6%
Want to Use for Self	70.9%	29.1%
Want to Use for All	69.2%	30.8%
Transparent Costs $3M More	59.4%	40.6%

Table 4: Respondents Prefer Transparent Algorithms (When Ranking Procedures)

	Prefer Non-Transparent	Prefer Transparent	Standard Error
Transparent More Accurate	13.3%	86.7%	1.4%
Same Accuracy	23%	77%	1%
Non-Transparent More Accurate	63.4%	36.6%	1.4%

Table 5: Respondents Prefer Non-Transparent Algorithms When They Are the More Accurate Procedure, But at Lower Rates Than When the Transparent Algorithm Is More Accurate (Percent Favoring More Accurate Procedure in Grey)

on whether they wanted to use an algorithm, they gave preferences on a five-point scale that ranged from "strongly do not want to use" (-2) to "strong want to use" (+2). A score of 0, the midpoint, indicated that respondents were unsure about whether they wanted to use it. But as shown in Figure 7, most respondents did not want to use either algorithm (both values are below zero). The transparent algorithm was the "least bad" outcome, not one they actually desired.[13]

As we saw in Experiment 1, respondents generally select the procedure with the higher accuracy level. However, 87% of respondents preferred the transparent algorithm when it was the more accurate one, but only 63% preferred the non-transparent algorithm when it was the more accurate one, as displayed in Table 5. Similar to the results in Experiment 1, these differences are *not* due to the fact that the treatment effect of accuracy differs by transparency. Rather, these differences stem from differences in the base level of support when the two procedures are stipulated to have the same level of accuracy: instead of an even split, 77% prefer transparent algorithms.

Experiment 3: Disparate Impact

In late 2017, a committee created by the California Supreme Court recommended that the state adopt a system of individualized risk assessments, including risk assessment algorithms. In its report, the committee specifically noted that "tools that demonstrate any implicit or explicit bias must not be used" (Workgroup, 2017, pg. 58).

[13] These results do not quite align with those in Experiment 1. There, when respondents were asked to rate how much they would be willing to use an algorithm (without being given any information about transparency and with the algorithm manipulated to have the same level of accuracy as in this experiment), the average rating was 0.06. (The difference from 0 was not statistically significant at $p < 0.1$.) Here, ratings of both algorithms are lower. This might occur because some respondents to Experiment 2 also responded to Experiment 1 first, so they were exposed to the fact that alternatives existed.

As a legal matter, such bias goes to the constitutionality of algorithms under the Equal Protection clause of the Fourteenth Amendment. But precisely mapping that relationship is difficult: as Huq (2019) describes, the existing Equal Protection doctrine is ill-suited to assess the use of algorithms. Both the standard of bad intent and that of bad classification "transfer poorly, if at all" to the use of algorithms (page 29). For an algorithm developer, the classification problem is easy to avoid: she can simply avoid the explicit use of race, gender, or other protected classes as an input.[14] The issue of intent—a lodestone of Equal Protection jurisprudence—is harder. Machines themselves lack intent, but neither the people who develop algorithms (e.g., engineers tweaking hyper-parameters) nor those who produced the data on which the algorithm was developed (e.g., court officials and police officers producing data on who is a flight risk) are so impartial. Huq concludes that "legal policy-makers should not ask *whether* algorithmic criminal justice will be racially nonneutral, but rather *how* it will be so biased" (pg. 6).

While Huq was primarily referring to adjudication between different metrics of fairness for the same algorithm, other scholars have argued that we should not evaluate algorithms in isolation. Rather than labeling algorithms as simply "biased" or "unbiased," the bias of an algorithm should be compared to the bias of a "counterfactual" human decision-maker, such as a judge or psychologist (Cowgill and Tucker, 2017). That is, a jurisdiction might legitimately decide as a policy matter to adopt a "biased" algorithm—as long as the algorithm's bias is less than that of the procedure it replaces.

But this approach seems to assume that we should respond to disparate impact in similar ways whether it stems from the unconscious bias of a human decision-maker or, say, data generated from racially-biased policing. Answering this normative question is outside the scope of this project, but a useful starting point is to understand, as an empirical matter, how the public actually evaluates algorithmic "bias." In particular, when algorithms result in different outcomes across different groups of people, does the public view these outcomes as *bias*ed (and thus a source of unfairness) or merely as a set of disparate outcomes (without impacting fairness)? And how do the public's reactions compare to their reactions if a psychological assessment creates the same (disparate) outcomes?

Put another way, the motivating question of this experiment is whether the public responds differently to disparate impact based on whether its source is an algorithm or a human decision-maker.

Design

The structure of Experiment 3 is largely identical to Experiment 1 and Experiment 2, with two main exceptions. First, instead of varying the procedure between Algorithms, Psychologists, and Guidelines, respondents are randomly assigned to either evaluate an Algorithm or Psychologist only. Second, instead of varying accuracy levels, the experiment varies the level of disparate impact.

[14] This strategy avoids but does not *resolve* the classification concern. As Huq discusses in some depth, a rule requiring algorithm developers to avoid such covariates "would be unmoored from the justifications for an anticlassificatory rule, at odds with the purported justifications . . . and would engender results that contradict the assumed purposes of the rule" (pg. 33).

Specifically, respondents were shown the following text:

> Your city will use a `[ALGORITHM|PSYCHOLOGIST]`. Past data shows that, if released, `[GROUP 1]` are 10% more likely to commit a crime than `[GROUP 2]`. However, academic studies of this procedure show that `[GROUP 1]` are `[20|30]`% more likely to be kept in jail.[15]

This text contains three separate and independent randomizations: (1) the source of the disparate impact was either attributed to algorithms or psychologists; (2) the groups were either randomized to be about race (comparing White and Black Americans) or class (comparing low-income and middle- or high-income Americans); (3) the level of disparate impact was either "high" (30%) or "low" (20%). In terms of the different metrics for "algorithmic fairness," this experiment manipulates disparate levels of false positives (predicting someone is high-risk and thus deserving of jail when they are not). Given that the literature has not settled on a standard metric (or even set of metrics) to assess disparate impact, selecting any metric is somewhat arbitrary, and this metric at least seemed relatively straightforward to understand.

This experiment has 3,453 respondents, drawn from the same source as Experiment 1 and Experiment 2.

Results

Respondents seem to react differently to disparate impact based on whether its source is an algorithm or a psychologist. To measure this effect, we estimate the regression

$$Y = \text{Procedure} + \text{Disparate Impact} + \text{Procedure} * \text{Disparate Impact} + \text{Covariates}$$

where Y is the respondent's rating after viewing a single procedure, Procedure is whether the procedure is an algorithm or psychologist, and Disparate Impact refers to the effect of moving from a low level of disparate impact to a higher level of disparate impact. The variable of interest is the coefficient on the interaction term (Procedure $*$ Disparate Impact): a significant interaction term indicates that the effect of moving from a low to a higher level of disparate impact changes based on whether the procedure is an algorithm or a psychologist.

As Table 6 shows, the effect of disparate impact differs by whether its source is an algorithm or a psychologist. The interaction term is significant at the $p < 0.05$ level when evaluating whether individuals want to use an algorithm. The interaction term is not significant at $p < 0.1$ in terms of individuals' assessment of the fairness of the procedure. But the point estimate is in the same direction as the model based on individuals' desire to use the procedure.

[15] The full question text is available in the Appendix.

	Fairness Rating	Want to Use
Procedure (Psychologist)	0.25^{***}	0.38^{***}
	(0.06)	(0.06)
Disparate Impact	0.03	0.04
	(0.06)	(0.06)
Procedure * Disparate Impact	-0.13	-0.17^{**}
	(0.08)	(0.08)
Observations	3,453	3,453

Note: *p<0.1; **p<0.05; ***p<0.01
Covariate controls and intercept omitted.

Table 6: Respondents React Differently to the Presence of Disparate Impact Based on Source

	Procedure	Effect	Standard Error	Test Statistic	P-Value
Fairness	Algorithm	0.03	0.06	0.47	0.64
	Psychologist	-0.10	0.06	-1.69	0.09
Want to Use	Algorithm	0.04	0.06	0.69	0.49
	Psychologist	-0.13	0.06	-2.11	0.04

Table 7: Increases in Disparate Outcomes Decrease Support for Psychologists, But Have No Effect on Support for Algorithms

The marginal effects of these models, which allow us to more clearly appreciate the distinct reactions towards algorithms versus psychologists, are displayed in Table 7. Respondents react negatively when psychologists exhibit a higher level of disparate impact. For example, when assessing whether they want to use a particular procedure, respondents rate psychologists with a higher disparate impact 0.13 points lower on a -2 to $+2$ scale than a psychologist with a lower level of disparate impact. However, respondents are indifferent along both fairness and policy preference when algorithms exhibit higher levels of disparate impact.[16]

It is worth noting that the effect of disparate impact is relatively small in this experiment.[17] These small results could demonstrate that the public is indifferent to disparate impact in all its forms, whether from a psychologist or an algorithm. An alternate explanation is that disparate impact is a difficult concept to explain, and the treatment effects may have been

[16] Careful readers may note that, in both models, the effect of disparate impact when an algorithm is used is not significant while the effect *is* significant when a psychologist is used. However, the difference between statistically significant and not statistically significant may not itself be statistically significant (Gelman and Stern, 2006). But sometimes it is, and the correct way to adjudicate is to examine interaction term displayed in Table 6. Table 7's marginal effects are displayed solely for interpretability.

[17] The overall effect of having a higher disparate impact, aggregated across algorithms and psychologists, is about 0.05 points on a scale from -2 to $+2$ for fairness and 0.06 points on the same scale for policy preference.

attenuated due to the complexity of the treatment. Both explanations are compatible with the results of the experiment, and given the statistical (if not substantive) significance of the results, further study is warranted.

Discussion and Conclusion

Algorithms are rapidly becoming a ubiquitous feature of the criminal justice system. Much of the literature on the use of algorithms have focused on their substantive "fairness," as assessed by differential accuracy rate. Without commenting on the importance of such metrics, this project addresses a second, unexamined issue: whether the public views such algorithms as fair.

The results are not promising. Before policymakers rush to implement algorithms, they should note that the public views the use of algorithms as unfair, undesirable, and illegitimate. This dislike of algorithms persists even when members of the public are told that algorithms are transparent and have the same level of accuracy or disparate impact as other procedures.

But members of the public are not *indifferent* to these other factors. When algorithms are more accurate than other procedures, public support for algorithms increases. This relationship suggests that people are balancing two factors: an intrinsic distrust of algorithms and a desire to use the most accurate procedure. Absent any information on accuracy, these beliefs are not in conflict because people both dislike using algorithms and view them as inaccurate. If they are told that algorithms are more accurate than other procedures, they balance their desire for the most accurate procedure against their dislike of algorithms. The result is a preference for algorithms when they are the more accurate procedure and only when they are the more accurate procedure. To the extent that algorithms are or become more accurate than existing procedures,[18] policymakers should address and refute public beliefs. If this is done, algorithms become more acceptable.

Similarly, members of the public are more likely to accept algorithms when they are transparent. When individuals are told that algorithm creators divulge what goes into an algorithm and the statistical techniques they apply, they are more likely to favor algorithms as a matter of fairness and policy. But even when algorithms are transparent, members of the public still do not want to use them. They merely prefer transparent to non-transparent algorithms.

Algorithms have a mixed effect in the case of disparate impact. Disparate impact by a human decision-maker undermines perceptions of fairness. Disparate impact by an impersonal machine does not. This accords with the general suggestion that people punish individuals who exhibit disparate impact, likely because they view such persons as having "bad" motives. But when they are presented with a machine who spits out the same outcomes, they are less likely to view it as prejudiced and therefore do not view such outcomes as unfair. This might be good news for algorithm proponents: while algorithms might be upsetting to the public for any number of reasons, disparate impact does not appear to be one of them. On the

[18] Whether algorithms are more accurate than human judgment is very much a contested proposition Dressel and Farid (2018).

other hand, this result should only increase the concern of scholars of “algorithmic fairness.” They must not only define “fairness” but also convince the public that it matters at all.

But these results don’t imply that the public uniformly prefers human decision-making. Indeed, the results indicate that members of the public are as comfortable with statutory guidelines as they are with expert psychologists. As a policy issue, this might imply that public officials need to educate the public more about what algorithms are, how they work, and who regulates them. By failing to consider public opinion around algorithms, scholars and policymakers alike are missing the point. Even when algorithms are as accurate as other procedures, even when they have the same level of disparate impact, and even when they are transparent—Americans still dislike algorithms, and they don’t want to use them in the bail context.

Thus, while we should not lose sight of the substantive fairness of algorithms, scholars must also investigate and mitigate their perceived (un)fairness. In particular, we should assess whether the use of algorithms, like any other criminal justice procedure, meets the familiar principles of procedural justice: the opportunity for participation, a neutral forum, trustworthy authorities, and treatment with dignity and respect. Otherwise, even if we fix the set of problems that scholars and policymakers have identified, the use of algorithms will be unpopular and may even damage the legitimacy of the judicial system.

References

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *arXiv:1703.09207 [stat]*, Mar. 2017. URL http://arxiv.org/abs/1703.09207. arXiv: 1703.09207.

R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. In *ACM Conference on Human Factors in Computing Systems*, 2018. URL http://arxiv.org/abs/1801.10408. arXiv: 1801.10408.

A. Christin. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, July 2017. doi: 10.1177/2053951717718855. URL http://journals.sagepub.com/eprint/SPgDYyisV8mAJn4fm7Xi/full.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of KDD'17*, 2017. URL http://arxiv.org/abs/1701.08230. arXiv: 1701.08230.

B. Cowgill and C. Tucker. Algorithmic Bias: A Counterfactual Perspective. Working Paper: NSF Trustworthy Algorithms, 2017. URL http://trustworthy-algorithms.org/whitepapers/Bo%20Cowgill.pdf.

B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*, 144(1):114–126, 2015. ISSN 1939-2222, 0096-3445. doi: 10.1037/xge0000033. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000033.

J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, Jan. 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aao5580. URL http://advances.sciencemag.org/content/4/1/eaao5580.

Electronic Privacy Information Center. Algorithms in the Criminal Justice System. Technical report, 2018. URL https://epic.org/algorithmic-transparency/crim-justice/. [Accessed May 2018].

A. Feller, E. Pierson, S. Corbett-Davies, and S. Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post*, Oct. 2016. URL https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

A. Gelman and H. Stern. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, 60(4):328–331, Nov. 2006. doi: 10.1198/000313006X152649. URL http://www.tandfonline.com/doi/abs/10.1198/000313006X152649.

B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50, Oct. 2017. doi: 10.1609/aimag.v38i3.2741. URL http://arxiv.org/abs/1606.08813. arXiv: 1606.08813.

U. Gromping. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, 63(4):308–319, Nov. 2009. ISSN 0003-1305. doi: 10.1198/tast.2009.08199. URL https://amstat.tandfonline.com/doi/abs/10.1198/tast.2009.08199.

M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. Wiley-Interscience, New York, 2 edition edition, Jan. 1999. ISBN 978-0-471-19045-5.

A. Z. Huq. Racial Equity in Algorithmic Criminal Justice. *Duke Law Journal*, 68, 2019. URL https://papers.ssrn.com/abstract=3144831.

D. Kehl, P. Guo, and S. Kessler. Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Technical report, Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School, 2017. URL http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of Innovations in Theoretical Computer Science*, 2017. URL http://arxiv.org/abs/1609.05807. arXiv: 1609.05807.

J. Logg, J. Minson, and D. A. Moore. Algorithm Appreciation: People Prefer Algorithmic To Human Judgment. SSRN Scholarly Paper ID 2941774, Social Science Research Network, Rochester, NY, Apr. 2018. URL https://papers.ssrn.com/abstract=2941774.

R. J. MacCoun and T. R. Tyler. The Basis of Citizens' Perceptions of the Criminal Jury: Procedural Fairness, Accuracy, and Efficiency. *Law and Human Behavior*, 12(3):333–352, 1988. ISSN 0147-7307. URL http://www.jstor.org/stable/1393683.

T. Mathiesen. Selective Incapacitation Revisited. *Law and Human Behavior*, 22(4):455–469, 1998. ISSN 0147-7307. URL http://www.jstor.org/stable/1394595.

L. W. Miratrix, J. S. Sekhon, A. G. Theodoridis, and L. F. Campos. Worth Weighting? How to Think About and Use Weights in Survey Experiments. *arXiv:1703.06808 [stat]*, Mar. 2017. URL http://arxiv.org/abs/1703.06808. arXiv: 1703.06808.

ODonnell v. Harris County. 251 F.Supp.3d 1052 (S.D.Tex. 2017).

S. Picard-Fritsche, M. Rempel, J. A. Tallon, J. Adler, and N. Reyes. Demystifying Risk Assessment: Key Principles and Controversies. Technical report, Center for Court Innovation, 2017. URL https://www.courtinnovation.org/sites/default/files/documents/Monograph_March2017_Demystifying%20Risk%20Assessment_1.pdf.

J. Ramji-Nogales, A. I. Schoenholtz, and P. G. Schrag. Refugee Roulette: Disparities in Asylum Adjudication. *Stanford Law Review*, 60:295, 2007. URL https://heinonline.org/HOL/Page?handle=hein.journals/stflr60&id=303&div=&collection=.

N. Ritter. Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise. *National Institute of Justice*, 271:4–13, 2013. URL https://www.ncjrs.gov/pdffiles1/nij/240696.pdf.

R. Simmons. Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System. *University of California Davis Law Review*, 52, 2018. URL https://papers.ssrn.com/abstract=3156510.

S. B. Starr. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review*, 66:803–872, Sept. 2014. URL https://papers.ssrn.com/abstract=2318940.

State v. Loomis. 881 N.W.2d 749 (Wis. 2016).

T. R. Tyler. *Why People Obey the Law*. Princeton University Press, Mar. 1990.

T. R. Tyler and J. Sevier. How do the Courts Create Popular Legitimacy-The Role of Establishing the Truth, Punishing Justly, and/Or Acting through Just Procedures. *Albany Law Review*, 77:1095, 2013.

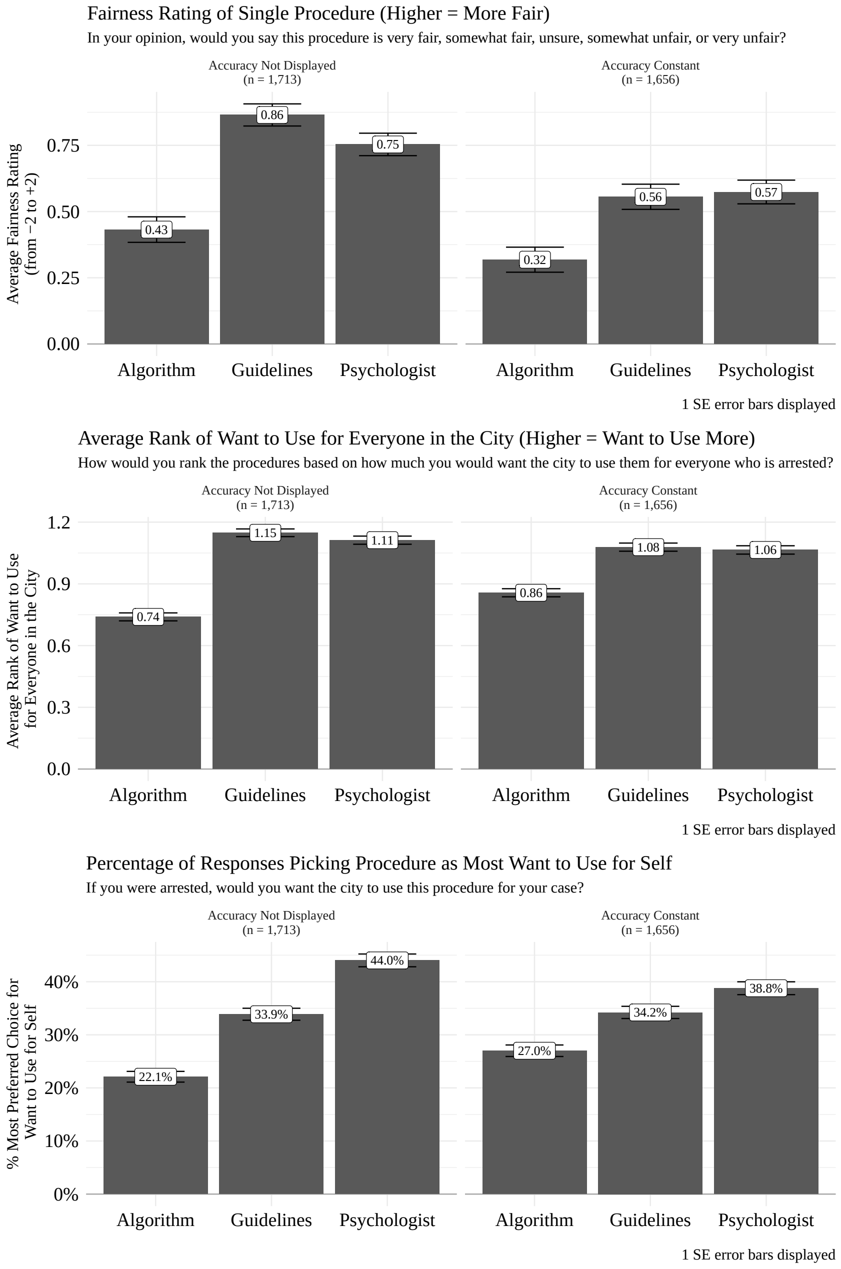
Workgroup. Pretrial detention reform: Recommendations to the chief justice. Technical report, Judicial Branch of California, Oct. 2017. URL https://newsroom.courts.ca.gov/internal_redirect/cms.ipressroom.com.s3.amazonaws.com/262/files/20179/PDRReport-FINAL%2010-23-17.pdf.

Appendix

Additional Tables and Figures

Accuracy of Psychologist	Accuracy of Guidelines	Accuracy of Algorithm	Compared to Psychologist	Compared to Guidelines
Low	Low	Low	Same	Same
High	High	High	Same	Same
Low	Low	High	Algorithm Higher	Algorithm Higher
Low	High	Low	Same	Guidelines Higher
Low	High	High	Algorithm Higher	Same
High	Low	Low	Psychologist Higher	Same
High	Low	High	Same	Algorithm Higher
High	High	Low	Psychologist Higher	Guidelines Higher

Table A1: Mapping Accuracy to Pairwise Comparisons



Fairness Rating of Single Procedure (Higher = More Fair)

In your opinion, would you say this procedure is very fair, somewhat fair, unsure, somewhat unfair, or very unfair?

1 SE error bars displayed

Average Rank of Want to Use for Everyone in the City (Higher = Want to Use More)

How would you rank the procedures based on how much you would want the city to use them for everyone who is arrested?

1 SE error bars displayed

Percentage of Responses Picking Procedure as Most Want to Use for Self

If you were arrested, would you want the city to use this procedure for your case?

1 SE error bars displayed

Figure A1: Ranking order is the same regardless of whether respondents have information about accuracy

Questionnaire

In the United States, when a person is arrested for a crime, they may wait weeks or months for a trial to occur. While the person waits for trial, a judge decides whether to keep the person in jail or release them. A judge must consider whether the person would commit another crime if released. If a judge releases a person, they may require the person to pay a certain amount of money, called **bail**, to ensure that they return for trial. Judges hold **bail hearings** to decide who is released in jail and who is released on bail.

Experiment 1

Suppose that your city needs to decide how to hold bail hearings. One input in this decision is what the public thinks of different options. You'll be shown a series of proposed procedures and asked questions about them.

Randomize to one of the following

`[PSYCHOLOGIST]`. Your city will use a **court-provided psychologist**, who conducts an interview to predict whether someone will commit another crime if released.

The psychologist evaluates the person's offense, criminal history, home conditions, and support network.

Academic studies of this method show that the accuracy rate is `[ACCURACY_LEVEL]`.

`[GUIDELINE]`. Your city will use a set of **mandatory guidelines** developed by the city government to predict whether someone will commit another crime if released.

The guidelines evaluates the person's offense, criminal history, home conditions, and support network.

Academic studies of this method show that the accuracy rate is `[ACCURACY_LEVEL]`.

`[ALGORITHM]`. Your city will use a **statistical algorithm**, a computer program that analyzes a set of offender-related data to predict whether someone will commit another crime if released.

The algorithm evaluates the person's offense, criminal history, home conditions, and support network.

Academic studies of this method show that the accuracy rate is `[ACCURACY_LEVEL]`.

After viewing a single procedure

In your opinion, would you say this procedure is . . . Very Fair, Somewhat Fair, Unsure, Somewhat Unfair, or Very Fair?

If you were arrested, would you want the city to use this procedure for your case? Strongly want to use, Somewhat want to use, Unsure, Somewhat don't want to use, Strongly don't want to use

If your city adopted this procedure, how much would you agree with the statement "**You should accept the decisions made by courts, even if you think they are wrong**." Strongly agree, somewhat agree, Unsure, Somewhat Disagree Strongly Disagree.

Now we're going to ask you to compare the procedure you just saw with two other procedures. **Other two procedures are presented in a random order.**

How would you rank these procedures in terms of accuracy? Drag and drop to change their ranking so that 1 is Most Accurate and 3 is Least Accurate.

How would you rank these procedures in terms of fairness? Drag and drop to change their ranking so that 1 is Most Fair and 3 is Least Fair.

How would you rank the procedures based on whether you would want the city to use them **if you were arrested**? Drag and drop to change their ranking so that 1 is Most Want to Use and 3 is Least Want to Use.

How would you rank the procedures based on how much you would want to the city to use them **for everyone that is arrested**? Drag and drop to change their ranking so that 1 is Most Want to Adopt and 3 is Least Want to Adopt).

Please provide a brief description of why you ranked these choices as you did.

Experiment 2

Suppose that your city has decided to use a statistical algorithm, a computer program that analyzes a set of offender-related data to predict whether someone will commit another crime if released. This algorithm is **developed by a private company**.

Some people argue that the company should publicly describe how the algorithm works so people can correct errors and challenge decisions made by it.

Others argue that companies should be able to keep their algorithms private to protect trade secrets, just like Coca-Cola or McDonald's keep their recipes private.

You will be shown a series of proposed algorithms and asked questions about them.

Randomize to one of the following

`[TRANSPARENT]`. Your city plans to purchase an algorithm from a company that **publicly describes** how its algorithm works. Academic studies of this algorithm show that the accuracy rate is `[ACCURACY_LEVEL]`.

[NON-TRANSPARENT]. Your city plans to purchase an algorithm from a company that **does NOT publicly describe** how its algorithm works. Academic studies of this algorithm show that the accuracy rate is [ACCURACY_LEVEL].

After viewing a single procedure

In your opinion, would you say this procedure is . . . Very Fair, Somewhat Fair, Unsure, Somewhat Unfair, or Very Fair?

If you were arrested, would you want the city to use this procedure for your case? Strongly want to use, Somewhat want to use, Unsure, Somewhat don't want to use, Strongly don't want to use

Now we're going to ask you to compare the procedure you just saw with a new procedure.

Which procedure is more fair?

Which procedure would you want the city to use for your case **if you were arrested**?

Which procedure would you want the city to use for your case **for everyone who is arrested**?

Purchasing the algorithm that is publicly described is **more expensive** than the algorithm that is not publicly described. If the city chooses the algorithm that is NOT publicly described, it will save $3,000,000 over the next five years. Which procedure would you prefer that the city adopt?

Experiment 3

Now you're going to learn about a completely different set of procedures for holding bail hearings.

Your city is concerned that different types of bail hearings might impact groups differently. We're going to ask you some questions about procedures your city might adopt.

Randomize to one of the following

[PSYCHOLOGIST]. Your city will use a court-provided psychologist, who conducts an interview to predict whether someone will commit another crime if released. Past data shows that, if released, [African-Americans/low-income people]. are 10% more likely to commit a crime than [Whites/middle-class or high-income people]. However, academic studies of this procedure show that [African-Americans/low-income people] are [DISPARITY_LEVEL] more likely to be kept in jail.

[ALGORITHM]. Your city will use a statistical algorithm, a computer program that predicts whether someone will commit another crime if released by analyzing a set of offender-related data. Past data shows that, if released, [African-Americans/low-income people]. are 10% more likely to commit a crime than [Whites/middle-class or high-income peo-

`ple]`. However, academic studies of this procedure show that `[African-Americans/low-income people]` are `[DISPARITY_LEVEL]` more likely to be kept in jail.

After viewing a single procedure

In your opinion, would you say this procedure is . . . Very Fair, Somewhat Fair, Unsure, Somewhat Unfair, or Very Fair?

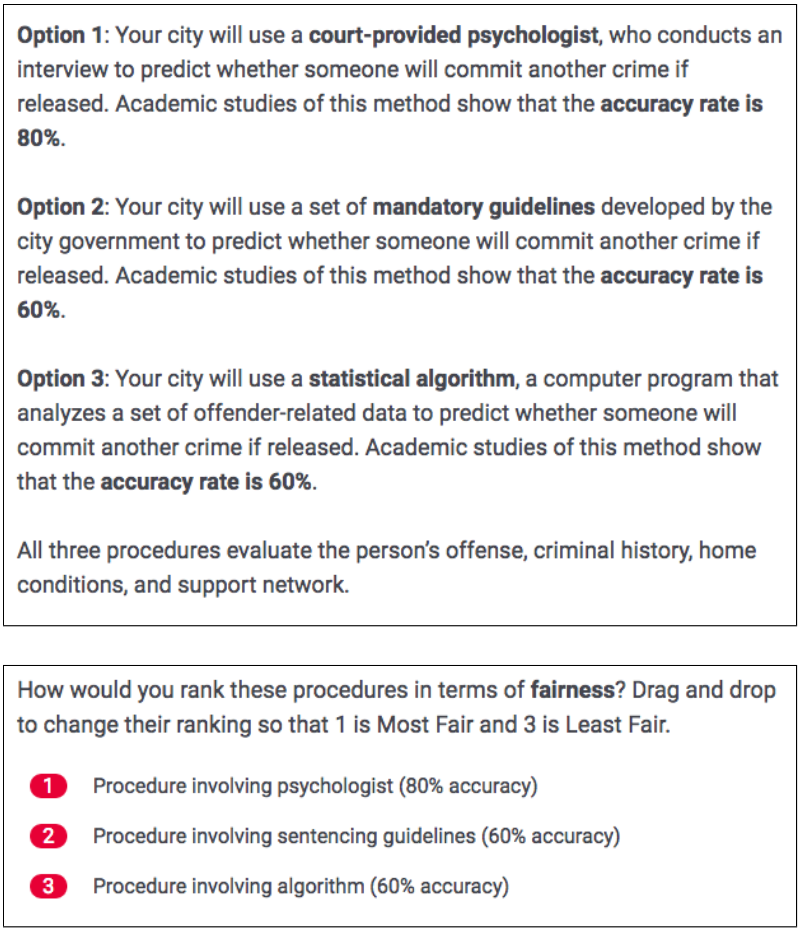
If you were arrested, would you want the city to use this procedure for your case? Strongly want to use, Somewhat want to use, Unsure, Somewhat don't want to use, Strongly don't want to use

Now we're going to ask you to compare the procedure you just saw with another procedure.

Which procedure is more fair?

Which procedure would you want the city to use for your case **if you were arrested**?

Which procedure would you want the city to use for your case **for everyone who is arrested**?



Option 1: Your city will use a **court-provided psychologist**, who conducts an interview to predict whether someone will commit another crime if released. Academic studies of this method show that the **accuracy rate is 80%**.

Option 2: Your city will use a set of **mandatory guidelines** developed by the city government to predict whether someone will commit another crime if released. Academic studies of this method show that the **accuracy rate is 60%**.

Option 3: Your city will use a **statistical algorithm**, a computer program that analyzes a set of offender-related data to predict whether someone will commit another crime if released. Academic studies of this method show that the **accuracy rate is 60%**.

All three procedures evaluate the person's offense, criminal history, home conditions, and support network.

How would you rank these procedures in terms of **fairness**? Drag and drop to change their ranking so that 1 is Most Fair and 3 is Least Fair.

1 Procedure involving psychologist (80% accuracy)

2 Procedure involving sentencing guidelines (60% accuracy)

3 Procedure involving algorithm (60% accuracy)

Figure A2: Screenshot from Experiment 1