Discussion Papers of the
Max Planck Institute for
Research on Collective Goods
2022/4

Fair Governance with Humans and Machines

Yoan Hermstrüwer
Pascal Langenbach

MAX PLANCK
SOCIETY

Discussion Papers of the
Max Planck Institute
for Research on Collective Goods 2022/4

Fair Governance with Humans and Machines

Yoan Hermstrüwer / Pascal Langenbach

May 2022

This version: October 2022

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
https://www.coll.mpg.de

Abstract

How fair do people perceive government decisions based on algorithmic predictions? And to what extent can the government delegate decisions to machines without sacrificing perceived procedural fairness? Using a set of vignettes in the context of predictive policing, school admissions, and refugee-matching, we explore how different degrees of human-machine interaction affect fairness perceptions and procedural preferences. We implement four treatments varying the extent of responsibility delegation to the machine and the degree of human involvement in the decision-making process, ranging from full human discretion, machine-based predictions with high human involvement, machine-based predictions with low human involvement, and fully machine-based decisions. We find that machine-based predictions with high human involvement yield the highest and fully machine-based decisions the lowest fairness scores. Different accuracy assessments can partly explain these differences. Fairness scores follow a similar pattern across contexts, with a negative level effect and lower fairness perceptions of human decisions in the context of predictive policing. Our results shed light on the behavioral foundations of several legal human-in-the-loop rules.

Keywords: algorithms; predictive policing; school admissions; refugee-matching; fairness.

Machine-learning algorithms are increasingly used to predict risks and assist public officials in their decisions. While the initial discussion has focused on the algorithmic assistance of judges in sentencing, pre-trial, or bail decisions (Kleinberg et al., 2018), similar algorithmic decision aids are rapidly expanding to other areas of public decision-making (Engstrom et al., 2020; Huq, 2020a). Some of the most notable applications include the prediction of crime in order to deploy police forces more effectively (Joh, 2016; Simmons, 2018), the matching of refugees with municipalities based on predicted integration success (Acharya et al., 2022; Ahani et al., 2021; Bansak et al., 2018), and the admission of students to schools based on their chances of completing their degree (Kearns & Roth, 2019; Muratov et al., 2017).

With the increasing application of machine-learning algorithms in public-sector decision-making, the fairness of these decision aids becomes increasingly important. Vivid illustrations can be found in the growing number of court cases touching upon fairness problems. In *Houston Federation of Teachers v. Houston Independent School District*, for example, a group of teachers argued that the score of an algorithmic system used to evaluate their performance and terminate their contracts was the source of an unfair deprivation of their constitutional rights (see, Crawford & Schultz, 2019).[1] And following the Wisconsin Supreme Court's verdict in *State v. Loomis*,[2] a group of researchers argued that the COMPAS software used to predict recidivism was no fairer than predictions made by humans, not even than those made by lay people (Dressel & Farid, 2018). Yet, it remains unclear what drives assessments of algorithmic fairness and whether the absence of human judgment in algorithmic procedures affects the sentiment of not being treated fairly.

1 Houston Federation of Teachers v. Houston Independent School District, 251 F Supp 3d 1168 (SD Tex 2017).
2 State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

Perceptions of procedural fairness have been found to shape the perceived legitimacy of government actions and compliance with the law (Tyler, 2006, 2003). Hence, people's subjective evaluations are normatively relevant not only because democratic governance has to be in some way responsive to citizens' procedural preferences, but also because people's fairness perceptions might define the success of algorithmic governance because of their link to legitimacy and compliance (Scurich & Krauss, 2020; Cuéllar & Huq, 2022; Simmons, 2018; see also, Wang, 2018). Fairness perceptions could, therefore, be predictive for the future role of algorithms in public-sector decision-making (Nagtegaal, 2021; Simmons, 2018).

Despite recent trends towards partial or even full algorithmic governance (see, Cuéllar & Huq, 2022; Engstrom et al., 2020), its effects on the perceived fairness of public decision-making are not yet well understood. In particular, current studies do not reflect the growing applications of algorithmic decision-aids in different policy fields. Moreover, the design of human-machine interactions when an algorithm assists a human decision-maker has received comparatively little attention so far.

In this article, we address the perceived fairness of algorithmically assisted decision procedures in the public sector. We study how procedural fairness perceptions vary with the degree of machine involvement in the decision-making process, and whether fairness perceptions systematically differ across different policy contexts. The broad scope of our study including a diverse set of policy areas and different human-algorithmic decision procedures is likely to contribute to a better understanding of what drives the acceptance of algorithmic decisions and how far algorithmic governance can go without sacrificing procedural fairness.

The Psychology of Algorithmic Fairness

Perceived Procedural Fairness in Algorithmic Public Decision-Making

Algorithmic fairness can be conceptualized in different ways. One line of research studies the fairness of algorithmic predictions from a normative perspective. This research ultimately tries to improve algorithmic predictions measured by some normative standard, such as statistical parity, equality of false-positives, equality of false-negatives, or equality of predictive accuracy (see, e.g., Barocas et al., 2021; Berk et al., 2021; Hellman, 2020; Kleinberg et al., 2017; Chouldechova, 2017; Corbett-Davies et al., 2017). In the tradition of fairness research in social psychology, we are concerned with the perceived fairness of algorithmic decisions. A common distinction is made according to the object of fairness judgments, that means, whether they refer to decision outcomes (*distributive fairness*) or to the decision-making procedure (*procedural fairness*) (see, Lind & Tyler, 1988; Walker et al., 1979). While achieving distributive fairness may be an important element of legitimacy, for example by defining a social-welfare function that captures a preference for more equitable outcomes (Rambachan et al., 2020), it is argued that the guarantees of procedural fairness are no less important for the law and define the level of compliance and cooperation with legal authorities (see, Tyler, 2006; Tyler & Jackson, 2014).

Previous findings in procedural justice research suggest that part of the effect of fair procedures on perceived legitimacy is mediated by the fact that fair procedures yield more accurate outcomes (Tyler & Sevier, 2014). Likewise one could expect that more accurate procedures are already perceived as fairer: In the literature on algorithmic fairness, Wang (2018) has shown that accuracy indeed affects fairness judgments. Studying the fairness of specific features used in algorithmic predictions, Grgić-Hlača, Zafar, et al. (2018) find that

fairness ratings increase when it is assumed that a feature enhances the accuracy of the prediction. In the study of Albach and Wright (2021), how a specific feature contributes to the accuracy of the decision emerges as people's main concern when they form their fairness assessment of the use of this feature in an algorithmic decision-making process.[3] The introduction of algorithms into a decision-making procedure might change the perceived accuracy of this procedure. According to objective measures, algorithms regularly outperform human decision-makers in prediction tasks (Meehl, 1954; Grove et al., 2000; Kleinberg et al., 2018). Yet, these objective measures need not be congruent with people's subjective evaluations.

Moreover, as procedural justice theory emphasizes, people are not merely concerned with outcomes, but independently evaluate the fairness of the decision procedures (Tyler, 2006; Lind & Tyler, 1988). In the procedural justice research two components, among others, emerged as especially important for the perceived fairness of a procedure: decision-making, that means the way in which legal authorities come to their decisions (e.g., neutral, aimed at accuracy, transparent), and treatment, that means how legal authorities treat people when they interact with them (dignity, respect) (Tyler & Blader, 2000; see also, Trinkner et al., 2018). Both components might convey information about one's status in the social group (Lind & Tyler 1988; Tyler & Lind, 1992; Tyler & Blader, 2000). However, decision-making and personal treatment could be affected when an algorithm becomes part of the decision process (Simmons, 2018; Wang, 2018). People might potentially perceive algorithmic decision-making as rather neutral and free of personal bias, but also as discriminatory; and the fact that a computer program processes one's case as a mere data point might also affect whether a person

3 For a non-legal setting, Yin et al. (2019) report experimental evidence that the stated accuracy of a machine-learning model may affect self-reported trust in the model.

feels treated with human dignity (Simmons, 2018). Therefore, it seems plausible that fairness evaluations could change when algorithms enter public decision-making processes. Knowing how the public's perceptions of fairness and accuracy are affected by the role algorithms play in public decision-making may ultimately help policy-makers in designing algorithmically enhanced decision-procedures more effectively and in accordance with people's preferences (Scurich & Krauss, 2020).

Recognizing the behavioral dimension of fairness perceptions, a growing literature has turned its attention to the perceived fairness of algorithmic decision procedures (for a summary, see, Starke et al., 2021). On the one hand, empirical evidence in the legal domain suggests a preference for human decision-making processes.[4] B. M. Chen, et al. (n.d.) report evidence from a vignette study – with three scenarios covering a consumer refund, a pre-trial bail decision, and a custodial sentencing decision – and show that a human judge is perceived as fairer than an algorithmic judge. Similarly, Yalcin et al. (2022) find higher trust in a human judge than in an algorithmic judge in divorce cases in two vignette experiments. Focusing on decisions in the criminal-justice context, and using a representative sample of the US population, Wang (2018) reports in several vignette studies that the use of a computer algorithm in bail decisions is disliked compared to other expert procedures, with fairness perceptions being affected by information about the accuracy of the procedure. Yet, people's dislike for algorithms in bail decisions seems not to depend just on the inaccuracy of such decisions, but also on the distribution of false-positive rates across groups (Harrison et al., 2020). Within a sample of 600 participants, Simmons (2018), however, reports no differences in fairness

4 This strand of literature is in line with more general evidence showing that people prefer human over algorithmic decisions (see, e.g., Lee & Baykal, 2017; Lee et al., 2019) and that humans tend to distrust algorithmic outputs, a phenomenon sometimes referred to as *algorithm aversion* (Dietvorst et al., 2015).

perceptions between bail decisions made by a judge with or without the assistance of a "computer program".

On the other hand, several studies also show that people assess automated decision-making as fairer than the human alternative.[5] Araujo et al. (2020), for example, report similar fairness ratings for algorithmic and human decisions across different contexts. Yet, when the consequences of decisions are severe, people in their study judge algorithmic decision-making as fairer (for example, the administrative decision whether to issue a fine for wrong parking versus the prosecutorial decision to bring criminal charges). In an experiment on policing by Miller and Keiser (2021), black participants prefer traffic control by automated red-light cameras to a police officer when shown a picture that suggests an underrepresentation of black citizens in the municipal police department. In a survey study, Marcinkowski et al. (2020) find that students rate university admissions decisions made by an algorithm as fairer, and the procedure as less biased, compared to a human admissions committee. Studying fairness perceptions of public employees, Nagtegaal (2021) reports that human decision-making is perceived as fairer than fully algorithmic decision-making for more complex tasks that cannot easily be quantified, whereas the ranking was the other way around for simpler tasks. Descriptively, a combination of a human and an algorithm was in the middle but not statistically different from human decision-making.

Decision Context

In light of these inconclusive results, further empirical investigations of the procedural fairness of algorithmic legal decision-making are inherently valuable. One key insight of the

5 This strand of literature is in line with evidence showing that humans tend to appreciate the use of algorithms in specific commercial contexts, a phenomenon sometimes dubbed *algorithm appreciation* (Logg et al., 2019).

previous literature, however, is that fairness perceptions seem to be context-dependent (Starke et al., 2021). Yet, the overwhelming majority of the studies in the legal domain focus on the criminal-justice system (see, e.g., Imai et al., 2021; Harrison et al., 2020; Dodge et al., 2019; Grgić-Hlača, Redmiles, et al., 2018; Grgić-Hlača, Zafar, 2018; Simmons, 2018; Wang, 2018). While algorithmically assisted decision-making has indeed been very prominent in the context of criminal justice, it is difficult to extrapolate the results to other domains. Only a few studies have extended this relatively narrow contextual scope, exploring fairness perceptions in the context of university admissions (Marcinkowski et al., 2020), parking offenses and criminal charges (Araujo et al., 2020), child protective services and unemployment aid (Albach & Wright, 2021),[6] the enforcement of traffic laws (Miller & Keiser, 2021), and divorce cases (Yalcin et al., 2022). Our study is designed to generate additional evidence that is robust across different areas of the law. With our set of vignettes covering predictive policing, school admissions, and refugee-matching, we can compare the fairness of the different algorithmic decision aids in three practically relevant public-law contexts.

Human-Machine Interaction

Finally, we add to a newer strand of research that explicitly focuses on the interaction between algorithms and human decision-makers (Imai et al., 2021; Green & Y. Chen, 2019a, 2019b). Considering the risks of discrimination, in-group bias, or automation bias in algorithmic decision-making, legal scholars have been discussing whether and to what extent the law actually grants a right to a human decision (Williams, 2022; Huq, 2020b). Computer scientists have also voiced claims in favor of human-in-the-loop, human-on-the-loop, or

6 Albach and Wright (2021) additionally investigate the fairness of specific features in the context of bail, hospital resources, insurance rates, and loans.

human-in-command requirements (Binns, 2022; Yaghini et al., 2021). This corresponds to the basic model of Art. 22 (1) EU General Data Protection Regulation, formulating the principle that no person shall be subject to a decision based on fully automated data processing.[7] Under Art. 14 (1) of the proposed EU Artificial Intelligence Act (AI Act), high-risk AI systems, such as predictive schooling systems like the one we explore in this study, shall be designed and developed in such a way that they can be effectively overseen and fully understood by humans. Others have been more optimistic about the future of purely machine-made decisions and have argued that the outputs generated by machine-learning algorithms should be used as micro-directives (Casey & Niblett, 2017).

Most current algorithmic decision-making practices are based on the premise that decisions cannot or should not be entirely delegated to a machine. Rather, there is some interaction between a human decision-maker and an algorithmic decision aid. While the recent literature has included these hybrid decisions as a third category in the spectrum spanning fully human and fully algorithmic decisions (Nagtegaal, 2021), only relatively little attention has been paid to the effects of different degrees of human control when humans and algorithms work together. Therefore, in addition to comparing the perceived fairness of human and algorithmic decision procedures, we explore procedures in which human decision-makers are assisted by algorithmic decision aids and exert different levels of control over the final outcome.

Overview of the Study

We report results from an online vignette experiment covering three areas of public-sector decisions: predictive policing, school admissions, and refugee-matching. Treatments

7 Art. 22 (3) GDPR contains several exceptions to this principle. This indicates that the material scope of the right to a human decision may be context-dependent rather than universal, as several use cases will likely be exempted from the right.

differ in whether (i) a human, (ii) an algorithm, or (iii) a human assisted by an algorithm makes the decision. The latter case is split into two treatments: one in which the algorithm's assessment of the facts only provides additional information for the human decision-maker, and one in which the human more often than not just relies on the algorithm's assessment, hence practically delegating the decision to the machine in most of the cases. Without knowing the outcomes of the procedure for any particular case, participants judge the fairness of the procedure they are presented with.

Previous research suggests that whether an algorithm is part of a decision-making procedure affects participants' fairness perceptions. While some of these results seem to be context and task dependent, many studies in the realm of decision-making in (criminal) law find a preference for human decision-makers. This preference should lead to an increase of fairness ratings with the degree of human involvement in the process. However, even if people generally prefer human decision-making, they might also prefer a procedure that processes as much information as possible without sacrificing human control. Therefore, human-machine interactions – at least with high human involvement – might also be judged as fairer than procedures in which only humans or algorithms decide. We test our hypotheses in different contexts of public decision-making. We do not have clear predictions on how context and procedure might interact, but we know from the literature that the perceived relative fairness of machine-based and human decision procedures might change with context.

Overall, our results suggest a prevalence of strong fairness preferences for hybrid machine-human decision procedures with a high degree of human involvement. However, people do not seem to care much whether a *human does all* the work or whether a *machine does most* of the work. These results provide important guidance for the interpretation and for the

design of legal rules aimed at organizing the division of labor between humans and algorithms in the public sector.

Method

In this section, we report the design of our study as well as the experimental procedures, including all data exclusions, all manipulations, and all measures in the study.

Design

Treatments. Our main research question concerns the fairness perceptions of different forms of algorithmic assistance in public-sector decision-making. We explore these differences based on between-subject treatment comparisons. In a between-subjects design, we study four treatments that differ in the extent to which the decision is based on algorithmic assistance. This design choice is motivated by the observation that machine assisted decision procedures vary in the level of automation (see, Cummings, 2004; Manzey et al., 2012). Moreover, many human decision procedures cover algorithm-based executions that the human needs to *approve*, while many algorithmic decision procedures allow humans to *veto* an otherwise automatic execution.

In our *HUMAN* treatment, the decision is entirely made by a human decision-maker and solely based on a human assessment of the facts. Participants therefore read that a human decision-maker will *conduct an in-depth analysis of the case material* and *assess the risk* or the *success probability*. Participants also read that the human decision-maker *has discretion* in making the decision. On the other end of the spectrum, in the *MACHINE* treatment, the decision is entirely controlled by a computer algorithm. Participants read that a computer algorithm will conduct the *in-depth analysis of the case material* and *assess the risk* or the *success probability*.

Further, the computer algorithm will make the final decision that no human decision maker can override.

Between those extremes, we implement two treatments with algorithmically assisted decision-making. In both treatments, a human who has discretion in making the decision has the final say. Yet, the degree of algorithmic assistance and the level of human involvement and control – high or low – differs between treatments. In the *HIGH* treatment, the computer's assessment of the facts and the resulting probabilities are *always* accompanied by a human assessment. Participants therefore read that *the decision will never be based on the computer algorithm alone*, but that the human decision-maker *will always conduct his or her own analysis* before making the final decision. In the *LOW* treatment, by contrast, the human input in the decision-making process is heavily reduced as *the decision will usually be based on the computer algorithm alone*. The human decision-maker will only *sometimes conduct his or her own analysis*, meaning that the human decision-maker *will in some cases conduct an in-depth analysis of the case material, and assess the risk/success probabilities*. An example of the wording in the different treatments is shown in Figure 1.

The descriptions of the computer algorithm and of the human assessment are identical across all treatments (when applicable). While our vignettes contain a precise description of the facts that the algorithm and the human decision-maker use to make their assessments and how these facts are elicited, by design, we keep the mechanics of the algorithm vague. Given that we are interested in the fairness evaluations of lay people, we deem it externally valid to give no further information about the technical details of the algorithm, since the public will most likely not have more detailed knowledge about how a computer algorithm assisting a government official produces its results.

Scenarios. In order to enhance the robustness of our findings across different practically relevant areas of the law, we implement each of the four treatments in three different scenarios. In this within-subjects component of our experiment, participants in a session respond to one treatment presented in three different legal contexts: a predictive-policing scenario, a school-admissions scenario, and a refugee-matching scenario. Hence, for a given treatment, each participant reads all three scenarios. Scenarios are presented in randomized order.

Apart from representing different policy contexts, the three scenarios, of course, also diverge in other regards. First, the task of the computer algorithm and the goal of the human assessment slightly differ across the scenarios. In the predictive-policing scenario, it is the risk of violent crimes in specific areas of the city that needs to be predicted. In the school-admissions scenario, the probability of graduation is assessed, whereas in the refugee scenario the probability of employment for a refugee in a certain location is of interest. Second, in the predictive-policing and the refugee scenario, a single human decision-maker, either a police or an immigration officer, is in charge. In the school scenario, a school admissions board manages the application procedures and decisions. Third, while the tasks used in all our vignettes are not purely mechanical and therefore not easily quantifiable, their level of complexity differs to some extent. Predicting crime may be simpler than predicting the probability of employment of refugees, as the latter is likely to depend on individual characteristics as well as fluctuations in supply and demand in labor markets. Predicting the employment of refugees may in turn be simpler than predicting success at school, as this depends on individual characteristics and the evolution of skills over a long period of time. Task complexity might also affect the relative evaluation of human or algorithmic decision procedures (Nagtegaal, 2021; Yalcin et al., 2022).

Scenario: *Police*			
One of the main tasks of the police is to prevent criminal behavior. In order to deploy their forces in an optimal manner, the police need to assess the risk that criminal behavior will occur. This risk assessment refers to various types of criminal behavior, including the risk of violent assaults. Suppose the local police want to assess the risk of violent assaults in certain areas of the city - including the probable type, location, and time of the assault - and perform bodily searches of all persons within a small and well-defined area of the city. The purpose of these bodily searches is to track down weapons used for violent assaults.			
Treatment: Human	**Treatment: Machine**	**Treatment: High**	**Treatment: Low**
The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.	The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer. The decision **will be based** on the computer algorithm's assessment **alone**.	The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer. The decision **will never be based** on the computer algorithm **alone**. The police officer will **always** conduct his or her own analysis, that means, the police officer will in each case conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.	The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer. The decision will **usually be based** on the computer algorithm **alone**. The police officer will **sometimes** conduct his or her own analysis, that means, the police officer will in some cases conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.
Based on his or her risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.	Based on its risk assessment, the computer algorithm will order or not order bodily searches in a certain area of the city. The police officer cannot override the decision of the computer algorithm and has no discretion in this decision.	Based on the risk assessment of the computer algorithm and his or her own risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.	Based on the risk assessment of the computer algorithm and - only if conducted - his or her own risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.

Figure 1: Vignettes for the Four Different Treatments in the Police Scenario. [8]

8 Wording for the treatments in the other two scenarios can be found in the Instructions in Appendix A.

Measures

After each vignette describing one of the scenarios, participants answered four questions. First, we asked participants to indicate the fairness of the procedure by which the decision-maker would come to her decision. Participants could choose one of seven possible answers ranging from *very unfair* (1) to *very fair* (7). Second, as a control variable, we asked participants for their accuracy assessment of the probability estimate on which the decision-maker would base her decision. Different evaluations of the accuracy of a procedure might help explain differences in fairness ratings (Wang, 2018). Participants could choose one of seven possible answers, ranging from *not accurate at all* (1) to *extremely accurate* (7).

Additionally, we elicited responses to two questions designed to identify whether the participants' fairness assessments differ between situations in which they are personally involved or not involved (see, Wang, 2018). Therefore, we asked participants whether they would want the decision-making procedure to be implemented in case they were personally affected by the decision. Finally, we asked whether they would want the procedure to be used for the public. In both cases, participants could choose one of seven possible answers, ranging from *not at all* (1) to *to a large extent* (7).

To control for socio-demographic characteristics, after the last vignette, we also collected individual-level covariates, including age, education, gender, ethnicity, political affiliation, and the weekly hours spent on remunerated tasks like those offered on MTurk.[9]

9 In the questionnaire, participants were additionally asked to rank the three scenarios according to their severity and had the option to give a short written explanation for their responses in the experiment.

Participants

Our sample consists of 1,598 participants, recruited from the MTurk marketplace in the US, as all algorithmic decision support systems we explore in our study have been either developed or predominantly applied in the US or Switzerland to this date (for the matching of refugees see, Acharya et al., 2022; Ahani et al., 2021; Bansak et al., 2018). Of our participants 6% stated an age below 25 years, 46% between 25 and 34 years, 24% between 35 and 44, 14% between 45 and 54, 7% between 55 and 64, and 2% above 65 years. With 63%, men are over-represented in our sample. 67% of our sample identify as *White*, 26% as *Black or African American*, and 5% as *Asian*. Roughly, 60% report a *four-year college degree* as their highest education, and over 18% report a *professional degree*.

Procedure

Our study was programmed in Qualtrics and deployed through CloudResearch to ensure a reliable recruitment of participants. Aware of the challenges posed by MTurk (see, Horton et al., 2011), we implemented measures to enhance the validity of our results. To mitigate further potential self-selection problems, we ran the study in different sessions on different days and at different times of day to ensure a diverse composition of the participant pool. In order to motivate participants to engage seriously with the vignettes, we made sure to keep our vignettes short and paid a competitive participation fee. On average, participants spent approximately 8 minutes on the vignettes and earned 1.50 USD after completing the study. In addition, we implemented an attention check before participants began reading the vignettes. Failure to pass the attention check resulted in immediate exclusion from the experiment without any payment and without recording further input from the participants. We opted for only one, but rigorously applied attention check before subjects read our vignettes in order to strike a balance between

the interest of screening out inattentive subjects and fairness concerns towards the experimental participants. Specifically, because the experiment is rather short a second attention check (and the possible exclusion without pay) later in the survey raises the concern that participants then already completed a considerable part of their task. We also imposed a time constraint of 45 minutes to exclude participants who left the task for a considerable amount of time and to allow that new participants are admitted to the task. Only participants who completed the experiment entered our final sample.[10]

Results

Our main research question pertains to the effects of different forms of algorithmic assistance in public-sector decision-making. These results are captured by the between-subjects treatment differences in our experiment. We begin by reporting analyses of these treatment differences on the pooled data over all scenarios (*HUMAN*: $n = 401$, *HIGH*: $n = 397$, *LOW*: $n = 401$, *MACHINE*: $n = 399$). These analyses also include discussions of the relationship between the perceived accuracy of the different procedures and procedural fairness and the role of socio-demographic characteristics. We then delve deeper into the context-specific effects of the four treatments in the three different scenarios. In all our analyses, we focus on fairness ratings of the different decision-making procedures.[11]

10 In total 2035 participants started the experiment. 2028 participants consented to take part in the experiment (3 did not consent and 4 abandoned the survey). Out of these, 1657 participants passed the attention check (363 failed and 8 abandoned the survey before). Until this point no decision data was elicited. From the remaining sample, we excluded anyone who did not finish the experiment (so abandoned the experiment at any time; these were additional 59 participants), which gives us our final sample of 1598 participants.

11 Procedural preferences do not seem to differ between cases with personal involvement and cases applied to the general population. We relegate the summary analysis of our results on procedural preferences regarding the involvement of oneself or others to Appendix B.

Overall Treatment Effects

According to our experimental design, each participant responds to the same treatment (in a different scenario) at three points in time. We observe that the participants' first response differs from the other two responses (average fairness ratings over all treatments at position $1/2/3 = 5.05/4.85/4.78$). However, these differences in fairness ratings seem to be mere level effects resulting from the timing of the response. There appears to be no systematic difference between responses at different points in time related to the treatments.[12] Treatment-specific order effects being absent, we run our analyses at the group level on the data pooled from all responses across time. Note that we present both the results of conservative estimates obtained from non-parametric Mann-Whitney U (MWU) tests and the results obtained from (random-effects) linear regression models.

Treatment Effects. As can be seen in Figure 2, fairness ratings are highest in the *HIGH* treatment ($M = 5.20$) with a human-computer interaction and high human control over the decision-making procedure. By contrast, participants judge the *MACHINE* treatment ($M = 4.64$) as the least fair. Participants relatively dislike when human decision-makers totally relinquish decision control. The *HUMAN* ($M = 4.89$) and the *LOW* treatment ($M = 4.84$) with human-computer interaction and low human control are in between. In sum, however, fairness ratings are relatively high in all treatments. More specifically, they are above the midpoint of the scale in all treatments, which suggests that all decision-making procedures seem to be acceptable in terms of procedural fairness.

12 We refer to Appendix C for analyses of potential differential effects of the point in time of the response according to treatment. We only find a marginally significant difference between the effects of the position of the response in the *HIGH* treatment compared to the *MACHINE* treatment.

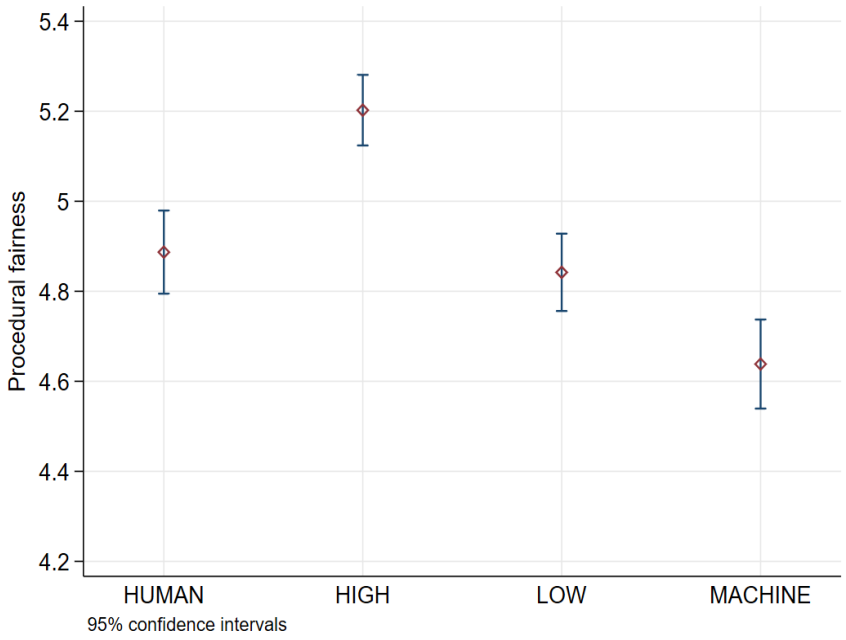

Figure 2: Procedural Fairness Across All Scenarios

Overall, treatment differences are statistically significant. Fairness ratings in the *HIGH* treatment are significantly higher than fairness ratings in all other treatments ($ps < .001$, MWU, *HUMAN*: $r = .086$; *LOW*: $r = .121$; *MACHINE*: $r = .155$). Participants seem to value the extensive human involvement in the decision-making process. Consequently, the purely algorithmic decision procedure in the *MACHINE* treatment yields significantly lower fairness ratings than all other treatments (*HUMAN*: $p < .001$, MWU, $r = .068$, *LOW*: $p = .031$, MWU $r = .044$). The difference in fairness ratings between the *HUMAN* and the *LOW* treatment, however, does not reach statistical significance ($p = .168$, MWU). This might support the interpretation that people accept a certain delegation of decisions to an algorithmic decision aid. Even a procedure in which the human decision-maker regularly just follows the machine advice leads to similar fairness ratings than a purely human decision procedure.

In order to further determine the robustness of our results, we perform a two-sample sensitivity analysis (see, Faul et al., 2007; Perugini et al., 2018), investigating the sensitivity of

the effect size for fairness rating means in response to variations in the power levels (between $1-\beta=0.6$ and $1-\beta=0.95$) and in the significance levels (between $\alpha=0.1$ and $\alpha=0.001$). We calculate the minimum detectable change in the mean of the fairness ratings between a treatment and a control group ($\delta=M_2-M_1$). For the sake of simplicity, we use the approximate mean of the MACHINE treatment of $M_1=4.60$ as the mean of the control group, and assume an equal standard deviation of $\sigma=1.55$ across samples, and a sample size of $n=400$. The results of this sensitivity analysis show that the minimum detectable change in the fairness rating mean ranges between $\delta=4.87-M_1=0.27$ and $\delta=5.05-M_1=0.45$ at the conventional power level of $1-\beta=0.8$ and significance levels ranging between $\alpha=0.1$ and $\alpha=0.001$ (Figure D1 in Appendix D). Even at the highest power level of $1-\beta=0.95$, the minimum detectable change in the fairness rating mean never increases above the $\delta=5.143-M_1=0.54$ threshold. This analysis corroborates the robustness of our results and supports the conclusion that the differences we observe are not random and not a mere artifact of specific parameter values.

Finally, generalized least squares random-effects regression models confirm our fairness results. All model specifications are displayed in Table 1. In Model 1, we regress fairness ratings on treatment dummies and dummies for the decision point in time. We control for the different scenarios and for socio-demographic characteristics elicited in the post-experimental survey in Model 2. To be specific, we include dummy variables for the scenarios, as well as participants' gender, ethnicity, political preferences, age, and education. We also include the self-reported amount of time the participants in our sample spend on paid online tasks. To explore possible explanations for our treatment effects, we add the participants' accuracy ratings to the regression estimation in Model 3.

Baseline: HUMAN DV: Procedural fairness	(1)	(2)	(3)
HIGH	0.315***	0.254***	0.107*
	(0.095)	(0.085)	(0.057)
LOW	-0.045	-0.047	-0.105*
	(0.094)	(0.085)	(0.057)
MACHINE	-0.249***	-0.249***	-0.210***
	(0.095)	(0.085)	(0.057)
Seq2	-0.199***	-0.201***	-0.143***
	(0.036)	(0.035)	(0.033)
Seq3	-0.264***	-0.255***	-0.149***
	(0.036)	(0.035)	(0.033)
Schools		0.304***	0.172***
		(0.035)	(0.033)
Refugees		0.289***	0.189***
		(0.035)	(0.033)
Republicans		0.594***	0.287***
		(0.066)	(0.045)
Other party		-0.598***	-0.229***
		(0.115)	(0.078)
Gender (f)		-0.210***	-0.067
		(0.062)	(0.042)
Black or African American		0.381***	0.084*
		(0.074)	(0.050)
American Indian or Alaska Native		0.193	0.086
		(0.271)	(0.183)
Asian		-0.325**	-0.105
		(0.143)	(0.097)
Other ethnicity		-0.730***	-0.436***
		(0.234)	(0.158)
Age		-0.069***	-0.068***
		(0.027)	(0.018)
Education		0.178***	0.072***
		(0.028)	(0.019)
Hours		YES	YES
Accuracy			0.601***
			(0.012)
Constant	5.041***	3.890***	1.693***
	(0.070)	(0.182)	(0.132)
R^2	0.021	0.178	0.516
Wald χ^2	95.671***	621.146***	3591.745***
Wald tests (p-values):			
HIGH vs LOW	$< .001$	$< .001$	$< .001$
HIGH vs MACHINE	$< .001$	$< .001$	$< .001$
LOW vs MACHINE	.031	.017	.068
N Observations		4794	
N Groups		1598	

*** $p < .01$; ** $p < .05$; * $p < .1$
Random-effects GLS regression. Standard errors in parentheses.

Table 1: Treatment Effects on Procedural Fairness Across Scenarios

With the *HUMAN* treatment as the reference category, we observe that the reported fairness differences between our treatments are robust to the inclusion of all control variables added in Model 2. The coefficients for the *HIGH* treatment dummy and the *MACHINE* treatment dummy are positive and negative, respectively, and turn out to be statistically

significant. The coefficient for the *LOW* treatment, in contrast, is close to zero and statistically insignificant. Wald tests, run after the estimation of Model 2, confirm the treatment differences between the *HIGH* treatment and either the *LOW* or the *MACHINE* treatment ($ps < .001$), as well as between the latter treatments ($p = .017$).

This leads to the following main results of our study:

Result 1: *Fairness ratings are responsive to different degrees of human involvement in the decision procedures.*

Result 2: *A human-machine interaction with high human involvement is judged as fairer than the decision procedures in all other treatments.*

Result 3: *Purely machine-based decision procedures receive the lowest fairness scores of all procedures.*

Result 4: *Purely human decision-making and human-machine interactions with low human involvement are perceived as equally fair.*

Decision Accuracy. In Model 3, we observe a significant effect of the participants' accuracy assessments on fairness ratings. Controlling for accuracy considerably changes the coefficients of our treatment dummies. However, coefficients for the *HIGH* and *MACHINE* treatment dummies keep their sign and remain (marginally) significant (*HIGH*: $p = .063$, *MACHINE*: $p < .001$), whereas the coefficient for the *LOW* treatment is now clearly negative and marginally significant (*LOW*: $p = .067$). Post-regression Wald tests confirm the further treatment differences, also after controlling for expected accuracy.

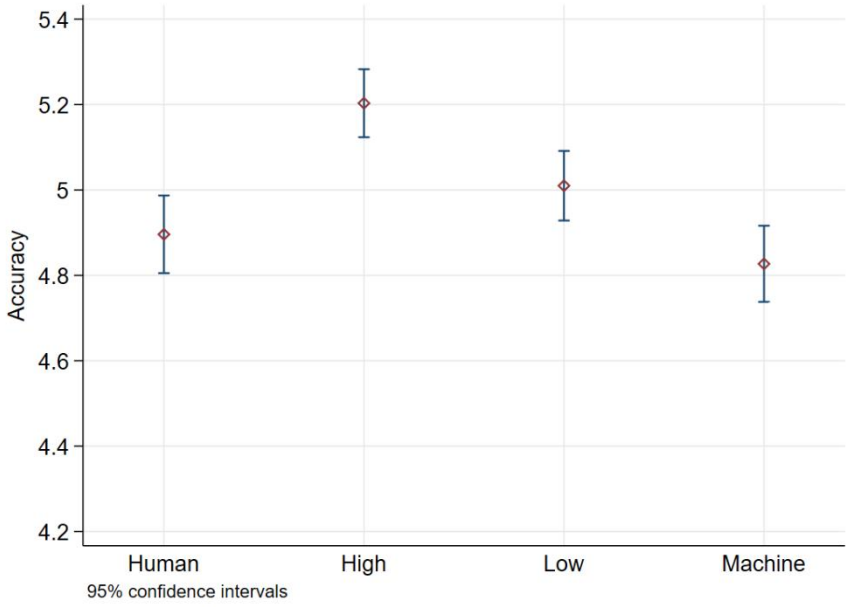

Figure 3: Perceived Accuracy Across all Scenarios

These results lead to the interpretation that people largely seem to prefer the human-computer interaction in the *HIGH* treatment because they think this specific interaction leads to a more accurate result. However, the relative dislike of a purely algorithmic decision in the *MACHINE* treatment is practically not affected by the inclusion of the accuracy assessments (coefficients for the *MACHINE* treatment are of roughly equal size in Model 2 and Model 3 and significant in both models). The difference between the *HIGH* and the *LOW* treatment remains significant after controlling for accuracy ($p < .001$, Wald test).

This suggests that participants are sensitive to variations in the degree of human involvement in algorithmically assisted decision procedures and base their accuracy assessments and fairness ratings on it. Moreover, these results indicate that the rejection of purely algorithmic decisions is not merely driven by the expectation that algorithms make more mistakes.

It seems that a combination of algorithmic and human decision inputs is perceived to produce more accurate factual assessments. As can be seen in Figure 3, high human involvement in the algorithmic decision procedure, as in the *HIGH* treatment ($M = 5.2$), is likely to foster the perceived accuracy of the procedure, as compared to all other conditions ($ps < .001$, MWU, *HUMAN*: $M = 4.90$, $r = .087$; *LOW*: $M = 5.01$, $r = .072$; *MACHINE*: $M = 4.83$, $r = .115$). Participants report no difference in expected accuracy between the *HUMAN* and the *MACHINE* treatment ($p = .223$, MWU). This seems noteworthy because arguably with these ratings our participants would underestimate the capability of the algorithmic prediction – were it a real-world application – as a long-standing literature indicates that, usually, statistical models do better than humans in prediction tasks (Meehl, 1954; Grove et al., 2000; Kleinberg et al., 2018).

Our results so far corroborate that accuracy may play a role in people's fairness assessments of different decision procedures in human-machine interactions. To explore the conjecture further that the perceived accuracy of the procedure mediates fairness ratings in the context of our study, we conduct a mediation analysis to measure the *direct effect* of our treatments (x_i) on fairness ratings (y_i) and the *indirect effect* of our treatments on fairness ratings through accuracy assessments as a mediator (z_i). Following a standard approach in psychology research (Baron & Kenny, 1986; Holmbeck, 1997; Frazier, 2004; Danner et al., 2015), we estimate these effects in a structural equation model, with the treatment effect on the mediator given by:

$$z_{it} = \alpha_0 + \alpha_x x_{it} + u_{it} + \varepsilon_i \ ,$$

where u_{it} denotes the residual error between individuals and ε_i denotes the individual-specific error. The full structural equation model can be specified as follows:

$$y_{it} = \beta_0 + \beta_x x_{it} + \beta_z z_{it} + +u_{it} + \varepsilon_i \, .$$

The direct treatment effect is given by β_x, denoting the pathway from treatment to fairness ratings while controlling for accuracy assessments. The indirect treatment effect is given by $\gamma_I = \alpha_x \cdot \beta_z$, denoting the pathway through accuracy assessments.[13]

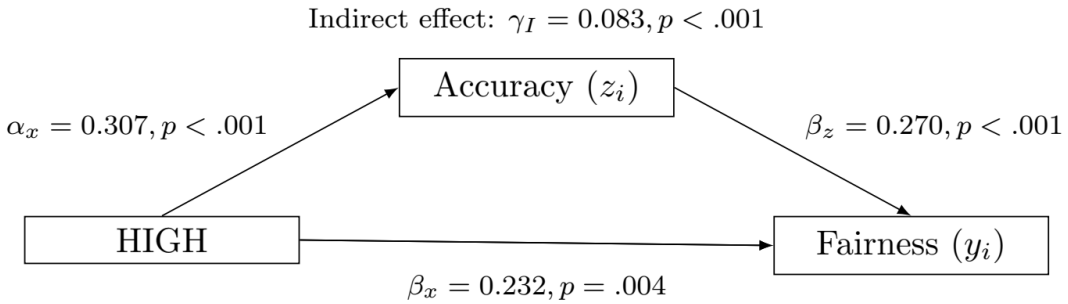

Figure 4: Mediation Analysis HIGH vs. HUMAN

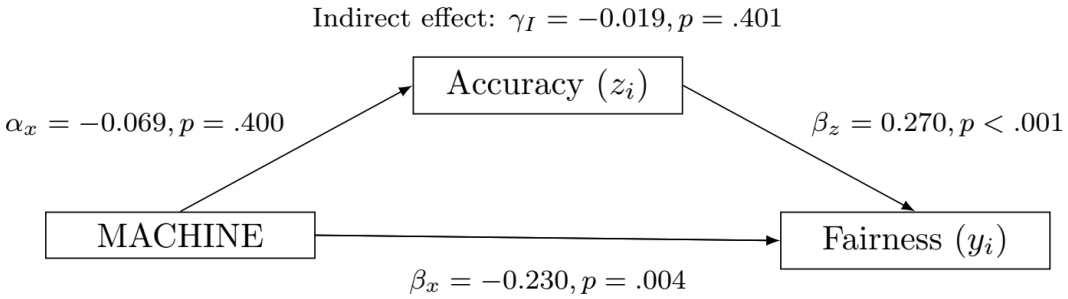

Figure 5: Mediation Analysis MACHINE vs. HUMAN

The results of our mediation analysis show that a considerable part of the *HIGH* treatment effect compared to the *HUMAN* treatment follows the indirect path through accuracy assessments (Figure 4). In the *MACHINE* treatment, by contrast, we observe no significant indirect effect mediated by accuracy (Figure 5). This supports the conclusion that the decrease

13 The total treatment effect is given by $\gamma_T = \beta_x + \alpha_x \cdot \beta_z$ and already reported, for slightly different model specifications, in Table 1.

of fairness ratings observed for purely algorithm-based decision procedures is mostly driven by cognitive or motivational effects that are unrelated to perceived accuracy.

> **Result 5:** *In parts, the HIGH treatment is judged as fairer than the HUMAN treatment because it is perceived as more accurate. The relative dislike of the MACHINE treatment is not affected by accuracy assessments.*

Impact of Socio-Demographic Characteristics. On further inspection of the control variables included in Model 2, we observe several effects of socio-demographic characteristics on fairness assessments. First, participants identifying as Republicans show higher fairness ratings than participants identifying as Democrats. Participants identifying as neither Republican nor Democrat, by contrast, report significantly lower fairness evaluations than Democrats. Second, the coefficient of the Gender dummy also turns out significant, with women reporting lower fairness evaluations than men. Third, we observe a positive correlation between identifying as African American and fairness ratings. Fourth, we find that fairness ratings decline with age. Finally, we observe a positive effect of education on fairness ratings.

A further observation from Model 3 is that the coefficient for the gender dummy is much smaller and no longer significant once we control for accuracy assessments ($p = .111$). Similarly, the coefficient for African Americans turns out much smaller and only marginally significant when controlling for accuracy ($p = .097$). This indicates that the higher fairness ratings of men compared to women and of participants identifying as African American are also in parts driven by the perceived accuracy of the procedure. We indeed find that female participants ($M = 4.80$) express significantly lower accuracy ratings than male participants ($M = 5.09,\ p < .001,$ MWU, $r = .093$). Moreover, accuracy assessments of participants

identifying as African American ($M = 5.65$) are significantly higher than the average accuracy assessments of participants belonging to all other ethnic groups ($M = 4.75, p < .001$, MWU, $r = .280$).[14]

While the effects for female and African American participants vanish once we control for accuracy assessments, the effects of party preference, age, and education seem more robust to the inclusion of all our covariates (Model 3). With Democrats as the reference category, the coefficient for Republicans remains consistently positive throughout all model specifications, whereas we observe a consistently negative effect of identifying with a political ideology beyond the bipartisan Democrat-Republican spectrum. Moreover, we observe a small but significant decline of fairness ratings with age. Finally, while the education coefficient becomes smaller when controlling for accuracy assessments, it remains positive and highly significant throughout all model specifications.

Scenario-Specific Effects

Descriptively, the overall pattern of the aggregated results is also present if we look at the treatments in the three scenarios individually.[15] Fairness ratings in the treatments for each scenario can be seen in Figure 6. In all three scenarios, fairness ratings are highest in the *HIGH* treatment and lowest in the *MACHINE* treatment. In the school-admissions and the refugee-

14 This also holds for the purely machine-based decision procedure (African American: $M = 5.34$, all other: $M = 4.68, p < .001$, MWU, $r = .200$).

15 As mentioned before, each participant answered the fairness question in the same treatment in three different scenarios. The effects of the timing of the decision seem to be generally unaffected by the different scenarios. In Appendix B, we report a random-effects generalized least squares regression model, in which all interactions of the decision point in time and the scenarios turn out insignificant, with the exception of the refugee-matching scenario, where presenting the scenario last yields a (marginally) significantly more negative effect than the other two treatments.

matching scenarios, the fairness ratings of the other two treatments are in between, with the *HUMAN* treatment being assessed as (slightly) fairer than the *LOW* treatment.[16]

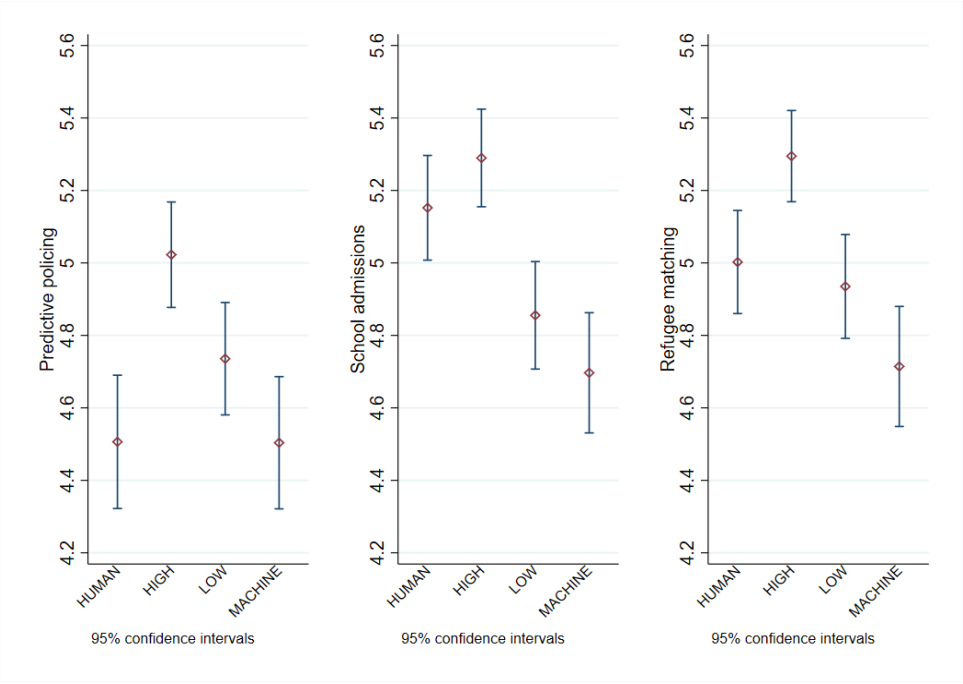

Figure 6: Procedural Fairness by Scenario

Predictive Policing. The predictive-policing scenario stands out in this regard, as decisions by a human police officer are considered less fair than decisions by human decision-makers in the other two scenarios.[17] Our analysis suggests that the fairness-enhancing effect of a human decision-maker is entirely captured by the school-admissions and refugee-matching

16 In the main text, we limit the analyses of the scenario-specific effects to non-parametric MWU tests. Further results obtained from ordinary least squares linear regression models also considering the elicited individual-level covariates can be found in Tables E1, E2, and E3 in Appendix E.

17 This can be shown in a generalized least squares random-effects regression model estimating treatment effects on fairness ratings, with the *MACHINE* treatment and the predictive-policing scenario as reference categories (Table E4 in Appendix E). We observe a significant effect of all treatments both in our base specification (Model 1) and in our specification including dummies as for the school admissions and the refugee-matching scenarios as controls (Model 2). When including an interaction term for treatment and scenario, however, the coefficient for the *HUMAN* treatment in the predictive policing scenario is no longer significant, whereas we observe a significant interaction between the *HUMAN* treatment and the school-admissions and the refugee-matching scenarios (Model 3).

context. Overall, there seems to be a context-specific difference between human police officers and other public officials.

Accordingly, in the predictive-policing scenario, we find that the average fairness ratings in the *HUMAN* ($M = 4.51$) and *MACHINE* treatment ($M = 4.50$) are virtually identical ($p = .883$, MWU). Moreover, we do not find a significant difference either between the *LOW* ($M = 4.74$) and the *MACHINE* treatment ($p = .199$, MWU) or between the *LOW* and the *HUMAN* treatment ($p = .288$, MWU). Yet, we observe significantly higher fairness ratings in the *HIGH* treatment ($M = 5.02$) than in all other treatments (*HUMAN*: $p < .001$, MWU, $r = .118$; *LOW*: $p = .013$, MWU, $r = .088$; *MACHINE*: $p < .001$, MWU, $r = .125$). We interpret these results as evidence of relatively strong fairness preferences for hybrid predictive-policing procedures involving the combined input of humans and algorithms.

The school-admissions and the refugee-matching scenario look much more similar, with the *HIGH* treatment being consistently perceived as the fairest and the *HUMAN* treatment performing consistently better in terms of fairness than the *MACHINE* treatment across both scenarios.

School Admissions. In the school-admissions scenario, fairness ratings are highest in the *HUMAN* ($M = 5.15$) and the *HIGH* ($M = 5.29$) treatment, with both treatments being rather close to each other ($p = .211$, MWU). The *HIGH* treatment yields significantly higher fairness ratings than the *LOW* ($M = 4.86$, $p < .001$, MWU, $r = .151$) and the *MACHINE* treatment ($M = 4.70$, $p < .001$, MWU, $r = .178$). Also, the *HUMAN* treatment leads to significantly higher fairness ratings than the *LOW* ($p = .004$, MWU, $r = .101$) and the *MACHINE* treatment ($p < .001$, MWU, $r = .133$). The pronounced difference between our treatments with strong human involvement and the other two (more algorithmic) treatments points to the particular importance of human oversight in areas as sensitive as school

admissions. The markedly positive effect of our *HUMAN* treatment may also be due to the fact that, unlike in the other scenarios, the decision is made by a group – the school-admissions board – rather than by an individual.

Refugee Matching. In the refugee-matching scenario, by contrast, the *HIGH* treatment ($M = 5.29$) produces significantly higher fairness ratings than all other treatments (*HUMAN*: $M = 5.00$, $p = .004$, MWU, $r = .102$, *LOW*: $M = 4.94$, $p < .001$, MWU, $r = .124$ *MACHINE*: $M = 4.71$, $p < .001$, MWU, $r = .164$). However, fairness ratings differ neither between the *HUMAN* and the *LOW* treatment ($p = .582$, MWU) nor between the *LOW* and the *MACHINE* treatment ($p = .165$, MWU). Moreover, even when comparing the *HUMAN* and the *MACHINE* treatment, we only find a marginally positive effect of an entirely human refugee-matching procedure ($p = .055$, MWU, $r = .068$). While a procedure based on human-computer interaction and high human control is viewed as bolstering procedural fairness, the degree of human involvement does not seem to matter much when it comes to refugee-matching. This may be because issues of distributive justice or participatory rights of those affected by the decision are less salient in refugee-matching procedures than in other contexts.

Discussion

In this article, we report experimental evidence on the importance of human involvement in algorithmically assisted public-sector decision-making for fairness perceptions. Working within the framework of procedural justice research, we find for several application contexts that procedures are perceived as fairest when an algorithmic decision aid is accompanied by high human involvement in the decision-making procedure. Arguably, this is the case to a large extent because people expect these procedures to be the most accurate. By contrast, we observe that purely algorithmic decisions are consistently judged as least fair.

Theoretical Implications

These results have important theoretical implications. While the perceived accuracy matters for fairness perceptions in our experiment, it cannot explain people's dislike for purely algorithmic decision-making as these evaluations seem to be largely independent of the perceived accuracy of the procedure. This is in line with previous findings on accuracy and procedural fairness in bail decisions (Wang, 2018), but contradicts the explanation that people dislike algorithmic decision-making for more complex tasks because they doubt that computers are fit to make these decisions correctly (Scurich & Krauss, 2020; Green & Y. Chen, 2019b). We have thus provided evidence that normative judgements of algorithmic decisions are not just a function of perceived accuracy.

Scurich & Krauss (2020) offer an additional explanation for the dislike of algorithms, arguing that while people might think that algorithms are fulfilling their task correctly, they might simultaneously perceive them as facilitating the discrimination of minorities and the socio-economically disenfranchised. A related concern is that human decision-makers interacting with an algorithmic decision-aid might even exacerbate algorithmic discrimination and the risk of procedurally unfair decisions, for example by being more likely to deviate from an algorithmic recommendation to the disadvantage of African American defendants and to the advantage of Caucasian defendants in pretrial bail decisions (Green & Y. Chen, 2019a). A tendency to opportunistic adjustments of algorithmic outcomes bears the risk of reinforcing the effects of motivated reasoning, decreasing predictive accuracy, and thus hampering procedural fairness (Cohen et al., 2016; Scurich & Krauss, 2013). Our results, by contrast, suggest that human involvement and oversight are key in sustaining perceived procedural fairness.

While a high level of human involvement boosts the procedural fairness of algorithmic assistance, it counteracts the efficiency promises of algorithmic decision aids. In the treatment

with high human involvement, human and algorithmic decision-making always coincide. There is no real substitution of human decision-making by the algorithm. However, our findings also lend support to the view that decision-making procedures with reduced human involvement might yield similar fairness perceptions as the status quo of purely human decision-making procedures.

Caveats

Our findings come with caveats, of course. One limitation of our study stems from the fact that, in all our treatments with human involvement, the human decision-maker at least theoretically retains final control. The human can reverse every decision by the algorithm. The delegation of decision power to the machine in our treatment with low human involvement is only factual. Human decision-makers de facto forgo the opportunity to evaluate the facts of the case, but they are not legally obliged to refrain from performing their own assessment.

Moreover, treatment differences are in some instances sensitive to the decision context. We find noteworthy differences between the three scenarios for predictive policing, school admissions, and refugee-matching. For example, assessments of human decisions considerably vary across contexts, with the predictive-policing scenario showing considerably lower fairness ratings for a human decision-maker as compared to the other two treatments. This difference may reflect a general loss of trust in human police officers in light of repeated abuses of police authority and increasing public awareness of police brutality, such as the murder of George Floyd in 2020.

Finally, it is important to note that our study, like any other vignette study, may be prone to hypothetical bias and may not fully capture evolving behavioral or emotional patterns in public responses to algorithm-assisted decisions. Yet, evidence from previous external

validation tests show that vignette studies encouraging comparisons between different attributes perform remarkably well in capturing actual behavior (Hainmueller et al., 2015). Having implemented our treatments across different scenarios, we are confident that we encouraged participants to seriously engage with the task at hand. And while it is true that the acceptance of algorithm-assisted decision procedures may evolve relatively quickly, it is important to stress that several empirical results in law and psychology are snapshots of evolving psychological patterns. Our study is a first useful step in evaluating the psychological forces underlying the behavioral response to algorithm-assisted decision procedures in three different contexts.

Policy Implications

The results of our study suggest that moving from the status quo of public decision-making by humans towards algorithmic decision-making procedures may be less disruptive in terms of procedural fairness than the law and policy debate sometimes suggests. In our treatment with low human involvement, the decision is usually based on the algorithmic advice alone, with the human decision-maker only sometimes engaging in a personal assessment of the facts. This leads to largely similar fairness ratings than an entirely human decision-making procedure. Hence, while human involvement matters to people, they are relatively open to moderate degrees of decision delegation to a machine.

These results also indicate that the recent trend towards fully automated decision-making and away from human discretion and intervention, especially in the US federal administration (Engstrom et al., 2020), comes at a cost. While exclusive reliance on algorithmic outputs reduces the room for undue political influence, agency capture, or personal biases (Cuéllar & Huq, 2022), it also abates procedural fairness. Legislative proposals of fully algorithmic decision procedures may therefore not adequately reflect public opinion and are

likely to entail stronger opposition and – if enacted – lower levels of compliance. Our results thus suggest that there is a trade-off between procedural justice and mechanic objectivity. Maintaining sufficient room for human discretion and the ability to override algorithmic outputs may be an important condition of public support and compliance (see, Williams, 2022; Garrett & Monahan, 2020; Oswald, 2018). Moreover, our results suggest that agencies and courts may be well-advised to adopt a principled approach to the right to a human decision enshrined in Art. 22 (1) EU General Data Protection Regulation. Justifications of exclusive reliance on algorithms under Art. 22 (2)(b) EU General Data Protection Regulation are likely to strike a better balance between procedural justice and mechanic objectivity if they also consider the psychological cost of full automation for those affected by the decision procedure.

The observed differences between the different scenarios indicate that there may be no *one-size-fits-all solution* for the use of algorithms in public-sector decision-making. For example, fairness perceptions of human decisions are rather high in the school-admissions context. While this may be due to the perceived importance of school admissions or the fact that the admissions decision is made by a collective in this scenario, our experiment is not designed to generate data in support of these interpretations. It is up to future research to explore the optimal mix of human and algorithmic involvement in decision-making procedures for specific policy fields.

In sum, the success or failure of algorithmic governance is likely to depend on the degree of human agency and human accountability. A complete absence of human involvement, human oversight, and human discretion is unlikely to square with due process requirements and legal human-in-the-loop or human-on-the-loop rules. More specifically, our results cast doubt on the idea that the human element can be limited to the definition of policy goals when designing an algorithmic governance system and be completely occluded from the view of those affected by

the decision (see, Cuéllar & Huq, 2022). The human mind at least for now seems to be hardwired to feel at ease with “government by humans, for humans”.

References

Acharya, A., Bansak, K., & Hainmueller, J. (2022). Combining outcome-based and preference-based matching: A constrained priority mechanism. *Political Analysis, 30*(1), 89–12. https://doi.org/10.1017/pan.2020.48

Ahani, N., Andersson, T., Martinello, A., Teytelboym, A., & Trapp, A. C. (2021). Placement optimization in refugee resettlement. *Operations Research, 69*(5), 1468–1486. https://doi.org/10.1287/opre.2020.2093

Albach, M., & Wright, J. R. (2021). The role of accuracy in algorithmic process fairness across multiple domains. In *Proceedings of the 22nd ACM Conference on Economics and Computation* (pp. 29–49). Association of Computing Machinery. https://doi.org/10.1145/3465456.3467620

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY, 35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science, 359*(6373), 325–329. https://doi.org/10.1126/science.aao4408

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. Retrieved from https://fairmlbook.org

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of*

Personality and Social Psychology, 51(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Binns, R. (2022). Human judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance, 16*(1), 197–211. https://doi.org/10.1111/rego.12358

Casey, A. J., & Niblett, A. (2017). The death of rules and standards. *Indiana Law Journal, 92*(4), 1401–1447.

Chen, B. M., Stremitzer, A. & Tobia, K. (n.d.). Having your day in robot court. *Harvard Journal of Law & Technology, 36,* forthcoming.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Cohen, T. H., Pendergast, B., & VanBenschoten, S. W. (2016). Examining overrides of risk classifications for offenders on federal supervision, *Federal Probation, 80*(1), 12–21.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. Z. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 797–806). Association of Computing Machinery. https://doi.org/10.1145/3097983.3098095

Crawford, K., & Schultz, J. (2019). AI systems as state actors. *Columbia Law Review, 119*(7), 1941–1972.

Cuéllar, M.-F., & Huq, A. Z. (2022). Artificially intelligent regulation. *Daedalus, 151*(2), 335–347. https://doi.org/10.1162/daed_a_01920

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313

Danner, D., Hagemann, D., & Fiedler, K. (2015). Mediation analysis with structural equation models: Combining theory, design, and statistics. *European Journal of Social Psychology, 45*(1), 460–481. https://doi.org/10.1002/ejsp.2106

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://dx.doi.org/10.1037/xge0000033

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E. & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 275–285). Association for Computing Machinery. https://doi.org/10.1145/3301275.3302310

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances,4*(1), Article eaao5580. https://doi.org/10.1126/sciadv.aao5580

Engstrom, D. F., Ho, D. E., Sharkey, C. M., Cuéllar, M.-F. (2020). *Government by algorithm: Artificial intelligence in federal administrative agencies.* Administrative

Conference of the United States. https://www.acus.gov/sites/default/files/documents/Government%20by%20Algorithm.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51*(1), 115–134. https://doi.org/10.1037/0022-0167.51.1.115

Garrett, B. L., & Monahan, J. (2020). Judging risk. *California Law Review, 108*(2), 439–493. https://doi.org/10.15779/Z38B56D515

Green, B., & Chen, Y. (2019a). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 90–99). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287563

Green, B., & Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction, 3,* Article 50. https://doi.org/10.1145/3359152

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 903–912). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186138

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 51–60. https://doi.org/10.1609/aaai.v32i1.11296

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, *112*(8), 2395–2400. https://doi.org/10.1073/pnas.1416587112

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., & Ur, B. (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* (pp. 392–402). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372831

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, *106*(4), 811–866.

Holmbeck, G. N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychology literatures. *Journal of Consulting and Clinical Psychology*, *65*(4), 599–610. https://doi.org/10.1037/0022-006X.65.4.599

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*, 399–425. https://doi.org/10.1007/s10683-011-9273-9

Huq, A. Z. (2020a). Constitutional rights in the machine-learning state. *Cornell Law Review, 105*(7), 1875–1954.

Huq, A. Z. (2020b). A right to a human decision. *Virginia Law Review, 106*(3), 611–688.

Imai, K., Jiang, Z., Greiner, J., Halen, R., & Shin, S. (2021). *Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment.* ArXiv. https://doi.org/10.48550/arXiv.2012.02845

Joh, E. E. (2016). The new surveillance discretion: Automated suspicion, Big Data, and policing. *Harvard Law & Policy Review, 10*(1), 15–42.

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. New York, NY: Oxford University Press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics, 133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference* (pp- 43:1–43:23). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1035–1048). Association for Computing Machinery. https://doi.org/10.1145/2998181.2998230

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3,* Article 182. https://doi.org/10.1145/3359284

Lind, E. A., & Tyler, T. R. (1988). *The Social Psychology of Procedural Justice*. New York, NY: Plenum Press.

Logg, J. M., Minson, J. A. & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151,* 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making, 6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 122–130). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372867

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota Press.

Miller, S. M., & Keiser, L. R. (2021). Representative bureaucracy and attitudes toward automated decision making. *Journal of Public Administration Research and Theory, 31*(1), 150–165. https://doi.org/10.1093/jopart/muaa019

Muratov, E., Lewis, M., Fourches, D., Tropsha, A., & Cox, W. C. (2017). Computer-assisted decision support for student admissions based on their predicted academic performance. *American Journal of Pharmaceutical Education, 81*(3), Article 46. DOI: https://doi.org/10.5688/ajpe81346

Nagtegaal, R. (2021). The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly, 38*(1), Article 101536. https://doi.org/10.1016/j.giq.2020.101536

Oswald, M. (2018). Algorithm-assisted decision-making in the public sector: Framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A, 376*, Article 20170359. https://doi.org/10.1098/rsta.2017.0359

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to Power Analysis for simple experimental designs. *International Review of Social Psychology, 31*(1), Article 20. https://doi.org/10.5334/irsp.181

Rambachan, A., Kleinberg, J., Ludwig, J., & Mullainathan, S. (2020). An economic perspective on algorithmic fairness. *AEA Papers and Proceedings, 110*, 91–95. https://doi.org/10.1257/pandp.20201036

Scurich, N., & Krauss, D. A. (2020). Public's views of risk assessment algorithms and pretrial decision making. *Psychology, Public Policy, and Law, 26*(1), 1–9. https://doi.org/10.1037/law0000219

Scurich, N., & Krauss, D. A. (2013). The effect of adjusted actuarial risk assessment on mock-jurors' decisions in a sexual predator commitment proceeding. *Jurimetrics, 53*(3), 395–413.

Simmons, R. (2018). Big Data, machine judges, and the legitimacy of the criminal justice system. *University of California Davis Law Review, 52*(2), 1067–1118.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). *Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature.* ArXiv. https://doi.org/10.48550/arXiv.2103.12016

Trinkner, R., Jackson, J., & Tyler, T. R. (2018). Bounded authority: Expanding "appropriate" police behavior beyond procedural justice. *Law and Human Behavior, 42*(3), 280–293. https://doi.org/10.1037/lhb0000285

Tyler, T. R. (2003). Procedural justice, legitimacy, and the effective rule of law. *Crime and Justice, 30,* 283–357.

Tyler, T. R. (2006). *Why people obey the law*. (2nd ed.). Princeton, NJ: Princeton University Press.

Tyler, T. R., & Blader, S. L. (2000). *Cooperation in groups: Procedural justice, social identity, and behavioral engagement.* Philadelphia, PA: Psychology Press.

Tyler, T. R., & Jackson, J. (2014). Popular legitimacy and the exercise of legal authority: Motivating compliance, cooperation, and engagement. *Psychology, Public Policy, and Law, 20*(1), 78–95. https://doi.org/10.1037/a0034514

Tyler, T. R. & Lind, E. A. (1992). A relational model of authority in groups. In M. P. Zanna (Ed.) *Advances in Experimental Social Psychology*, *25* (pp. 115–191). Academic Press https://doi.org/10.1016/S0065-2601(08)60283-X

Tyler, T. R., & Sevier, J. (2014). 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*(3), 1095–1137.

Walker, L., Lind, E. A. & Thibaut, J. (1979). The relation between procedural and distributive justice. *Virginia Law Review*, *65*(8), 1401–1420.

Wang, A. J. (2018). *Procedural justice and risk-assessment algorithms.* http://dx.doi.org/10.2139/ssrn.3170136

Williams, R. (2022). Rethinking administrative law for algorithmic decision making. *Oxford Journal of Legal Studies, 42*(2), 468–494. https://doi.org/10.1093/ojls/gqab032

Yaghini, M., Heidari, H., & Krause, A. (2021). A human-in-the-loop framework to construct context-dependent mathematical formulations of fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp.1023-1033). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462583

Yalcin, G., Themeli, E., Stamhuis, E., Philipsen, S., & Puntoni, S. (2022). Perceptions of justice by algorithms. *Artificial Intelligence and Law. https://doi.org/10.1007/s10506-022-09312-z*

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper No. 279). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300509

Appendix A

Instructions

Police

One of the main tasks of the police is to prevent criminal behavior. In order to deploy their forces in an optimal manner, the police need to assess the risk that criminal behavior will occur. This risk assessment refers to various types of criminal behavior, including the risk of violent assaults.

Suppose the local police want to assess the risk of violent assaults in certain areas of the city - including the probable type, location, and time of the assault - and perform bodily searches of all persons within a small and well-defined area of the city. The purpose of these bodily searches is to track down weapons used for violent assaults.

*Decision making procedure **(HUMAN)***

The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.
Based on his or her risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.

*Decision making procedure **(HIGH)***

The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent

assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer.

The decision will never be based on the computer algorithm alone. The police officer will always conduct his or her own analysis, that means, the police officer will in each case conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.

Based on the risk assessment of the computer algorithm and his or her own risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.

*Decision making procedure **(LOW)***

The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer.

The decision will usually be based on the computer algorithm alone. The police officer will sometimes conduct his or her own analysis, that means, the police officer will in some cases conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.

Based on the risk assessment of the computer algorithm and - only if conducted - his or her own risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.

*Decision making procedure **(MACHINE)***

The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer.

The decision will be based on the computer algorithm's assessment alone.

Based on its risk assessment, the computer algorithm will order or not order bodily searches in a certain area of the city. The police officer cannot override the decision of the computer algorithm and has no discretion in this decision.

How fair do you consider the procedure by which the police come to their decision? (1=very unfair, 7=very fair)

How accurately do you think the police will assess the risk of violent crimes in the city? (1=not accurately at all, 7=extremely accurately)

To what extent would you want the decision making procedure to be used in a case if you were personally concerned? (1=not at all, 7=to a large extent)

To what extent would you want the decision making procedure to be used in a case for the general public? (1=not at all, 7=to a large extent)

Schools

Many public schools have limited capacities. Accordingly, these schools are unable to accept all students who apply. Therefore, the school admissions boards have to select students based

on some criterion. One prominent criterion is the chance that an applicant will succeed within the respective school system. In applying this criterion, school admissions boards usually assess the probability that the applicant will eventually graduate.

Suppose a school admissions board wants to assess this success probability and decide whether to accept or reject an applicant.

*Decision making procedure **(HUMAN)***

The school admissions board will conduct an interview with the applicant and collect additional information on the applicant's school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will conduct an in-depth analysis of the case material and assess the applicant's success probability.

Based on its assessment of the success probability, the admissions board will accept or reject the applicant. The admissions board has discretion in this decision.

*Decision making procedure **(HIGH)***

The school admissions board will conduct an interview with the applicant and collect additional information on the applicant's school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will use a computer algorithm to assess the applicant's success probability. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the applicant's success probability to the admissions board.

The decision will never be based on the computer algorithm alone. The admissions board will always conduct its own analysis, that means, the admissions board will in each case conduct an in-depth analysis of the case material, and assess the applicant's success probability.

Based on the assessment of the success probability of the computer algorithm and its own assessment of the success probability, the admissions board will accept or reject the applicant. The admissions board has discretion in this decision.

*Decision making procedure **(LOW)***

The school admissions board will conduct an interview with the applicant and collect additional information on the applicant's school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will use a computer algorithm to assess the applicant's success probability. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the applicant's success probability to the admissions board.

The decision will usually be based on the computer algorithm alone. The admissions board will sometimes conduct its own analysis, that means, the admissions board will in some cases conduct an in-depth analysis of the case material and assess the applicant's success probability. Based on the assessment of the success probability of the computer algorithm and - only if conducted - its own assessment of the success probability, the admissions board will accept or reject the applicant. The admissions board has discretion in this decision.

*Decision making procedure **(MACHINE)***

The school admissions board will conduct an interview with the applicant and collect additional information on the applicant's school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will use a computer algorithm to assess the applicant's success probability. The computer algorithm will conduct

an in-depth analysis of the case material and present its assessment of the applicant's success probability to the admissions board.

The decision will be based on the computer algorithm's assessment alone.

Based on its assessment of the success probability, the computer algorithm will accept or reject the applicant. The admissions board cannot override the decision of the computer algorithm and has no discretion in this decision.

How fair do you consider the procedure by which the school admissions board comes to its decision? (1=very unfair, 7=very fair)

How accurately do you think the school admissions board will assess the probability that the applicant will eventually graduate? (1=not accurately at all, 7=extremely accurately)

To what extent would you want the decision making procedure to be used in a case if you were personally concerned? (1=not at all, 7=to a large extent)

To what extent would you want the decision making procedure to be used in a case for the general public? (1=not at all, 7=to a large extent)

Refugees

One of the main tasks of immigration authorities is to assign refugees to certain locations within the country of immigration. Refugee facilities have limited capacities. Therefore, the immigration authorities have to assign refugees based on some criterion. One prominent criterion is the chance that a refugee will be able to integrate herself into society. In applying this criterion, immigration authorities usually assess the probability that the refugee will successfully find employment when assigned to a certain location.

Suppose an immigration authority wants to assess this probability and decide to which location within the country of immigration a refugee should be assigned.

*Decision making procedure **(HUMAN)***

The case manager will conduct an interview with the refugee and collect information on the refugee's origin, education, linguistic skills, and family status. Then the case manager will conduct an in-depth analysis of the case material and assess the probability of successful employment.

Based on his or her assessment of the probability of successful employment, the case manager will assign the refugee to a certain location. The case manager has discretion in this decision.

*Decision making procedure **(HIGH)***

The case manager will conduct an interview with the refugee and collect information on the refugee's origin, education, linguistic skills, and family status. Then the case manager will use a computer algorithm to assess the probability of successful employment. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the probability of successful employment to the case manager.

The decision will never be based on the computer algorithm alone. The case manager will always conduct his or her own analysis, that means, the case manager will in each case conduct an in-depth analysis of the case material and assess the probability of successful employment.

Based on the assessment of the probability of successful employment of the computer algorithm and his or her own assessment of the probability of successful employment, the

case manager will assign the refugee to a certain location. The case manager has discretion in this decision.

*Decision making procedure **(LOW)***

The case manager will conduct an interview with the refugee and collect information on the refugee's origin, education, linguistic skills, and family status. Then the case manager will use a computer algorithm to assess the probability of successful employment. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the probability of successful employment to the case manager.

The decision will usually be based on the computer algorithm alone. The case manager will sometimes conduct his or her own analysis, that means, the case manager will in some cases conduct an in-depth analysis of the case material and assess the probability of successful employment.

Based on the assessment of the probability of successful employment of the computer algorithm and - only if conducted - his or her own assessment of the probability of successful employment, the case manager will assign the refugee to a certain location. The case manager has discretion in this decision.

*Decision making procedure **(MACHINE)***

The case manager will conduct an interview with the refugee and collect information on the refugee's origin, education, linguistic skills, and family status. Then the case manager will use a computer algorithm to assess the probability of successful employment. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the probability of successful employment to the case manager.

The decision will be based on the computer algorithm's assessment alone.

Based on its assessment of the probability of successful employment, the computer algorithm will assign the refugee to a certain location. The case manager cannot override the decision of the computer algorithm and has no discretion in this decision.

How fair do you consider the procedure by which the immigration authority comes to its decision? (1=very unfair, 7=very fair)

How accurately do you think the immigration authority will assess the probability that the refugee will successfully find employment? (1=not accurately at all, 7=extremely accurately)

To what extent would you want the decision making procedure to be used in a case if you were personally concerned? (1=not at all, 7=to a large extent)

To what extent would you want the decision making procedure to be used in a case for the general public? (1=not at all, 7=to a large extent)

[Questionnaire]

1. You have seen three different scenarios. Please rank these scenarios according to the severeness of the decision for the recipient from 1 (least severe) to 3 (most severe).
2. In this survey, you have been asked to assess the fairness of several decision making procedures by public officials. Please state shortly for what reasons you decided the way you did, especially on which criteria you based your evaluation of the fairness of the procedure (keywords are sufficient).
3. How old are you?
4. What is your highest educational degree?

5. What is your gender?

6. What is your ethnicity?

7. Which political party do you feel closest to?

8. How many hours per week do you spend online doing tasks for money?

Appendix B

Treatment Effects on Procedural Preferences

In this Appendix, we present a summary analysis of our results on procedural preferences regarding the involvement of oneself (Table B1) or others (Table B2).

Baseline: HUMAN DV: Procedural preferences (self)	(1)	(2)	(3)
HIGH	0.310***	0.237**	0.083
	(0.106)	(0.092)	(0.064)
LOW	-0.082	-0.079	-0.139**
	(0.106)	(0.092)	(0.064)
MACHINE	-0.288***	-0.288***	-0.248***
	(0.106)	(0.092)	(0.063)
Seq2	-0.146***	-0.147***	-0.087**
	(0.039)	(0.039)	(0.037)
Seq3	-0.210***	-0.203***	-0.093**
	(0.039)	(0.039)	(0.037)
Schools		0.228***	0.090**
		(0.039)	(0.037)
Refugees		0.227***	0.123***
		(0.039)	(0.037)
Republicans		0.608***	0.288***
		(0.072)	(0.050)
Other party		-0.794***	-0.409***
		(0.125)	(0.086)
Gender (f)		-0.204***	-0.056
		(0.068)	(0.047)
Black or African American		0.534***	0.224***
		(0.080)	(0.056)
American Indian or Alaska Native		0.157	0.046
		(0.294)	(0.203)
Asian		-0.415***	-0.186*
		(0.156)	(0.107)
Other ethnicity		-0.667***	-0.361**
		(0.254)	(0.175)
Age		-0.092***	-0.091***
		(0.029)	(0.020)
Education		0.257***	0.146***
		(0.031)	(0.021)
Hours		YES	YES
Accuracy			0.626***
			(0.014)
Constant	4.901***	3.433***	1.147***
	(0.078)	(0.197)	(0.147)
R^2	0.017	0.202	0.507
Wald χ^2	63.466***	652.082***	3361.841***
Wald tests (p-values):			
HIGH vs LOW	$< .001$	$< .001$	$< .001$
HIGH vs MACHINE	$< .001$	$< .001$	$< .001$
LOW vs MACHINE	.052	.023	.087
N Observations		4794	
N Groups		1598	

*** $p < .01$; ** $p < .05$; * $p < .1$
Random-effects GLS regression. Standard errors in parentheses.

Table B1: Treatment Effects on the Procedural Preferences Regarding Oneself Across Scenarios

Baseline: HUMAN DV: Procedural preferences (others)	(1)	(2)	(3)
HIGH	0.395***	0.324***	0.152***
	(0.104)	(0.090)	(0.057)
LOW	0.015	0.019	-0.049
	(0.104)	(0.091)	(0.057)
MACHINE	-0.211**	-0.205**	-0.160***
	(0.104)	(0.090)	(0.057)
Seq2	-0.112***	-0.115***	-0.047
	(0.038)	(0.038)	(0.035)
Seq3	-0.120***	-0.112***	0.011
	(0.038)	(0.038)	(0.035)
Schools		0.278***	0.123***
		(0.038)	(0.035)
Refugees		0.246***	0.129***
		(0.038)	(0.035)
Republicans		0.646***	0.288***
		(0.071)	(0.045)
Other party		-0.649***	-0.219***
		(0.122)	(0.077)
Gender (f)		-0.223***	-0.056
		(0.066)	(0.042)
Black or African American		0.581***	0.234***
		(0.079)	(0.050)
American Indian or Alaska Native		0.144	0.018
		(0.289)	(0.182)
Asian		-0.340**	-0.083
		(0.153)	(0.096)
Other ethnicity		-0.781***	-0.438***
		(0.249)	(0.157)
Age		-0.080***	-0.079***
		(0.029)	(0.018)
Education		0.226***	0.103***
		(0.030)	(0.019)
Hours		YES	YES
Accuracy			0.701***
			(0.013)
Constant	4.841***	3.405***	0.845***
	(0.077)	(0.194)	(0.132)
R^2	0.017	0.204	0.572
Wald χ^2	47.367***	659.020***	4639.201***
Wald tests (p-values):			
HIGH vs LOW	$< .001$	$< .001$	$< .001$
HIGH vs MACHINE	$< .001$	$< .001$	$< .001$
LOW vs MACHINE	.030	.014	.052
N Observations		4794	
N Groups		1598	

*** $p < .01$; ** $p < .05$; * $p < .1$
Random-effects GLS regression. Standard errors in parentheses.

Table B2: Treatment Effects on the Procedural Preferences Regarding Others Across Scenarios

Appendix C

Order Effects

In this Appendix, we present an analysis of order effects pooled over all scenarios (Table C1, left column) and for each scenario (Table C1, right column).

Overall DV: Procedural fairness		Scenario-specific DV: Procedural fairness	
Seq2	-0.252***	Seq2	-0.191**
	(0.071)		(0.075)
Seq3	-0.327***	Seq3	-0.135*
	(0.071)		(0.073)
HIGH	0.333***	Schools	0.359***
	(0.111)		(0.074)
LOW	-0.120	Refugees	0.369***
	(0.111)		(0.073)
MACHINE	-0.345***	Seq2 × Schools	-0.013
	(0.111)		(0.112)
Seq2 × HIGH	-0.035	Seq2 × Refugees	-0.017
	(0.101)		(0.112)
Seq2 × LOW	0.110	Seq3 × Schools	-0.144
	(0.101)		(0.112)
Seq2 × MACHINE	0.137	Seq3 × Refugees	-0.219*
	(0.101)		(0.112)
Seq3 × HIGH	-0.018		
	(0.101)		
Seq3 × LOW	0.115		
	(0.101)		
Seq3 × MACHINE	0.154		
	(0.101)		
Constant	5.080***	Constant	4.801***
	(0.078)		(0.059)
R^2	0.022	R^2	0.014
Wald χ^2	101.21***	Wald χ^2	160.77***
Wald tests (p-values):		Wald tests (p-values):	
Seq2 vs Seq3	.294	Seq2 vs Seq3	.450
Seq2 × HIGH vs Seq3 × HIGH	.867	Seq2 × Schools vs Seq3 × Schools	.243
Seq2 × LOW vs Seq3 × LOW	.961	Seq2 × Refugees vs Seq3 × Refugees	.072
Seq2 × MACHINE vs Seq3 × MACHINE	.865		
Seq2 × HIGH vs Seq2 × LOW	.151		
Seq2 × HIGH vs Seq2 × MACHINE	.089		
Seq2 × LOW vs Seq2 × MACHINE	.790		
Seq3 × HIGH vs Seq3 × LOW	.188		
Seq3 × HIGH vs Seq3 × MACHINE	.089		
Seq3 × LOW vs Seq3 × MACHINE	.699		
N Observations	4794	*N Observations*	4794
N Groups	1598	*N Groups*	1598

*** $p < .01$; ** $p < .05$; * $p < .1$
Random-effects GLS regression. Standard errors in parentheses.

Table C1: Order Effects

Appendix D

Sensitivity Analysis

Figure D1 presents the results of a two-sample sensitivity analysis investigating the sensitivity of effect size for fairness rating means pooled across scenarios in response to variations in our study parameters, specifically for varying power levels (between $1 - \beta = 0.6$ and $1 - \beta = 0.95$) and varying significance levels (between $\alpha = 0.1$ and $\alpha = 0.001$). We calculate the minimum detectable change in the mean of fairness ratings between the treatment and the control group ($\delta = M_2 - M_1$), for a given fairness rating mean of $M_1 = 4.60$ (the approximate mean in the MACHINE treatment used as a control group), an equal standard deviation of $\sigma = 1.55$ across samples, and a sample size of $n = 400$.

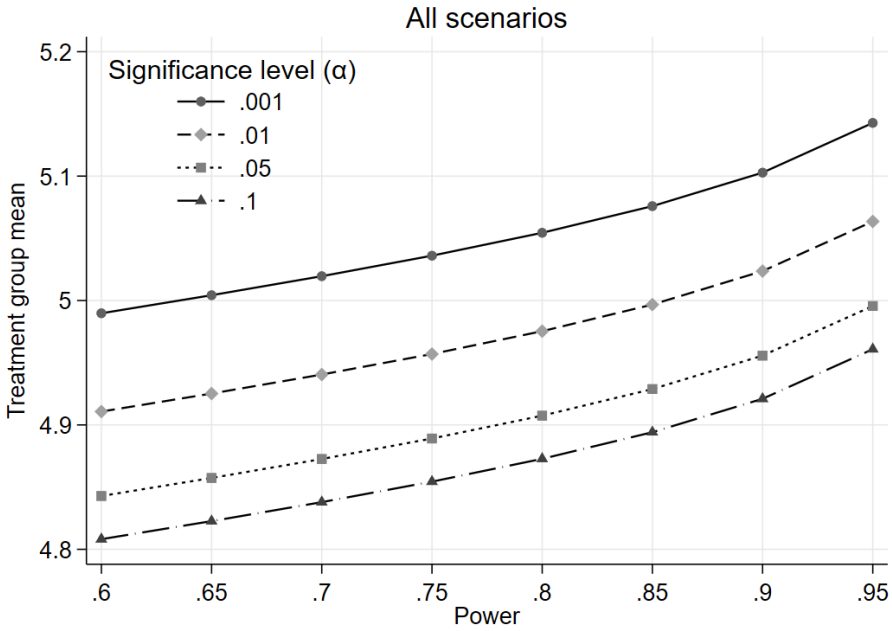

Figure D1: Sensitivity of Fairness Rating Means Pooled Across Scenarios

Appendix E

Additional Analyses of Scenario-Specific Effects

Tables E1, E2, and E3 present an analysis of treatment effects on perceived procedural fairness in each scenario used in our experiment (predictive policing, school admissions, refugee-matching).

Baseline: HUMAN DV: Procedural fairness	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIGH	0.516***	0.437***	0.447***	0.447***	0.450***	0.443***	0.221***
	(0.121)	(0.112)	(0.112)	(0.110)	(0.110)	(0.108)	(0.081)
LOW	0.229*	0.161	0.166	0.185*	0.207*	0.229**	0.036
	(0.120)	(0.112)	(0.112)	(0.110)	(0.110)	(0.108)	(0.081)
MACHINE	-0.002	-0.038	-0.028	-0.000	0.001	-0.005	-0.057
	(0.121)	(0.112)	(0.112)	(0.110)	(0.110)	(0.108)	(0.080)
Republicans		0.951***	0.938***	0.749***	0.772***	0.722***	0.265***
		(0.083)	(0.083)	(0.085)	(0.086)	(0.084)	(0.064)
Other party		-1.089***	-1.100***	-1.073***	-1.077***	-0.921***	-0.259**
		(0.150)	(0.149)	(0.147)	(0.147)	(0.146)	(0.110)
Gender (f)			-0.328***	-0.282***	-0.267***	-0.265***	-0.072
			(0.082)	(0.081)	(0.081)	(0.080)	(0.059)
Black or African American				0.605***	0.593***	0.478***	0.076
				(0.095)	(0.095)	(0.094)	(0.071)
American Indian or Alaska Native				0.281	0.273	0.236	0.315
				(0.352)	(0.351)	(0.346)	(0.257)
Asian				-0.260	-0.314*	-0.285	-0.117
				(0.185)	(0.186)	(0.183)	(0.136)
Other ethnicity				-1.018***	-1.025***	-0.991***	-0.478**
				(0.304)	(0.303)	(0.298)	(0.222)
Age					-0.100***	-0.111***	-0.119***
					(0.035)	(0.034)	(0.025)
Education						0.211***	0.084***
						(0.036)	(0.027)
Hours						YES	YES
Accuracy							0.697***
							(0.019)
Constant	4.506***	4.205***	4.327***	4.254***	4.616***	3.497***	1.178***
	(0.085)	(0.090)	(0.094)	(0.096)	(0.158)	(0.229)	(0.182)
R^2	0.015	0.149	0.157	0.187	0.191	0.218	0.568
F	8.278***	55.591***	49.445***	36.556***	34.143***	33.988***	148.666***
Wald tests (p-values):							
HIGH vs LOW	.018	.014	.012	.017	.027	.050	.022
HIGH vs MACHINE	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
LOW vs MACHINE	.055	.076	.082	.094	.061	.031	.251
N Observations				1598			

*** $p < .01$; ** $p < .05$; * $p < .1$
OLS regression. Standard errors in parentheses.

Table E1: Procedural Fairness in the Predictive-Policing Scenario

Baseline: HUMAN DV: Procedural fairness	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIGH	0.138	0.085	0.093	0.091	0.092	0.087	-0.091
	(0.107)	(0.103)	(0.103)	(0.102)	(0.102)	(0.100)	(0.080)
LOW	-0.297***	-0.339***	-0.335***	-0.320***	-0.307***	-0.289***	-0.336***
	(0.107)	(0.103)	(0.103)	(0.102)	(0.102)	(0.101)	(0.080)
MACHINE	-0.455***	-0.476***	-0.468***	-0.448***	-0.448***	-0.452***	-0.380***
	(0.107)	(0.103)	(0.103)	(0.102)	(0.102)	(0.100)	(0.080)
Republicans		0.660***	0.650***	0.499***	0.512***	0.472***	0.198***
		(0.076)	(0.076)	(0.079)	(0.079)	(0.078)	(0.063)
Other party		-0.581***	-0.590***	-0.571***	-0.573***	-0.445***	-0.153
		(0.138)	(0.138)	(0.136)	(0.136)	(0.136)	(0.109)
Gender (f)			-0.263***	-0.235***	-0.227***	-0.225***	-0.085
			(0.075)	(0.075)	(0.075)	(0.074)	(0.059)
Black or African American				0.440***	0.433***	0.338***	0.060
				(0.088)	(0.088)	(0.088)	(0.071)
American Indian or Alaska Native				-0.053	-0.058	-0.089	-0.250
				(0.326)	(0.325)	(0.321)	(0.257)
Asian				-0.555***	-0.587***	-0.561***	-0.208
				(0.171)	(0.172)	(0.170)	(0.136)
Other ethnicity				-0.484*	-0.488*	-0.459*	-0.274
				(0.281)	(0.281)	(0.277)	(0.222)
Age					-0.058*	-0.067**	-0.055**
					(0.032)	(0.032)	(0.025)
Education						0.169***	0.067**
						(0.034)	(0.027)
Hours						YES	YES
Accuracy							0.646***
							(0.022)
Constant	5.152***	4.925***	5.023***	4.996***	5.204***	4.294***	1.713***
	(0.076)	(0.083)	(0.087)	(0.089)	(0.146)	(0.213)	(0.190)
R^2	0.023	0.093	0.100	0.124	0.126	0.149	0.458
F	12.771***	32.650***	29.420***	22.418***	20.70***	21.265***	95.433***
Wald tests (p-values):							
HIGH vs LOW	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	.002
HIGH vs MACHINE	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
LOW vs MACHINE	.138	.183	.195	.209	.169	.107	.588
N Observations				1598			

*** $p < .01$; ** $p < .05$; * $p < .1$
OLS regression. Standard errors in parentheses.

Table E2: Procedural Fairness in the School-Admissions Scenario

Baseline: HUMAN DV: Procedural fairness	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIGH	0.292***	0.234**	0.240**	0.239**	0.240**	0.234**	0.135*
	(0.104)	(0.100)	(0.100)	(0.099)	(0.099)	(0.098)	(0.074)
LOW	-0.067	-0.112	-0.109	-0.102	-0.097	-0.081	-0.040
	(0.104)	(0.099)	(0.099)	(0.099)	(0.099)	(0.098)	(0.074)
MACHINE	-0.288***	-0.310***	-0.304***	-0.286***	-0.285***	-0.290***	-0.183**
	(0.104)	(0.100)	(0.099)	(0.099)	(0.099)	(0.098)	(0.074)
Republicans		0.755***	0.748***	0.618***	0.623***	0.588***	0.281***
		(0.074)	(0.074)	(0.076)	(0.077)	(0.076)	(0.059)
Other party		-0.550***	-0.556***	-0.540***	-0.541***	-0.429***	-0.127
		(0.133)	(0.133)	(0.132)	(0.132)	(0.132)	(0.101)
Gender (f)			-0.176**	-0.144**	-0.141*	-0.140*	0.009
			(0.073)	(0.072)	(0.073)	(0.072)	(0.055)
Black or African American				0.413***	0.410***	0.328***	0.002
				(0.085)	(0.085)	(0.085)	(0.065)
American Indian or Alaska Native				0.461	0.460	0.433	0.160
				(0.315)	(0.316)	(0.313)	(0.237)
Asian				-0.135	-0.147	-0.128	0.083
				(0.166)	(0.167)	(0.165)	(0.125)
Other ethnicity				-0.760***	-0.762***	-0.739***	-0.435**
				(0.272)	(0.272)	(0.270)	(0.205)
Age					-0.022	-0.029	-0.031
					(0.031)	(0.031)	(0.023)
Education						0.154***	0.025
						(0.033)	(0.025)
Hours						YES	YES
Accuracy							0.682***
							(0.020)
Constant	5.002***	4.732***	4.797***	4.745***	4.824***	4.015***	1.454***
	(0.074)	(0.080)	(0.084)	(0.086)	(0.142)	(0.207)	(0.174)
R^2	0.019	0.108	0.111	0.131	0.131	0.149	0.511
F	10.494***	38.479***	33.132***	23.879***	21.745***	21.302***	118.321***
Wald tests (p-values):							
HIGH vs LOW	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	.001	.019
HIGH vs MACHINE	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
LOW vs MACHINE	.034	.047	.050	.063	.057	.033	.054
N Observations				1598			

$^{***}\ p < .01$; $^{**}\ p < .05$; $^{*}\ p < .1$
OLS regression. Standard errors in parentheses.

Table E3: Procedural Fairness in the Refugee-Matching Scenario

Table E4 presents an analysis of treatment effects on perceived procedural fairness for each scenario with a treatment-scenario interaction in Model 3.

Baseline: MACHINE DV: Procedural fairness	(1)	(2)	(3)
HUMAN	0.249***	0.249***	0.002
	(0.095)	(0.095)	(0.111)
HIGH	0.564***	0.564***	0.519***
	(0.095)	(0.095)	(0.111)
LOW	0.204**	0.204**	0.232**
	(0.095)	(0.095)	(0.111)
Schools		0.307***	0.193***
		(0.036)	(0.071)
Refugees		0.295***	0.211***
		(0.036)	(0.071)
HUMAN x Schools			0.453***
			(0.100)
HUMAN x Refugees			0.286***
			(0.100)
HIGH x Schools			0.074
			(0.100)
HIGH x Refugees			0.062
			(0.100)
LOW x Schools			-0.073
			(0.100)
LOW x Refugees			-0.011
			(0.100)
Constant	4.638***	4.438***	4.504***
	(0.067)	(0.070)	(0.078)
R^2	0.016	0.024	0.027
Wald χ^2	36.28***	132.01***	166.32***
Wald tests (p-values):			
HUMAN vs HIGH	$< .001$	$< .001$	$< .001$
HUMAN vs LOW	.635	.635	.038
HIGH vs LOW	$< .001$	$< .001$	.010
N Observations		4794	
N Groups		1598	

*** $p < .01$; ** $p < .05$; * $p < .1$
Random-effects GLS regression. Standard errors in parentheses.

Table E4: Procedural Fairness with Treatment-Scenario Interaction