Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts

BEN GREEN, University of Michigan, USA
YILING CHEN, Harvard University, USA

Governments are increasingly turning to algorithmic risk assessments when making important decisions, such as whether to release criminal defendants before trial. Policymakers assert that providing public servants with algorithmic advice will improve human risk predictions and thereby lead to better (e.g., fairer) decisions. Yet because many policy decisions require balancing risk-reduction with competing goals, improving the accuracy of predictions may not necessarily improve the quality of decisions. If risk assessments make people more attentive to reducing risk at the expense of other values, these algorithms would diminish the implementation of public policy even as they lead to more accurate predictions. Through an experiment with 2,140 lay participants simulating two high-stakes government contexts, we provide the first direct evidence that risk assessments can systematically alter how people factor risk into their decisions. These shifts counteracted the potential benefits of improved prediction accuracy. In the pretrial setting of our experiment, the risk assessment made participants more sensitive to increases in perceived risk; this shift increased the racial disparity in pretrial detention by 1.9%. In the government loans setting of our experiment, the risk assessment made participants more risk-averse; this shift reduced government aid by 8.3%. These results demonstrate the potential limits and harms of attempts to improve public policy by incorporating predictive algorithms into multifaceted policy decisions. If these observed behaviors occur in practice, presenting risk assessments to public servants would generate unexpected and unjust shifts in public policy without being subject to democratic deliberation or oversight.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *Law, social and behavioral sciences*; • **Social and professional topics** → *Government technology policy*.

Additional Key Words and Phrases: decision-making; risk assessments; public policy

ACM Reference Format:
Ben Green and Yiling Chen. 2021. Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 418 (October 2021), 33 pages. https://doi.org/10.1145/3479562

1 INTRODUCTION

Following recent advances in the quality and accessibility of algorithms, governments increasingly use machine learning when making high-stakes decisions [21, 26]. Many applications of algorithms involve risk assessments, which predict the risk of some adverse outcome. These predictions are then presented to human decision-makers to inform consequential decisions about individuals. Applications of public sector risk assessments include informing pretrial and sentencing decisions with a criminal defendant's likelihood to recidivate [55, 73], targeting public health inspections

Authors' addresses: Ben Green, University of Michigan, Ann Arbor, MI, USA, bzgreen@umich.edu; Yiling Chen, Harvard University, Cambridge, MA, USA, yiling@seas.harvard.edu.

2573-0142/2021/10-ART418 $15.00
https://doi.org/10.1145/3479562

based on a child's risk of lead poisoning [56], and directing child welfare interventions based on a child's risk of being abused or neglected [21].

Although machine learning algorithms are adopted and celebrated for their accuracy in predicting policy outcomes [42, 43], in practice algorithms typically assist people in making informed yet ultimately normative decisions. Risk predictions and policy decisions represent distinct tasks: unlike predictions of risk, which can be directly optimized for accuracy, many policy decisions require balancing competing goals and therefore lack a straightforward correct answer [76].

The particular balancing act for any policy decision is of normative significance and is often subject to vigorous debate. How frontline government officials weigh competing values when making decisions effectively determines the implementation of public policy [49]. This makes it imperative that government decision-makers strike the appropriate balance between the conflicting goals embedded in public policy [76]. In settings such as pretrial detention, "[h]ow this balance is struck [...] has enormous implications" [51].

The distinction between predictions and decisions means that using risk assessments to improve people's predictions may not necessarily improve people's decisions. In particular, the normative multidimensionality inherent in many government decisions can create conflicts between risk-reduction and other values. For instance, although pretrial decisions must limit the risk of defendants being rearrested or not returning to court for trial, they must also prioritize the liberty of defendants [2]. Similarly, although government loan decisions must limit the risk of recipients defaulting on their loans, they must also promote equity by supporting low-income applicants [69].

Thoroughly evaluating the impacts of risk assessments therefore requires considering not just whether these algorithms improve the accuracy of human *predictions*, but also whether they improve the quality of human *decisions*. One notable concern regarding risk assessments is that these tools could alter how people balance risk with other considerations when making decisions [25, 62]. Although improving the accuracy of risk predictions is appropriate for policies that include risk as a consideration, any systematic change in the salience of risk in decision-making would amount to a shift in public policy, with potentially disparate impacts. As a result, more than 100 civil rights and social justice organizations have raised concerns that the emphasis on risk by pretrial risk assessments will prompt judges to treat defendants more harshly [66].

Policymakers and others advocating for risk assessments assert that these algorithms merely improve human predictions without altering how people factor risk into their decisions [21, 37, 55, 62, 73]. Proponents claim that as long as decision-makers are granted autonomy and discretion over final decisions, providing public servants with accurate risk predictions will lead these individuals to make better (e.g., fairer and less punitive) decisions [21, 37, 55, 62, 73]. Yet despite being central to arguments for adopting public sector algorithms, this assumption has not been rigorously tested. In fact, this claim is called into question by mounting evidence that the implementation of algorithms often relies on untested—and false—assumptions about human-algorithm collaborations, leading to unjust outcomes [30]. For instance, criminal justice risk assessments have failed to generate the intended benefits because judges use these tools in unexpected ways [1, 6, 64, 65].

This paper tests the assumption that improving human prediction accuracy with algorithms will necessarily improve human decisions. In particular, we study how presenting a risk assessment's advice influences the decision-making processes of laypeople. We ran an online experiment with 2,140 U.S.-based participants recruited from Amazon Mechanical Turk (a widely used online platform for human subjects research [13, 45]). We used this experiment to explore the influence of risk assessments in two high-stakes government settings where decision-making involves balancing risk-reduction with other factors: a pretrial setting and a home improvement loans setting.

We had three central goals for the study: 1) determine whether risk assessments merely improve human risk predictions, as is commonly asserted, or also alter how people weigh risk in the decision-making process itself; 2) characterize the effects of risk assessments on decision-making processes; and 3) determine how these effects impact outcomes such as racial disparities. We hypothesized that presenting risk assessments would alter decision-making processes, prompting participants to become more attentive to avoiding risk when making decisions. We also hypothesized that this effect would exacerbate racial disparities in decisions.

We found that risk assessments altered human decision-making processes in both settings. Presenting risk assessments to participants systematically changed how they factored risk into policy-relevant decisions in ways that can lead to harmful outcomes. Although the risk assessments improved the accuracy of human predictions, they also induced shifts in human decision-making processes that counteracted the potential benefits of these enhanced predictions. In the pretrial setting, the risk assessment made participants more sensitive to increases in perceived risk, increasing the racial disparity in pretrial detention by 1.9%. In the loans setting, the risk assessment made participants more risk-averse at all levels of perceived risk, reducing government aid by 8.3%.

These results challenge a central assumption behind support for algorithmic decision-making aids in government. They demonstrate that improving human prediction accuracy with risk assessments does not necessarily improve human decisions and instead can have unexpected adverse consequences. If these observed behaviors arose in practice, presenting algorithms to government decision-makers would generate unintended and unjust shifts in the application of public policy without being subject to democratic deliberation or oversight.

Because we study laypeople in a lab setting, our results do not reflect the behaviors of experts making real decisions. Practitioners differ from laypeople in numerous ways, as their specialized knowledge and professional identity shape their responses to risk assessments [6]. However, there are several reasons to believe that our results could align with real-world outcomes and complement studies of expert decision-making in practice. First, research has found that both judges and financial professionals exhibit many of the same behaviors as laypeople. Judges are susceptible to cognitive and racial biases when making decisions in much the same manner as laypeople [33, 57, 58]. Similarly, financial professionals are susceptible to priming effects and loss aversion [12, 23, 34]. Financial professionals exhibit these behaviors to a greater extent than laypeople [23, 34]; furthermore, greater professional experience among financial professionals does not mitigate priming effects [12]. Second, experimental studies using procedures very similar to those in this study [28, 29] have found behaviors among laypeople using risk assessments that align closely with behaviors observed among judges using risk assessments in practice [1, 14].

Experimental trials with laypeople therefore present a promising approach for evaluating and improving proposed human-algorithm collaborations before algorithmic decision-making aids are adopted. Performing these experimental studies would provide diagnostic knowledge that can inform experimental studies with experts and, in turn, the development, implementation, and evaluation of real-world systems. Although the gold standard is data on how experts use risk assessments in practice, such evidence relies on retrospective data about systems that have been in use for years [6, 64, 65, 72]. By the time breakdowns in real-world human-algorithm collaborations are exposed, many people will have already been affected. Although some adverse behaviors would arise only in practice, many could potentially be detected before an algorithm's adoption via preliminary lab studies. Experimental trials can serve as an integral component of a broader evaluation pipeline, facilitating more rigorous and proactive scrutiny of whether algorithms actually improve decision-making.

2 BACKGROUND AND RELATED WORK

The increasing use of algorithmic decision-making aids across government has placed novel human-algorithm collaborations at the center of consequential policy decisions. In the context of machine learning, the standard framework for human-algorithm collaborations is "human-in-the-loop" systems. In these settings, people are incorporated into the machine learning pipeline to produce the best possible algorithmic output. Algorithms make the final decisions, with humans assisting (e.g., labeling training data and reviewing low-confidence classifications) [3]. In some cases, such tasks are completed by crowds of people, with crowdsourcing techniques used to support machine learning models [39, 71].

In public policy contexts, however, human-algorithm collaborations involve a different process and goal: algorithms are incorporated into human decision-making processes to generate the best possible human decision. People make the final decisions, with algorithms assisting (e.g., making accurate predictions based on patterns within datasets). Instead of the human-in-the-loop frame, therefore, most uses of algorithms in government call for a complementary paradigm: "algorithm-in-the-loop" decision-making [28]. Algorithm-in-the-loop settings center human decisions—rather than algorithmic decisions—as the most important outcome, orienting attention to how algorithms influence human decisions.

Despite the potential promise of algorithms aiding human predictions, experimental studies have uncovered numerous limits in people's ability to make appropriate and effective use of algorithmic advice. Several studies have found that algorithmic advice can improve the accuracy of human predictions, but people's decisions about when and how to diverge from algorithmic recommendations are typically incorrect [28, 29, 32, 46]. People struggle to evaluate the quality of algorithmic advice [24, 28, 29, 46], often discount accurate algorithmic recommendations [18, 48, 75], and exhibit racial biases in their responses to risk assessments [28, 29]. Although evidence suggests that experts are capable of overriding some erroneous predictions in practice [15], other evidence demonstrates that incorrect predictions reduce the quality of expert judgments [40] and that experts make less effective use of algorithmic forecasts than laypeople [50].

These breakdowns in human-algorithm collaboration demonstrate that algorithmic interventions are indeterminate: the effects of using an algorithm often diverge in practice from what was expected based on the algorithm's technical characteristics [31]. Numerous evaluations of how judges use pretrial risk assessments indicate that providing accurate risk predictions does not generate the intended improvements in decision-making (i.e., reduced pretrial detention, recidivism, and racial disparities). In jurisdictions across the country, judges disproportionately override release recommendations to detain defendants, leading to much higher than expected pretrial detention rates [60, 63–65, 72]. Several studies have shown that risk assessments exacerbate rather than diminish racial disparities in pretrial detention, in part because judges often make more punitive decisions about Black defendants than similar white defendants [1, 14, 64]. Furthermore, ethnographic work has found that judges often resist using risk assessments because they dislike the idea of these tools replacing or surveilling them [6].

Despite growing knowledge about how people use algorithms in both experimental and real-world settings, a significant open question is whether presenting algorithmic advice alters the process through which people make decisions. Because risk assessments emphasize the likelihood of specific adverse outcomes (such as a pretrial defendant failing to appear for trial or being rearrested), scholars and activists have raised concerns that risk assessments could make risk a more salient factor in decision-making processes [25, 62, 66]. Such concerns follow closely from the existing literature on framing and priming effects. Prior research demonstrates that framing decisions around losses motivates decision-makers (including judges) to avoid those losses [58, 59, 67]. Similarly,

priming people (including financial professionals) to consider risks makes them less likely to make or support decisions that involve risk [12, 19, 20, 23].

Initial studies have observed behaviors that are consistent with the possibility of risk assessments altering decision-making processes. A small experiment with 83 law students making simulated sentencing decisions found that presenting a risk assessment increased the sentence given to a high-risk defendant and decreased the sentence given to a low-risk defendant [62]. A later experiment with 340 judges making simulated sentencing decisions found that presenting a risk assessment increased the likelihood that a low socioeconomic status defendant would be incarcerated and decreased the likelihood that a high socioeconomic status defendant would be incarcerated [61].

Although these results suggest that showing a risk assessment heightens the salience of risk, they cannot conclusively determine whether risk assessments alter how people weigh risk when making decisions. Because these studies looked only at people's decisions with and without a risk assessment, neither can distinguish between two potential explanations: a) the risk assessment actually increased the weight that people gave to reducing risk when making decisions, and b) the risk assessment merely influenced the estimates of risk that people factored into their decisions. Distinguishing between these two explanations requires accounting for a risk assessment's effects on human risk predictions rather than directly comparing human decisions with and without a risk assessment.

3 HYPOTHESES AND FRAMEWORK

In this study, we investigate whether risk assessments systematically alter how people factor risk into their decisions. Because risk assessments emphasize the risk of an adverse outcome, we hypothesized that presenting risk assessments would make people more attentive to avoiding risk when making decisions. Given existing racial disparities in risk, we also hypothesized that this effect would exacerbate racial disparities in decisions.

Based on how policy documents [55] and court rulings [37, 73] describe the use of risk assessments, we analyzed decisions as being made through a two-stage process (Figure 1). First is the risk-prediction process (RPP), which represents a quantitative prediction task. The RPP evaluates the attributes of a given subject (e.g., pretrial defendant) to predict that person's risk of an adverse outcome (e.g., failing to return to court for trial or being arrested before trial). This process yields a decision-maker's "perceived risk" about the subject. Second is the decision-making process (DMP), which involves a normative balancing act between numerous considerations rather than a straightforward translation of risk into a decision. The DMP incorporates the perceived risk alongside other relevant factors (e.g., the harms associated with pretrial detention) to make a decision about the subject (e.g., whether to release the defendant before their trial). A systematic change to the DMP reflects a shift in how decision-makers balance risk with other factors, which amounts to a shift in public policy [51, 76]. We instantiated the DMP as a function that determines the probability of detaining a defendant or rejecting a loan applicant conditioned on the perceived risk about the subject in question.[1]

The central question of this study is whether risk assessments influence only the RPP, as is typically assumed, or instead affect both the RPP *and* the DMP. We categorize the influence of risk assessments into four possible "scenarios," as summarized in Table 1. Scenario 1 represents the baseline condition without any risk assessment. When a risk assessment's advice is presented, it could lead to either Scenario 3 or Scenario 4.[2] Scenario 3 is the commonly assumed outcome: risk

[1]From a decision-making standpoint, what matters as an input to the DMP is the decision-maker's perceived risk, as that is what decision-makers act on. The perceived risk can be influenced by the risk assessment, among other factors.

[2]Scenario 2, in which risk assessments alter the DMP but not the RPP, is ruled out by prior research demonstrating that risk assessments influence human predictions [28, 29, 32, 65].

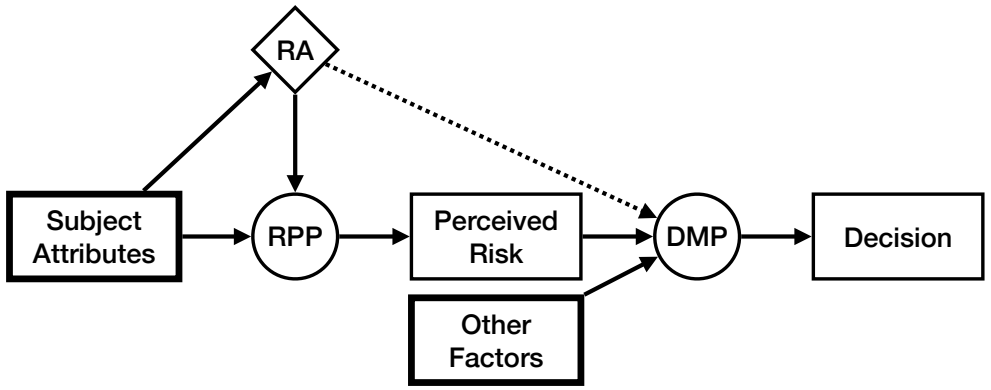

Fig. 1. How subject attributes are translated into a decision with the aid of a risk assessment, as conceptualized in law and policy. Circles represent the two stages of human cognitive processing: the risk-prediction process (RPP) and the decision-making process (DMP). The dashed line from the risk assessment (RA) to the DMP represents the key question of this study: whether the RA alters the DMP. The absence of this influence (i.e., the absence of the dashed line) represents Scenario 3. The presence of this influence (i.e., the presence of the dashed line) represents Scenario 4. Bold lines indicate that the rectangle represents a set of multiple attributes or factors.

Table 1. The four possible "scenarios" of how a risk assessment (RA) can affect the risk-prediction process (RPP) and the decision-making process (DMP). Scenario 1 represents a baseline process without a risk assessment. Scenarios 2–4 represent the possible outcomes when decision-makers are presented with a risk assessment (Scenario 2 is ruled out by prior research, however).

	DMP unaffected by RA	**DMP affected by RA**
RPP unaffected by RA	Scenario 1 (Baseline: RA does not affect RPP or DMP)	Scenario 2 (Implausible: RA affects only DMP)
RPP affected by RA	Scenario 3 (Common assumption: RA affects only RPP)	Scenario 4 (Hypothesis: RA affects both RPP and DMP)

assessments alter the RPP but not the DMP, meaning that improvements in prediction accuracy lead directly to more informed decisions. The assumption that algorithms lead to Scenario 3 is central to support for algorithmic decision-making aids in government [21, 37, 55, 62, 73]. In this scenario, which represents the absence of the dashed line in Figure 1, improving prediction accuracy with risk assessments directly improves decisions. Scenario 4 is our hypothesized outcome: risk assessments alter both the RPP and the DMP, meaning that shifts in the DMP could counteract any gains in prediction accuracy. In this scenario, which represents the presence of the dashed line in Figure 1, improving prediction accuracy with risk assessments may not improve decisions.

Our goal is to test whether showing a risk assessment alters the DMP, which amounts to distinguishing between Scenario 3 and Scenario 4. However, we cannot directly observe the DMP because it is a latent form of cognitive processing. Because risk assessments alter the RPP [28, 29, 32, 65], simply showing that a risk assessment changed participant decisions is insufficient to demonstrate that the risk assessment affected the DMP. The challenge arises because a risk assessment could alter decisions in two distinct but superficially indistinguishable ways. First, it

could influence the RPP alone (Scenario 3), which would lead to decisions based on different risk estimates without changing how perceived risk factors into decisions. Second, it could influence both the RPP and the DMP (Scenario 4), which would lead to decisions based on different risk estimates *and* would change how perceived risk factors into decisions.

The only way to determine whether risk assessments affect the DMP is to compare decisions made with and without a risk assessment *while accounting for the risk assessment's effects on predictions*. Accomplishing this requires access to decision-makers' perceptions of risk about each subject. However, this information is not produced in practice and is difficult to obtain experimentally without influencing people's behavior. Determining the risk assessment's effects on the DMP thus requires a more complex experimental setup than prior work that has directly compared the decisions that people make with and without a risk assessment's advice. As described in more detail below, we designed our experiment to elicit both decisions and predictions from participants, enabling us to infer the effects of risk assessments on human decision-making processes.

4 METHODS

Our study progressed in two stages. The first stage involved developing risk assessments for pretrial detention and home improvement loans. The second stage involved running an experiment on Amazon Mechanical Turk to evaluate how people interact with these risk assessments when making predictions and decisions. The full study was approved by the Harvard University Institutional Review Board and the National Archive of Criminal Justice Data (which manages the data used for the pretrial setting).

4.1 Study Settings

Our experiment simulated two settings of government decision-making: pretrial detention and government home improvement loans. Within the context of this study, these settings are structurally similar. In both settings, decision-makers must balance risk-reduction with conflicting normative considerations. Decisions are made about an individual "subject" and can be "positive decisions" or "negative decisions." Subjects with high risk are more likely to receive negative decisions. In the pretrial setting, the subject of the decision is a criminal defendant, the positive decision is to release the defendant before trial, and the negative decision is to detain the defendant before trial. In the loans setting, the subject of the decision is a loan applicant, the positive decision is to approve the loan application, and the negative decision is to reject the loan application.

4.1.1 Pretrial Detention Setting. After someone is arrested in the United States, they must await trial. Courts can either hold the criminal defendant in jail until their trial or release them with a mandate to return for their trial.[3] Pretrial detention decisions involve balancing competing goals. Courts aim to ensure that defendants will return to court for trial and will not commit any crimes if released. The higher the risk that a defendant will fail to return to court for their trial or will commit any crimes, the more likely a judge is to detain the defendant until their trial. The interest in reducing risk is enhanced by detaining defendants. However, pretrial decisions are also made with an interest in protecting the liberty of defendants, ensuring that defendants are able to mount a proper legal defense, and reducing the hardship to defendants and their families [2]. Pretrial detention is associated with a range of negative outcomes that include longer prison sentences, sexual abuse, and limited employment opportunities [27]. The interests in protecting liberty and avoiding the harms of pretrial detention are advanced by releasing defendants.

[3]In practice, many defendants are also released under conditions such as paying a cash bond or being subject to electronic monitoring.

In recent years, many jurisdictions across the U.S. have turned to risk assessments as a tool to make more accurate and objective predictions of risk. These improvements in prediction are intended to reduce racial biases and increase pretrial release rates [35, 38, 55].

4.1.2 Government Home Improvement Loans Setting. Many people apply for a loan to improve their house (e.g., to rehabilitate a home or to make a home energy efficient). When someone applies for a loan, it is common for the lender to assess the risk that the borrower will fail to pay back the money. This is known as defaulting on the loan. The higher the risk that the potential borrower will default on the loan, the less likely the lender generally is to provide money to that person. The U.S. government provides many types of home improvement loans in order to support low-income applicants who are unable to obtain affordable loans from banks [69]. This sets up a balancing act between conflicting aims. On the one hand, the goal of limiting loan default risk is enhanced by declining loans to low-income applicants. On the other hand, the goals of promoting equity, economic development, and community stability are enhanced by providing loans to low-income applicants.

It is common for lenders to evaluate loan applicants using risk assessments that predict the likelihood of loan default. Although there are no known cases of governments using risk assessments when allocating home improvement loans, this setting is akin to government uses of risk assessments to determine who should receive other resources [21].

4.2 Risk Assessments

In order to test the effects of presenting risk assessment predictions to participants in our experiment, we first developed risk assessments for pretrial detention and government home improvement loans. See Section A of the Appendix for a more detailed description of the data we used and how we developed these models. Our goal in this stage was not to develop optimal risk assessments, but to develop risk assessments that resemble those used in practice and that could be presented to participants during the Mechanical Turk experiment. We used datasets with information about 47,141 felony defendants across the United States who had been released before trial [68] (Table A.1) and 45,218 recipients of home improvement loans via the peer-to-peer lending company Lending Club (Table A.2). The data included demographic information (including race, which we restricted to Black and white) for the felony defendants but not the loan applicants.

We developed risk assessments (i.e., machine learning classifiers) using gradient boosted trees with ten-fold cross-validation. Our models included five attributes of each defendant[4] and seven attributes of each loan application.[5] The pretrial risk assessment was trained to predict whether a defendant, if released before trial, would fail to appear in court for trial or would be arrested before trial. The loans risk assessment was trained to predict whether a loan applicant, if given the loan, would default on that loan. Both risk assessments exhibited similar accuracy to pretrial and loan risk assessments developed in research and practice (pretrial AUC=0.67, loans AUC=0.69). Drawing from the held-out validation sets in each setting, we selected samples of 300 defendants and 300 loan applicants whose profiles and risk predictions would be presented to participants during the experiment. When used in our experiment, the risk assessments presented numerical predictions of risk about subjects (i.e., 0%–100%, in intervals of 10%) but did not suggest what decision participants should make based on those predictions.

[4]Defendant factors: age, offense type, number of prior arrests, number of prior convictions, and whether that person has any prior failures to appear for trial.

[5]Loan application factors: the applicant's annual income, credit score, and home ownership, as well as the loan's value, interest rate, monthly installment, and term of repayment.

Table 2. Attributes of the participants in our experiment, by setting. Measures of familiarity, clarity, and enjoyment are based on participant self-reports measured on a Likert scale from 1 (low) to 7 (high).

	Pretrial N=1,040	**Loans** N=1,100
Demographics		
Male	59.8%	61.0%
Black	14.2%	11.9%
White	71.5%	72.9%
18-24 years old	7.4%	6.8%
25-34 years old	46.1%	45.0%
35-59 years old	43.0%	43.9%
60+ years old	3.6%	4.3%
College degree or higher	82.5%	81.9%
Criminal justice familiarity	5.1	5.1
Financial lending familiarity	4.9	5.1
Machine learning familiarity	4.7	4.8
Outcomes		
Average experiment time	19.1 minutes	19.0 minutes
Average hourly wage	$14.86	$15.16
Experiment clarity	6.4	6.4
Participant enjoyment	5.8	5.9

4.3 Experimental Design

We recruited 2,685 participants on Amazon Mechanical Turk over two weeks in May 2020, restricting our task to workers inside the United States who had a task approval rate of at least 75%.[6] Our analysis includes the results from the 2,140 participants who completed the experiment while also passing our quality control reviews (by correctly answering several comprehension questions and two attention-check questions). Across both settings, a majority of participants were male, white, and college graduates (Table 2). Participants were paid $3 for completing the experiment, and those making predictions received an additional payment of up to $1 based on the accuracy of their predictions. Bonus payments were allocated using a Brier score, which incentivizes participants to report their true estimate of risk. Participants completed the experiment in an average of 19.0 minutes and received an average wage of $15.02 per hour.

The experimental design consisted of three treatments. When participants entered the experiment, they were split evenly into either the pretrial or loans setting. We then followed a 2x2 design within each setting (Figure 2). Participants were split into control groups (which were not presented with a risk assessment's advice) and treatment groups (which were presented with a risk assessment's advice). Participants were also split into prediction groups (which were asked to make quantitative predictions about subjects) and decision groups (which were asked to make binary decisions about subjects). This design elicits sufficient information to determine whether the risk assessments altered the DMP and is described below in more detail.

[6] In light of the COVID-19 pandemic, immediately before running this experiment we replicated a trial experiment conducted in December 2019. Our tests demonstrated that the pandemic did not influence any of our results (see Section B in the Appendix).

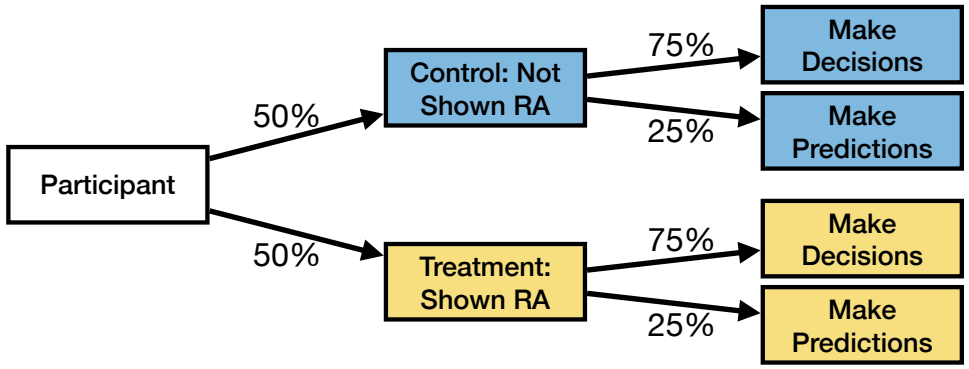

Fig. 2. Our 2x2 experimental design and the four conditions that participants were sorted into in each setting. Probabilities indicate the likelihood of each path. Within each setting, participants in all four conditions were presented with a sample of subjects drawn from the same set of 300 subjects.

The experimental procedure was the same in both the pretrial and the loans settings. After completing a consent page, participants entered a tutorial that described their setting and the predictions or decisions that they would be asked to make. These descriptions explained the key considerations (including but not limited to risk) that factor into decisions in the relevant setting. Participants who would be shown the risk assessment's advice were also presented with background information about the risk assessment. The description of the risk assessment included details about the algorithm's prediction task, training data, and accuracy, and invited participants to use the predictions in whatever manner they desired. Participants were unable to proceed beyond the tutorial until they correctly answered several questions demonstrating their comprehension. We ignored all data from participants who required more than four attempts to correctly answer all of the comprehension questions. Participants then completed an intro survey (to provide demographic information and other attributes), a prediction or decision task (described in detail below), and an exit survey (to provide reflections on the task).

The key component of the experiment was the prediction or decision task (Figure 3). Based on their assigned setting, participants were presented with narrative profiles describing seven features about defendants or applicants.[7] These defendants and applicants were drawn randomly from the assigned setting's 300-subject sample. Participants were tasked with making either numeric predictions of risk about 40 subjects or binary decisions about 30 subjects.[8] Prediction-makers were asked to predict risk on a scale from 0% to 100%, with options in 10% increments. Decisions in the pretrial setting entailed whether to release or detain criminal defendants before trial. Decisions in the loans setting entailed whether to approve or reject home improvement loan applications. This setup matches salient elements of real-world settings such as pretrial adjudication, in which risk assessments are introduced as important decision-making aids [27, 64] and in which decisions are often made in just a few minutes [5, 60].

The primary goal of our experiment was to determine the effects of the risk assessments on human decision-making processes. This requires comparing the decisions of participants with and without a risk assessment while accounting for the risk assessment's effects on predictions. We

[7]These features were the same as those used to develop the risk assessment in each setting, with the addition of race and gender in the pretrial setting.

[8]Participants making predictions were given a larger number of subjects because they received a bonus payment of up to $1 in addition to the $3 base payment received by all participants.

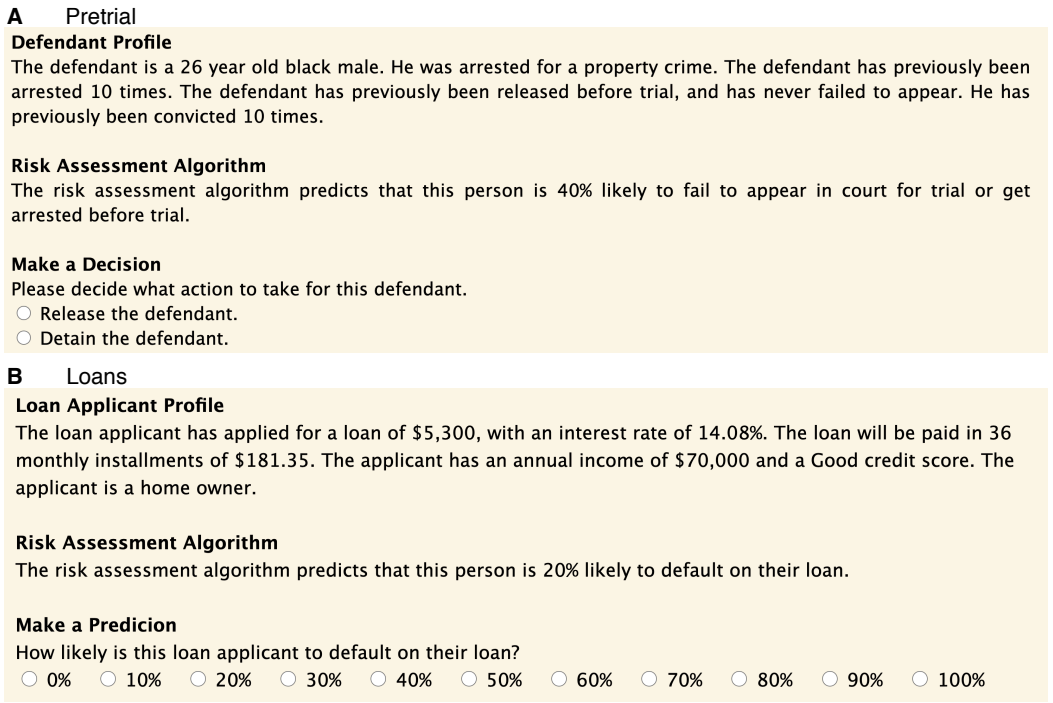

Fig. 3. Examples of the prompts presented to participants. (A) A profile presented to a decision-making participant in the pretrial setting. (B) A profile presented to a prediction-making participant in the loans setting. Both of these examples are for participants in the treatment group; participants in the control group saw the same prompt, but without the section about the risk assessment.

therefore followed a 2x2 experimental setup within each setting, splitting participants according to whether they are presented with the risk assessment and whether they make binary decisions or quantitative risk predictions (Figure 2).

Our first experimental condition in each setting was whether or not participants were presented with the predictions of a risk assessment. Participants in the control group were shown only the narrative profiles about subjects. Participants in the treatment group were shown the narrative profiles as well as the risk assessment's predictions about subjects (Figure 3). This first condition allows us to compare the behaviors of participants with and without the risk assessment.

However, directly comparing the decisions of the control and treatment groups cannot determine whether the risk assessment altered the DMP. Decisions could differ across the control and treatment groups because the risk assessment influenced the RPP but not the DMP. For instance, a risk assessment could increase the likelihood of a defendant being detained before trial by a) making decision-makers more risk-averse or b) causing decision-makers to increase their estimate of the defendant's risk. Determining a risk assessment's influence on the DMP therefore requires accounting for the risk assessment's influence on predictions. This means that we must obtain information regarding participants' perceived risk about subjects in addition to their decisions about subjects.

We could obtain information about the risks perceived by decision-makers in two ways. The first approach is to ask each participant to make both predictions and decisions about subjects. This approach would provide the most accurate measure of the perceived risk associated with each

decision. However, because this approach requires directly asking decision-making participants about risk, it would also prime them to consider risk whether or not they are shown the risk assessment. This priming would undermine the entire study by confounding our ability to detect how presenting a risk assessment influences the consideration of risk in the DMP. The second approach to measuring perceived risk—which we take in this study—is to have some participants make risk predictions and some participants make decisions. We use the risk predictions provided by prediction-makers to estimate the risks perceived by decision-makers. Although this approach means that we cannot directly measure decision-making participants' perceptions of risk, it provides a reasonable proxy while maintaining the integrity of our research question.

Our second experimental condition, therefore, was whether participants were asked to make predictions or decisions. To obtain risk estimates about each subject (both with and without a risk assessment), we asked 75% of participants to make binary decisions about subjects and 25% to make numerical predictions of risk about each subject (Figure 2).[9] We used the risk predictions elicited from the prediction-making participants to estimate the risks perceived by the decision-making participants. We estimated the perceived risk associated with a given decision as the average risk prediction made about the subject in question, grouping predictions and decisions based on whether the risk assessment was shown. For instance, the perceived risk assigned to a decision about a defendant made without the risk assessment was the average of the risk predictions made about that same defendant without the risk assessment. By eliciting many predictions about each subject, we obtained reliable measures of the average perceived risk about each subject (both with and without the risk assessment) without inappropriately influencing the behaviors of decision-making participants.

4.4 Analysis

To study whether and how a risk assessment alters the decision-making process, we modeled the DMP of participants with and without a risk assessment. We characterized negative decisions as a function of perceived risk and conducted Bayesian mixed-effects logistic regressions to learn this function.[10] Following the decision-making structure in Figure 1, we regressed participant decisions on three factors: the perceived risk about the subject in question, whether the risk assessment was shown, and the interaction between these two factors. Factors such as subject attributes and the risk assessment's prediction are incorporated into this decision function through *perceived.risk*, which is based on these elements. We also included three random effects to account for repeated samples in the data.

$$\begin{aligned} \text{negative.decision} \sim\ & \text{perceived.risk} + \text{show.RA} + \text{perceived.risk} * \text{show.RA} \\ & + (1|\text{participant}) + (1|\text{subject}) + (1|\text{progress.index}) \end{aligned} \tag{1}$$

This regression is structured to infer the DMP that participants followed and to determine whether the risk assessment altered this function, thus distinguishing between Scenario 3 and Scenario 4. If risk assessments present information that improves the RPP but does not influence the DMP (Scenario 3), we would expect to see that showing the risk assessment does not alter this regression. In this case, neither regression factor that includes *show.RA* would be significant, such that the relationship between decisions and perceived risk is the same whether or not the risk assessment is shown. However, if risk assessments influence the DMP as hypothesized (Scenario 4),

[9] We placed more participants into the decisions treatment because our analysis required more decisions than predictions to obtain robust results. The disparity in participants making decisions versus predictions is partially offset by the fact that participants making predictions evaluated more subjects (40) than participants making decisions (30).

[10] We used a Bayesian approach with weak priors to enable analyses based on posteriors. In all cases, the inferences made with Bayesian and non-Bayesian regressions were almost identical.

we would expect to see that showing the risk assessment alters this regression, making people more attentive to reducing risk when making decisions. This result could emerge through two different mechanisms: 1) the risk assessment makes participants more risk-averse at all levels of risk (in this case, the *show.RA* coefficient would be positive), or 2) the risk assessment makes participants more sensitive to increases in risk (in this case, the $perceived.risk * show.RA$ coefficient would be positive).

After observing that the risk assessments influenced the DMP (and thus generated Scenario 4 as hypothesized), we estimated the impacts of this influence. Our goal was to isolate the effects of the DMP change, controlling for the risk assessments' effects on the RPP. This analysis entailed comparing outcomes from the observed Scenario 4 behaviors with outcomes from the commonly expected Scenario 3 behaviors. Because our control group participants exhibited Scenario 1 and our treatment group participants exhibited Scenario 4, we did not observe Scenario 3 behaviors and could not directly compare Scenario 3 and Scenario 4 outcomes. We therefore estimated the differences between Scenario 3 and Scenario 4 outcomes through simulations. We began by fitting models for the RPP and DMP in the pretrial and loans settings, both with and without the risk assessment's advice. We then ran 1,000 trials simulating the outcomes for more than 4,000 defendants and loan applicants in the four scenarios described in Table 1.

See Section C of the Appendix for additional details about our analyses.

5 RESULTS

5.1 Effects of the Risk Assessments on the Risk-Prediction Process

We looked first at how the risk assessments affected predictions of risk. We evaluated participant "prediction quality" using a reverse Brier score bounded between 0 (worst possible performance) and 1 (best possible performance). In both settings, presenting the risk assessment reduced estimates of risk, improved prediction accuracy, and aligned the RPP more closely with the risk assessment's calculations. These results are consistent with prior work [28, 29].

In the pretrial setting, the risk assessment reduced perceived risk for 54.0% of defendants. Overall, defendants received an average reduction in perceived risk of 1.6% (from 40.6% to 38.9%, P=.001, d=0.19). While the reduction in perceived risk was significant for white defendants (38.4% to 35.7%, P=.003, d=0.30), Black defendants received a smaller and nonsignificant reduction (41.7% to 40.7%, P=.085, d=0.12). Bayesian linear regression (Equation A.1) found that showing the risk assessment altered the risk-prediction process, most notably prompting participants to consider the age of defendants and to reduce the risk associated with violent crime and prior failures to appear (Table A.3). Through these changes, presenting the risk assessment increased the average participant prediction quality from 0.72 to 0.75 (P<.001, d=0.11).

In the loans setting, the risk assessment altered predictions of risk more dramatically. The risk assessment reduced the perceived risk for 92.3% of loan applicants and generated an overall average reduction of 14.2% for each applicant (from 38.5% to 24.3%, P<.001, d=1.54). Bayesian linear regression (Equation A.2) found that showing the risk assessment altered the RPP by significantly reducing participants' baseline risk predictions, increasing the salience of annual income and interest rate, and prompting participants to consider the length of loans (Table A.3). In turn, showing the risk assessment increased participant prediction quality from 0.75 to 0.83 (P<.001, d=0.31).

5.2 Effects of the Risk Assessments on the Decision-Making Process

We next analyzed how the risk assessments affected participant decisions and decision-making processes.

5.2.1 Effects on Decisions. We first compared participant decisions with and without a risk assessment. Our goal was to investigate whether the shifts in decisions induced by the risk assessments align with the shifts in predictions induced by the risk assessments. If risk assessments lead to Scenario 3, we would expect to see that shifts in decisions closely track the shifts in predictions described above. In particular, reductions in perceived risk due to the risk assessment would be associated with reductions in negative decisions due to the risk assessment, and vice versa. Our results do not closely follow this pattern, however, indicating that the risk assessment's effects on decisions cannot be explained by shifts in the RPP alone.

In the pretrial setting, the risk assessment reduced pretrial detention rates but increased racial disparities. The risk assessment reduced each defendant's likelihood of pretrial detention by an average of 2.4% (from 44.5% to 42.1%, P<.001, d=0.21). White defendants received a 27% larger average reduction (38.7% to 35.9%, P=.014, d=0.24) than Black defendants (47.7% to 45.5%, P=.007, d=0.20). As a result, the overall racial disparity in pretrial detention increased by 18.8% from 8.8% to 10.4%. In addition, the risk assessment increased the "accuracy" of decisions from 56.7% to 58.4% (P=0.009, h=0.03) and reduced the "false positive rate" from 26.9% to 24.4% (P<.001, h=0.06).[11]

In the loans setting, the risk assessment's effects on negative decisions contrasted with the risk assessment's effects on perceived risk. Although the risk assessment dramatically reduced risk predictions, the risk assessment did not significantly alter each loan applicant's likelihood of rejection (loan rejection rates went from 22.1% to 23.1%, P=.159, d=0.08). Furthermore, although the risk assessment significantly increased the accuracy of risk predictions, the risk assessment reduced the "accuracy" of decisions from 72.2% to 70.5% (P=.002, h=0.04) and increased the "false positive rate" from 17.3% to 18.5% (P=.015, h=0.03). In sum, the risk assessment *notably reduced* perceived risk yet *did not reduce* rejection rates, and similarly *increased* prediction accuracy yet *decreased* decision "accuracy."

To further investigate the relationship between predictions and decisions, we then compared how the risk assessments altered the predictions and decisions made about each individual subject. We found that shifts in perceived risk did not translate to equivalent shifts in negative decisions (Figure 4). In both settings, subjects for whom the risk assessment decreased perceived risk did not reliably receive lower negative decision rates due to the risk assessment. Among the 54.0% of pretrial defendants for whom the risk assessment reduced perceived risk, only 59.3% received a reduced likelihood of pretrial detention when the risk assessment was shown. Among the 92.3% of loan applicants for whom the risk assessment reduced perceived risk, only 52.0% received a reduced likelihood of rejection when the risk assessment was shown. Overall, shifts in decisions were relatively insensitive to shifts in predictions, with regression coefficients less than 1 in both settings (0.23 in pretrial, P=.003; 0.42 in loans, P<.001). For instance, a 10% reduction in average perceived risk due to the risk assessment was associated with a 4.4% reduction in the pretrial detention rate and a 2.8% *increase* in the loan rejection rate.

These results depart notably from what we would expect to see if the risk assessments induced a shift to Scenario 3. These patterns demonstrate that reductions in perceived risk do not lead directly to reductions in pretrial detention or loan application rejections. Instead, these changes in perceived risk must be mediated through changes to the DMP before yielding decisions.

[11] These measures are described in quotes to reflect that decisions were made with risk as just one of several factors and that it is a simplification to evaluate decisions as if they were predictions of risk.

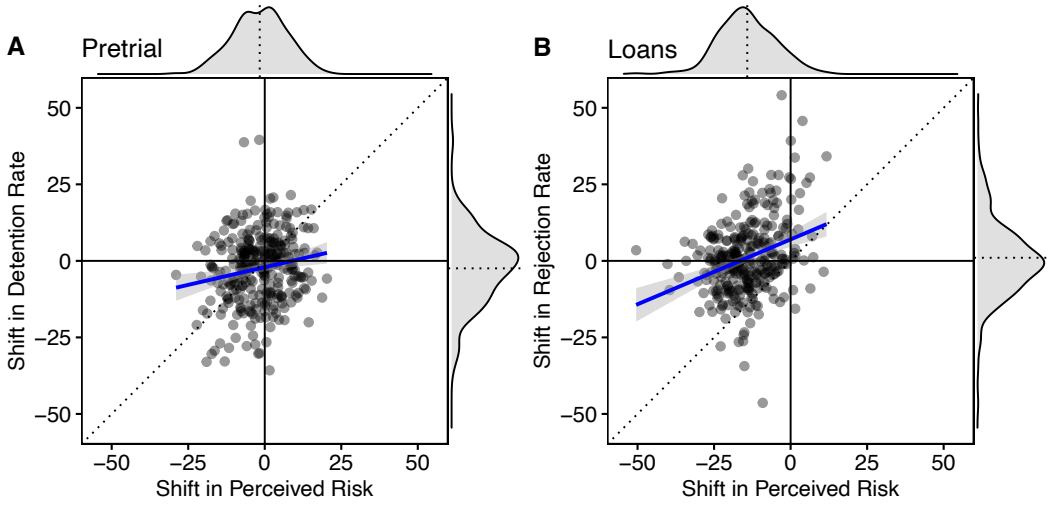

Fig. 4. Shifts in perceived risk and negative decision rates for each subject caused by showing the risk assessment to participants. (A) Pretrial setting. (B) Loans setting. Each point represents a single defendant or applicant, with marginal density plots along each axis (in which the dotted lines represent the average values). Positive values on the x-axis indicate that the risk assessment increased the average risk prediction about a subject. Positive values on the y-axis indicate that the risk assessment increased the negative decision rate for a subject. Blue lines represent linear regression fits of shifts in negative decisions versus shifts in perceived risk. These results indicate that decisions are relatively insensitive to shifts in predictions and that reductions in perceived risk do not necessarily lead to reductions in negative decisions.

5.2.2 Effects on the Decision-Making Process. We next analyzed the risk assessments' effects on the DMP. Bayesian mixed-effects logistic regressions (Equation 1) found that the risk assessment altered the decision-making process in both settings, making participants more attentive to risk when making decisions. These results demonstrate that the risk assessments prompted the hypothesized shift to Scenario 4 rather than Scenario 3.

In the pretrial setting, the risk assessment made participants more sensitive to increases in risk (Figure 5). Presenting the risk assessment increased the odds ratio associated with a 10% increase in perceived risk from 1.82 to 2.39 (Table 3). These results mean that the risk assessment made perceived risk a stronger determinant of whether defendants were released or detained: the risk assessment reduced pretrial detention rates for defendants with low perceived risk and increased pretrial detention rates for defendants with high perceived risk. For example, the risk assessment reduces the detention likelihood by 6.3% for a defendant with a perceived risk of 30% but increases the detention likelihood by 8.7% for a defendant with a perceived risk of 60% (Table A.4).

In the loans setting, the risk assessment made participants more risk-averse at all levels of risk (Figure 5). Presenting the risk assessment increased the odds of rejecting loan applications by a factor of 2.09 (Table 3). For all levels of perceived risk up to 46.0% (covering 97.3% of risk estimates with the risk assessment), participants were more than twice as likely to reject loan applications if they were shown the risk assessment (Table A.4). For instance, an applicant with a perceived risk of 30% would have an 8.7% likelihood of being rejected by a participant not shown the risk assessment and an 18.8% likelihood of being rejected by a participant shown the risk assessment.

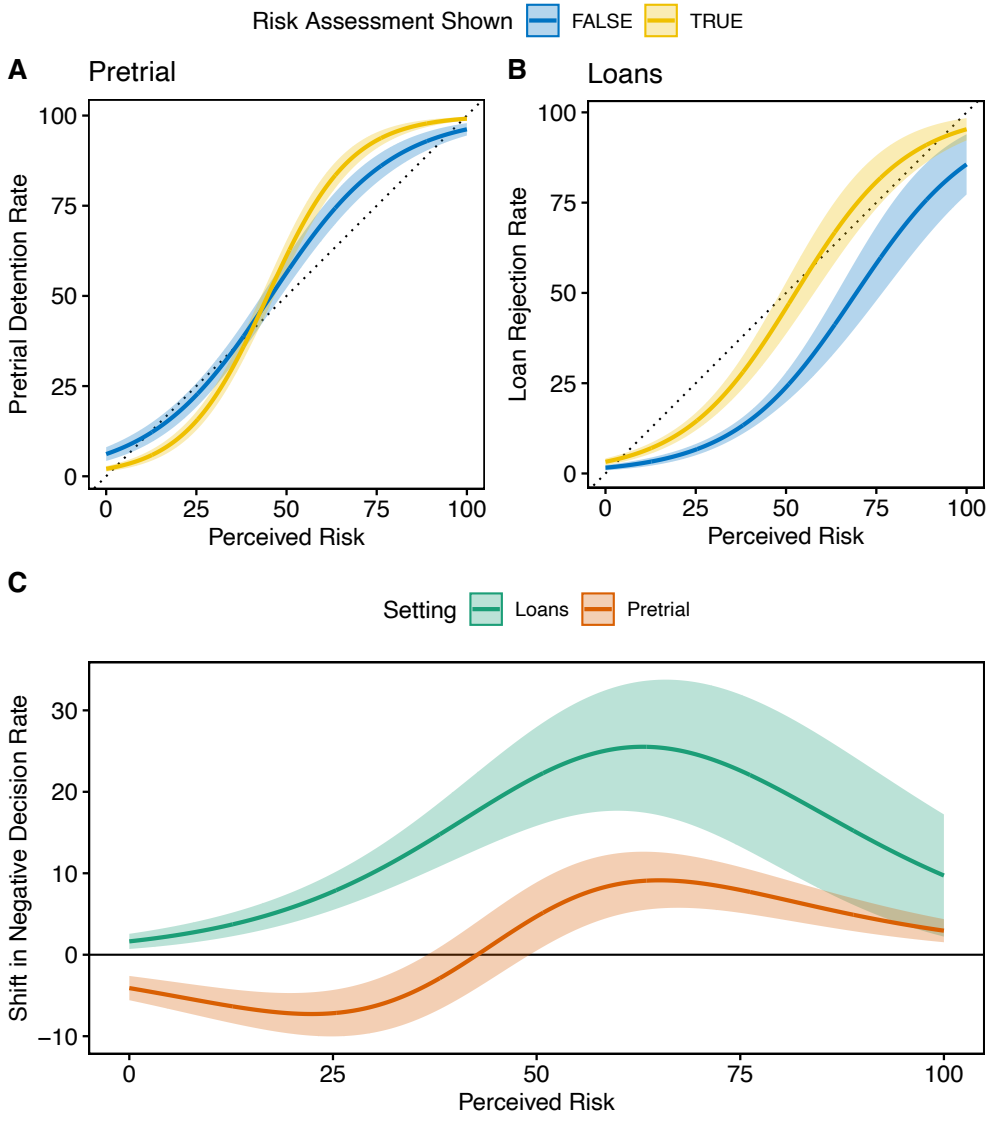

Fig. 5. Changes in the decision-making processes caused by showing the risk assessments to participants. (A) Decision functions indicating the likelihood of detaining a pretrial defendant based on the perceived risk of that defendant (see Table 3 for model coefficients). The risk assessment made participants more sensitive to increases in perceived risk, reducing detention at low risk and increasing detention at high risk. (B) Decision functions indicating the likelihood of rejecting a loan application based on the perceived risk of that applicant (see Table 3 for model coefficients). The risk assessment caused rejection rates to increase at all levels of perceived risk. (C) Shift in negative decision (i.e., pretrial detention or loan rejection) probability due to the shift in the DMP caused by showing the risk assessment. Given a perceived risk of 50%, for instance, the DMP shift increased the likelihood of pretrial detention by 4.7% and the likelihood of loan rejection by 21.9%. Bands indicate 95% confidence intervals in all panels. The values behind this figure are summarized in Table A.4.

Table 3. Bayesian mixed-effects logistic regression results estimating the likelihood of a negative decision about defendants and loan applicants as a function of perceived risk, following Equation 1. The first column presents the coefficient of each factor; the second column presents the coefficient of the interaction between that factor and the risk assessment being shown. The second column thus describes how showing the risk assessment altered each factor. Parenthetical terms represent standard errors and terms in brackets represent odds ratios. The intercept represents modeled participant responses at a perceived risk of 0%, with perceived risk measured in units of 10%. In the pretrial setting, presenting the risk assessment reduced the likelihood of detention for 0% risk but increased participants' sensitivity to increases in risk. In the loans setting, presenting the risk assessment increased the odds of rejecting loan applications by a factor of 2.09. These patterns are plotted in Figure 5. . P<0.1; * P<0.05; ** P<0.01; *** P<0.001

	Not Shown RA	Shown RA (interaction)
Pretrial		
Intercept	−2.74 (0.17) ***	−1.14 (0.14) [0.32] ***
Perceived Risk	0.60 (0.04) [1.82] ***	+0.27 (0.03) [1.31] ***
Loans		
Intercept	−4.15 (0.24) ***	+0.74 (0.22) [2.09] ***
Perceived Risk	0.60 (0.05) [1.82] ***	+0.05 (0.05) [1.05]

When asked to reflect on their behavior after making decisions, participants did not seem to recognize that the risk assessment had altered how they consider risk when making decisions. Despite becoming more attentive to risk when making decisions, participants presented with a risk assessment expressed less support for basing decisions on risk (Pretrial: P=.003, d=0.21; Loans: P=.001, d=0.23). Furthermore, the risk assessments did not alter participant reports regarding the priority that decision-makers should assign to key considerations such as risk (Table A.5).

5.3 Impacts of the Shifts in the Decision-Making Process

We used simulations to estimate the impacts of each risk assessment's influence on the DMP. Our goal was to isolate the effects of the DMP shifts by controlling for the concurrent RPP shifts. We accomplished this through simulations that enabled us to compare the observed Scenario 4 outcomes with the commonly expected Scenario 3 outcomes.

In the pretrial setting, the risk assessment's influence on the DMP reduced the average detention rate but exacerbated racial disparities (Figure 6). Had the risk assessment affected only the RPP (i.e., created a shift from Scenario 1 to Scenario 3), none of the "accuracy," "false positive rate," nor detention rates for either race would have changed. The shift in the DMP (i.e., from Scenario 3 to Scenario 4) increased decision "accuracy" from 57.7% to 60.4% (P<.001, d=4.43), decreased the "false positive rate" from 27.4% to 24.2% (P<.001, d=6.20), and reduced detention by 4.9% for white defendants and by 3.0% for Black defendants (P<.001, d=1.52). Thus, although the DMP shift improved some outcomes, it also increased the racial disparity by 1.9% and by a factor of 1.34 from 5.6% in Scenario 3 to 7.5% in Scenario 4 (P<.001, d=1.06; Figure 6).

In the loans setting, the change in the DMP caused by the risk assessment generated a notable decrease in "accuracy" and increase in rejections (Figure 6). Had the risk assessment affected only the RPP and thus prompted a shift from Scenario 1 to Scenario 3, the decision "accuracy" would have increased from 70.8% to 75.6% (P<.001, d=8.82), the "false positive rate" would have decreased from 17.5% to 11.5% (P<.001, d=12.31), and the rejection rate would have dropped from 22.2% to 14.9% (P<.001, d=13.09). The shift in the DMP negated these potential benefits, however, as the risk

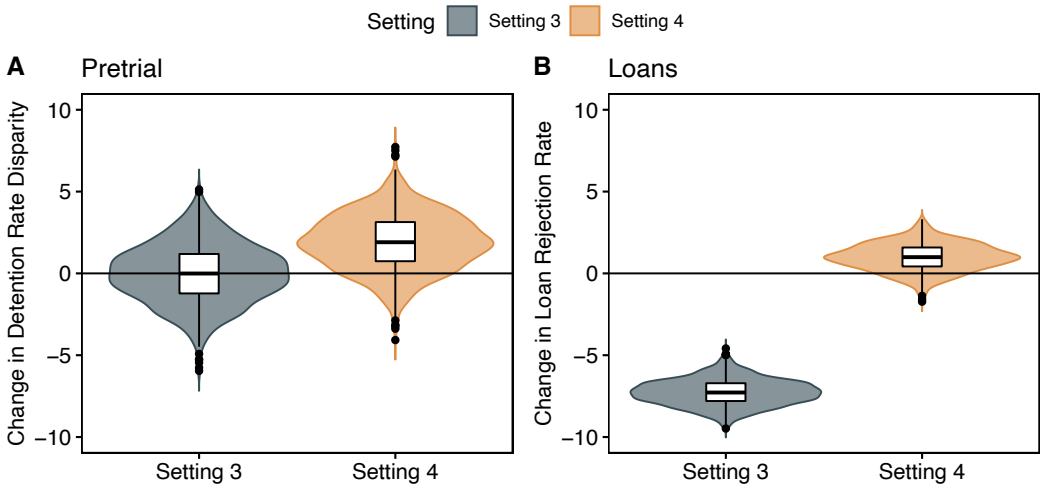

Fig. 6. Simulated outcomes in Scenarios 3 and 4 compared to Scenario 1. (A) Change in the Black-white detention rate disparity in the pretrial setting. Scenario 3 did not significantly alter the average racial disparity while Scenario 4 increased the average racial disparity by 1.9%. The shift in the DMP is therefore responsible for a 1.9% increase in racial disparities. (B) Change in the rejection rate in the loans setting. Scenario 3 reduced the average rejection rate by 7.3% while Scenario 4 increased the average rejection rate by 1.0%. The shift in the DMP is therefore responsible for an 8.3% increase in loan rejections.

assessment made participants more risk-averse. Moving from Scenario 3 to Scenario 4 decreased the decision "accuracy" from 75.6% to 70.7% (P<.001, d=8.83), increased the "false positive rate" from 11.5% to 18.1% (P<.001, d=13.26), and increased the rejection rate from 14.9% to 23.2% (P<.001, d=14.88). Overall, instead of simply improving risk predictions and thereby generating a 7.3% increase in loans granted, the risk assessment also increased risk-aversion and thereby actually *reduced* the loans granted by 1.0% (Figure 6). The shift in the DMP is therefore responsible for an 8.3% increase in loan rejections.

6 DISCUSSION

This paper provides the first direct evidence that risk assessments can systematically alter how people balance risk with other factors when making policy-relevant decisions. Even though our risk assessments improved the accuracy of human predictions, they also induced shifts in decision-making processes that counteracted the potential benefits of these improved predictions. Presenting a risk assessment increased participant sensitivity to risk in pretrial detention decisions (thus exacerbating racial disparities) and increased participant risk-aversion in government loan decisions (thus reducing the loans granted). These shifts mean that even when the risk assessments reduced participant predictions of risk about subjects, participants did not accordingly reduce the rate of negative decisions about those subjects. Alternative explanations, such as the risk assessments simply making participants more confident in their risk estimates, can be ruled out by our data (see Section D in the Appendix).

6.1 Policy Implications

Our results challenge the assumption that improving human predictions with risk assessments will necessarily improve human decision-making—an assumption that has been central to the adoption

of algorithmic decision-making aids by governments. These findings demonstrate the potential limits and harms of efforts to improve public policy by incorporating predictive algorithms into multifaceted policy decisions. If the observed changes were to occur in real-world settings, they would be notable for three primary reasons.

First, our findings indicate that government algorithms could generate unexpected shifts in public policy and jurisprudence. Although improving the accuracy of risk predictions is consistent with policies that include risk as a consideration, a systematic increase in the salience of risk amounts to a shift in the normative balancing act that comprises public policy in domains such as pretrial adjudication [44, 51]. Such a shift reduces the range of factors that decision-makers consider, diminishing the implementation of public policy [76]. In pretrial settings, increasing the weight that judges place on risk would generate undue social harms [74] and enhance the constitutionally contested policy of preventative detention (detaining defendants until trial due to their likelihood to commit future crimes) [27, 44]. In loans settings, greater risk-aversion would reduce government aid and would counteract the goal of promoting equity through giving loans to low-income (and hence high-risk) applicants.

Second, because risk is intertwined with legacies of racial discrimination in the criminal justice and financial systems, more heavily basing decisions on risk would likely exacerbate racial disparities in incarceration and government aid. Due to past and present oppression in the United States, Blacks have disproportionately higher risk levels than whites for being arrested and defaulting on loans, making them particularly vulnerable to increased attention to risk [27, 41]. Indeed, we found that the DMP shifts caused by the risk assessments increased the racial disparity in pretrial decisions and reduced government aid in loans decisions.

Third, because these two effects would arise as an unexpected byproduct of integrating an algorithm into decision-making, they would occur without deliberation or oversight. These shifts in policy and jurisprudence (and the resulting racial disparities) would be the consequence of an algorithm's unintended influence on human decision-making rather than a democratic policymaking process. Because these effects are unexpected, they would likely evade scrutiny, at least until their effects manifest in practice with sufficient evidence. Such changes would likely be further obscured by decision-makers not recognizing that the risk assessment had influenced their behavior, as observed both here and in prior work [28, 29]. These effects add another dimension to the unexpected and unaccountable policy distortions that emerge when laws are translated into code [11].

Together, these implications highlight harms that can arise when algorithms are incorporated into multifaceted policy decisions. If evaluations of algorithmic decision-making aids do not account for human-algorithm interactions and the many normative considerations relevant to policy decisions, they are likely to overestimate the benefits and underestimate the harms of incorporating algorithms into government decision-making [31, 65].

6.2 Future Work

There is an urgent need to uncover potential issues in human-algorithm collaborations *before* algorithms shape life-changing decisions. Risk assessments are increasingly being integrated into high-stakes decisions, yet consistently produce unexpected and unjust impacts in practice [1, 6, 64, 65]. Achieving a more responsible approach to algorithm-in-the-loop decision-making requires several areas of future work.

6.2.1 Open Research Questions. It is necessary to develop a deeper scientific understanding of how risk assessments and other algorithms influence human decision-making. Although we demonstrated that presenting risk assessments can alter human decision-making processes in harmful ways, many open questions remain.

One open question is how the effects of risk assessments vary across contexts. Notably, our risk assessments exerted different effects across the two settings studied, making participants more sensitive to increases in perceived risk in the pretrial setting and more risk-averse in the loans settings. We do not know what caused the observed differences across the two settings. One hypothesis is that the effects of a risk assessment depend on people's pre-existing notions of risk in that context. For instance, people may be strongly predisposed to consider risk in pretrial decisions, such that the risk assessment merely amplified this behavior, but not in government loans decisions, such that the risk assessment prompted heightened concern about mitigating risk.

An important role for future inquiry will be to study how algorithms alter decision-making processes in different settings and with different decision-makers. Algorithms are being deployed in many social contexts beyond government, such as schools [36], hospitals [40], and newsrooms [10]. Although these settings involve some straightforward prediction problems, in many cases people must integrate predictions with other considerations to make decisions. Determining the proper roles for algorithms in these and other settings thus requires a deeper understanding of how algorithmic predictions influence human decision-making across contexts.

A second open question is whether any mechanisms could mitigate the risk assessments' effects on decision-making processes, such that these algorithms do in fact lead to the widely expected Scenario 3 outcomes. It is possible that other approaches to presenting algorithms and structuring decision-making could improve how people incorporate algorithmic advice into their decisions. In the context of algorithm-aided human predictions, for instance, asking people to make preliminary predictions before being shown a risk assessment's predictions modestly improved accuracy and fairness, whereas providing feedback and explanations did not improve performance [29].

6.2.2 Developing a Proactive Pipeline of Evaluations. It is also necessary to develop a testing pipeline that evaluates human interactions with algorithmic decision-making aids before these tools are implemented in practice. Decisions to adopt algorithms should require a baseline of evidence suggesting that they are actually likely to improve decision-making. Our results show that a central assumption motivating risk assessments in public policy—that improving human predictions will improve human decisions—can be violated with laypeople. This finding suggests the need to investigate whether this assumption holds in practice. Furthermore, many regulations across the world point to human oversight as providing protections against algorithms, yet these protections rarely function as desired [30]. Rather than relying on untested assumptions, efforts to integrate algorithms into public policy should be grounded in proactive evaluations of proposed human-algorithm collaborations.

Attaining more thorough knowledge about the effects of algorithmic decision-making aids will require a pipeline of evaluations that combines several modes of analysis: experimental studies with laypeople in lab settings, experimental studies with domain experts in lab settings, and ethnographic and empirical studies of expert interactions with algorithms in practice. Each of these modes has particular strengths and weaknesses. Collectively, they can provide robust, proactive knowledge about how algorithms affect human decision-making and how to improve human-algorithm collaborations.

Developing this pipeline requires further exploring how public servants collaborate with algorithms in practice and how lab experiments can inform the implementation of algorithmic decision-making aids. The primary limitation of this paper is that our findings are based on the behaviors of Mechanical Turk workers in a lab experiment rather than judges or loan agents operating in real-world contexts. Our results do not directly reflect how algorithms affect the behaviors of experts making real decisions. There are likely to be significant differences between how

laypeople and trained experts make decisions with algorithms, particularly related to perceptions of professional identity and autonomy [6].

Despite these differences, experiments with laypeople can shed light on some behaviors of experts in practice. Research suggests that both judges [33, 57, 58] and financial professionals [12, 23, 34] are susceptible to priming and framing effects (alongside other cognitive biases) in much the same manner as laypeople. Prior studies of how laypeople interact with risk assessments [28, 29] have demonstrated racially biased behaviors similar to those observed among judges using risk assessments in practice [1, 14]. Furthermore, the results of this study align with prior experiments suggesting that risk assessments cause law students and judges to place a greater priority on reducing crime risk [61, 62] and that pretrial risk assessments have increased racial disparities in practice [1, 64].

Lab studies with laypeople therefore present a valuable approach for attaining preliminary insights about human-algorithm collaborations. Initial trials with laypeople can provide a foundation of knowledge about how algorithms influence decision-makers and whether there are mechanisms that can improve these collaborations. Compared to studies with experts, experimental studies with laypeople have several advantages. Most importantly, such experiments allow us to learn about human-algorithm collaborations before implementing an algorithm into real-world contexts. Furthermore, compared to lab and in situ evaluations with practitioners, experiments with laypeople can be conducted more quickly, with more participants, and with more precisely controlled experimental procedures. Insights from experiments with laypeople can inform the hypotheses and methods for studies with practitioners, which provide more precise knowledge about human-algorithm collaborations in a particular context but are more intensive to run.[12]

A proactive pipeline of evaluations along these lines should become a central component of proposals and policies for how governments use algorithmic decision-making aids. If algorithms such as risk assessments are to be implemented in a given policy context, there must first be rigorous evidence regarding what impacts they are likely to generate and democratic deliberation supporting those impacts.

ACKNOWLEDGMENTS

We thank the area chairs and reviewers for thoughtful feedback regarding how to improve the manuscript. We also thank Alan Altshuler, Evan Green, Ben Lempert, and Salomé Viljoen for their helpful comments on earlier drafts of this manuscript and Steve Worthington for consultation on statistical methodology. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745303. This work was also supported by the Michigan Society of Fellows.

REFERENCES

[1] Alex Albright. 2019. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. *The John M. Olin Center for Law, Economics, and Business Fellows' Discussion Paper Series* 85 (2019). https://thelittledataset.com/about_files/albright_judge_score.pdf

[2] American Bar Association. 2007. *ABA Standards for Criminal Justice: Pretrial Release* (3rd ed.). https://www.americanbar.org/groups/criminal_justice/publications/criminal_justice_section_archive/crimjust_standards_pretrialrelease_blk/

[3] Appen. 2019. What is Human-in-the-Loop Machine Learning? (2019). https://appen.com/blog/human-in-the-loop/

[4] Arnold Ventures. 2019. Public Safety Assessment FAQs ("PSA 101"). (2019). https://craftmediabucket.s3.amazonaws.com/uploads/Public-Safety-Assessment-101_190319_140124.pdf

[5] James Austin. 2014. Evaluation of Broward County Jail Population: Current Trends and Recommended Options. (2014). https://www.clearinghouse.net/chDocs/public/JC-FL-0008-0018.pdf

[12] Studies with practitioners can also evaluate the validity of experiments with laypeople and determine what kinds of knowledge such experiments can and cannot reliably provide.

[6] Sarah Brayne and Angèle Christin. 2020. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems* (2020). https://doi.org/10.1093/socpro/spaa004

[7] Paul-Christian Bürkner. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* 10, 1 (2018), 395–411. https://doi.org/10.32614/RJ-2018-017

[8] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76, 1 (2017), 1–32. https://doi.org/10.18637/jss.v076.i01

[9] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. 2020. xgboost: Extreme Gradient Boosting. (2020). https://CRAN.R-project.org/package=xgboost

[10] Angèle Christin. 2020. *Metrics at Work: Journalism and the Contested Meaning of Algorithms.* Princeton University Press.

[11] Danielle Keats Citron. 2008. Technological Due Process. *Washington University Law Review* 85, 6 (2008), 1249–1313. https://openscholarship.wustl.edu/law_lawreview/vol85/iss6/2/

[12] Alain Cohn, Jan Engelmann, Ernst Fehr, and Michel André Maréchal. 2015. Evidence for Countercyclical Risk Aversion: An Experiment with Financial Professionals. *American Economic Review* 105, 2 (2015), 860–885. https://doi.org/10.1257/aer.20131314

[13] Alexander Coppock. 2019. Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods* 7, 3 (2019), 613–628. https://doi.org/10.1017/psrm.2018.10

[14] Bo Cowgill. 2018. The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities. (2018). http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf

[15] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–12. https://doi.org/10.1145/3313831.3376638

[16] Matthew DeMichele, Peter Baumgartner, Michael Wenger, Kelle Barrick, and Megan Comfort. 2020. Public safety assessment: Predictive utility and differential prediction by race in Kentucky. *Criminology & Public Policy* 19, 2 (2020), 409–431. https://doi.org/10.1111/1745-9133.12481

[17] Sarah L. Desmarais and Jay P. Singh. 2013. Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States. *Council of State Governments Justice Center* (2013). https://csgjusticecenter.org/wp-content/uploads/2020/02/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf

[18] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. https://doi.org/10.1037/xge0000033

[19] David L. Eckles and Brian F. Schaffner. 2011. Priming Risk: The Accessibility of Uncertainty in Public Policy Decision Making. *Journal of Insurance Issues* 34, 2 (2011), 151–171. http://www.jstor.org/stable/41946320

[20] Hans-Peter Erb, Antoine Bioy, and Denis J. Hilton. 2002. Choice preferences without inferences: subconscious priming of risk attitudes. *Journal of Behavioral Decision Making* 15, 3 (2002), 251–262. https://doi.org/10.1002/bdm.416

[21] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

[22] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29, 2 (2001), 1189–1232. https://doi.org/10.1214/aos/1013203451

[23] Dalia Gilad and Doron Kliger. 2008. Priming the Risk Attitudes of Professionals in Financial Decision Making. *Review of Finance* 12, 3 (2008), 567–586. https://doi.org/10.1093/rof/rfm034

[24] Paul Goodwin and Robert Fildes. 1999. Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy? *Journal of Behavioral Decision Making* 12, 1 (1999), 37–53. https://doi.org/10.1002/(SICI)1099-0771(199903)12:1%3C37::AID-BDM319%3E3.0.CO;2-8

[25] Ben Green. 2018. 'Fair' Risk Assessments: A Precarious Approach for Criminal Justice Reform. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*. https://www.fatml.org/media/documents/fair_risk_assessments_criminal_justice.pdf

[26] Ben Green. 2019. *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future.* MIT Press.

[27] Ben Green. 2020. The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM, 594–606. https://doi.org/10.1145/3351095.3372869

[28] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 90–99. https://doi.org/10.1145/3287560.3287563

[29] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 50:1–50:24. https://doi.org/10.1145/3359152

[30] Ben Green and Amba Kak. 2021. The False Comfort of Human Oversight as an Antidote to A.I. Harm. *Slate* (2021). https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html

[31] Ben Green and Salomé Viljoen. 2020. Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM, 19–31. https://doi.org/10.1145/3351095.3372840

[32] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 178:1–178:25. https://doi.org/10.1145/3359280

[33] Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich. 2001. Inside the Judicial Mind. *Cornell Law Review* 86, 4 (2001), 777–830. https://scholarship.law.cornell.edu/facpub/814

[34] Michael S. Haigh and John A. List. 2005. Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis. *The Journal of Finance* 60, 1 (2005), 523–534. https://doi.org/10.1111/j.1540-6261.2005.00737.x

[35] Kamala Harris and Rand Paul. 2017. Pretrial Integrity and Safety Act of 2017. *115th Congress* (2017). https://www.congress.gov/bill/115th-congress/senate-bill/1593

[36] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2019. Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms. In *Artificial Intelligence in Education*. 157–171. https://doi.org/10.1007/978-3-030-23204-7_14

[37] Indiana Supreme Court. 2010. *Malenchik v. State.* 928 N.E.2d 564.

[38] Tom Jensen and John Tilley. 2012. HB 463 – Statement from the Sponsors. *Criminal Law Reform: The First Year of HB 463* (2012). https://cdn.ymaws.com/www.kybar.org/resource/resmgr/2012_Convention_Files/ac2012_2.pdf

[39] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining Human and Machine Intelligence in Large-scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '12)*. 467–474. http://dl.acm.org/citation.cfm?id=2343576.2343643

[40] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P. Langlotz, Robyn L. Ball, Thomas J. Montine, Brock A. Martin, Gerald J. Berry, Michael G. Ozawa, Florette K. Hazard, Ryanne A. Brown, Simon B. Chen, Mona Wood, Libby S. Allard, Lourdes Ylagan, Andrew Y. Ng, and Jeanne Shen. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *npj Digital Medicine* 3, 1 (2020), 1–6. https://doi.org/10.1038/s41746-020-0232-8

[41] Barbara Kiviat. 2019. The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores. *American Sociological Review* 84, 6 (2019), 1134–1158. https://doi.org/10.1177/0003122419884917

[42] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (2018), 237–293. https://doi.org/10.1093/qje/qjx032

[43] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction Policy Problems. *American Economic Review* 105, 5 (2015), 491–95. https://doi.org/10.1257/aer.p20151023

[44] John Logan Koepke and David G. Robinson. 2018. Danger Ahead: Risk Assessment and the Future of Bail Reform. *Washington Law Review* 93 (2018), 1725–1807. https://digitalcommons.law.uw.edu/wlr/vol93/iss4/4/

[45] Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. 2013. Crowdsourcing Performance Evaluations of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 207–216. https://doi.org/10.1145/2470654.2470684

[46] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 29–38. http://doi.acm.org/10.1145/3287560.3287590

[47] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[48] Joa Sang Lim and Marcus O'Connor. 1995. Judgemental Adjustment of Initial Forecasts: Its Effectiveness and Biases. *Journal of Behavioral Decision Making* 8, 3 (1995), 149–168. https://doi.org/10.1002/bdm.3960080302

[49] Michael Lipsky. 2010. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services* (30th Anniversary Expanded ed.). Russell Sage Foundation.

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

[51] Barry Mahoney, Bruce D. Beaudin, John A. Carver III, Daniel B. Ryan, and Richard B. Hoffman. 2001. Pretrial Services Programs: Responsibilities and Potential. *National Institute of Justice: Issues and Practices in Criminal Justice* (2001). https://www.ncjrs.gov/pdffiles1/nij/181939.pdf

[52] Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. 2019. Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology* 10 (2019), 1–14. https://doi.org/10.3389/fpsyg.2019.02767

[53] Dominique Makowski, Mattan S. Ben-Shachar, and Daniel Lüdecke. 2019. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software* 4, 40 (2019), 1–8. https://doi.org/10.21105/joss.01541

[54] myFICO. 2016. Understanding FICO Scores. (2016). https://www.myfico.com/Downloads/Files/myFICO_UYFS_Booklet.pdf

[55] New Jersey Courts. 2017. One Year Criminal Justice Reform Report to the Governor and the Legislature. (2017). https://www.njcourts.gov/courts/assets/criminal/2017cjrannual.pdf

[56] Eric Potash, Joe Brew, Alexander Loewi, Subhabrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Raed Mansour, and Rayid Ghani. 2015. Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2039–2047. https://doi.org/10.1145/2783258.2788629

[57] Jeffrey J. Rachlinski, Sheri Lynn Johnson, Andrew J. Wistrich, and Chris Guthrie. 2008. Does Unconscious Racial Bias Affect Trial Judges. *Notre Dame Law Review* 84 (2008), 1195–1246. https://scholarship.law.nd.edu/ndlr/vol84/iss3/4/

[58] Jeffrey J. Rachlinski and Andrew J. Wistrich. 2018. Gains, Losses, and Judges: Framing and the Judiciary. *Notre Dame Law Review* 94, 2 (2018), 521–582. https://scholarship.law.nd.edu/ndlr/vol94/iss2/2/

[59] Nicholas Scurich and Richard S. John. 2011. The Effect of Framing Actuarial Risk Probabilities on Involuntary Civil Commitment Decisions. *Law and Human Behavior* 35, 2 (2011), 83–91. https://doi.org/10.1007/s10979-010-9218-4

[60] Sheriff's Justice Institute. 2016. Central Bond Court Report. (2016). https://www.chicagoreader.com/pdf/20161026/Sheriff_s-Justice-Institute-Central-Bond-Court-Study-070616.pdf

[61] Jennifer Skeem, Nicholas Scurich, and John Monahan. 2019. Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants. *Law & Human Behavior* (2019). http://dx.doi.org/10.1037/lhb0000360

[62] Sonja B. Starr. 2014. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review* 66, 4 (2014), 803–872. https://www.stanfordlawreview.org/print/article/evidence-based-sentencing-and-the-scientific-rationalization-of-discrimination/

[63] David Steinhart. 2006. Juvenile detention risk assessment: A practice guide to juvenile detention reform. *The Annie E. Casey Foundation* (2006). https://www.aecf.org/m/resourceimg/aecf-juveniledetentionriskassessment1-2006.pdf

[64] Megan T. Stevenson. 2018. Assessing Risk Assessment in Action. *Minnesota Law Review* 103 (2018), 303–384. https://scholarship.law.umn.edu/mlr/58/

[65] Megan T. Stevenson and Jennifer L. Doleac. 2021. Algorithmic Risk Assessment in the Hands of Humans. (2021). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440

[66] The Leadership Conference Education Fund. 2018. The Use of Pretrial "Risk Assessment" Instruments: A Shared Statement of Civil Rights Concerns. (2018). https://leadershipconferenceedfund.org/pretrial-risk-assessment/

[67] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481 (1981), 453–458. https://doi.org/10.1126/science.7455683

[68] United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. 2014. State Court Processing Statistics, 1990-2009: Felony Defendants in Large Urban Counties. https://doi.org/10.3886/ICPSR02038.v5

[69] USDA Rural Development. 2020. Single Family Housing Repair Loans & Grants. (2020). https://www.rd.usda.gov/programs-services/single-family-housing-repair-loans-grants

[70] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 10–19. https://doi.org/10.1145/3287560.3287566

[71] Jiangtao Wang, Yasha Wang, and Qin Lv. 2019. Crowd-Assisted Machine Learning: Current Issues and Future Directions. *Computer* 52, 1 (2019), 46–53. https://doi.org/10.1109/MC.2018.2890174

[72] Human Rights Watch. 2017. "Not in it for Justice": How California's Pretrial Detention and Bail System Unfairly Punishes Poor People. (2017). https://www.hrw.org/report/2017/04/11/not-it-justice/how-californias-pretrial-detention-and-bail-system-unfairly

[73] Wisconsin Supreme Court. 2016. *State v. Loomis.* 881 Wis. N.W.2d 749.

[74] Crystal S. Yang. 2017. Toward an Optimal Bail System. *New York University Law Review* 92, 5 (2017), 1399–1493. https://www.nyulawreview.org/issues/volume-92-number-5/toward-an-optimal-bail-system/

[75] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* 32, 4 (2019), 403–414. https://doi.org/10.1002/bdm.2118

[76] Bernardo Zacka. 2017. *When the State Meets the Street: Public Service and Moral Agency.* Harvard University Press.

Received January 2021; revised April 2021; accepted July 2021

A DATA AND RISK ASSESSMENTS

A.1 Pretrial Detention

To create our pretrial risk assessment, we used the dataset "State Court Processing Statistics, 1990-2009: Felony Defendants in Large Urban Counties," which was collected by the U.S. Department of Justice [68]. The dataset contains court processing information about 151,461 felony cases filed in May in even years from 1990-2006 and in 2009 in 40 of the 75 most populous counties in the United States. The data contains information about each case that includes the arrest charges, the defendant's demographic characteristics and criminal history, and the outcomes of the case related to pretrial release (whether the defendant was released before trial and, if so, whether they were rearrested before trial or failed to appear in court for trial).

We first cleaned the dataset. We removed incomplete entries and restricted our analysis to defendants who were at least 18 years old and whose race was recorded as either Black or white. In order to have ground-truth data about whether a defendant was rearrested before trial or failed to appear for trial, we also restricted our analysis to defendants who were released before trial.

This yielded a dataset of 47,141 defendants (Table A.1). The defendants were primarily male (76.7%) and Black (55.7%), with an average age of 30.8 years. Among these defendants, 15.0% were rearrested before trial, 20.3% failed to appear for trial, and 29.8% exhibited at least one of these outcomes (which we defined as "violating" the terms of pretrial release).

We used this data to train a risk assessment (i.e., a machine learning classifier) that predicts whether each defendant will violate pretrial release. We trained the model using gradient boosted decision trees [22] with the `xgboost` implementation in `R` [9]. The classifier incorporated five features about each defendant: age, offense type, number of prior arrests, whether that person has any prior failures to appear, and number of prior convictions. Despite knowing the race and gender of defendants, we excluded these attributes from the model to match common practice among risk assessment developers [4].

We performed model selection and evaluated the model using ten-fold cross-validation. We first set aside a random sample of 10% of the data as a held-out validation set. We then took the remaining 90% of the data as the training data. We split this training data into ten folds, using cross-validation to find hyperparameters for the boosted trees model. Cross-validation on the final model yielded an average test AUC of 0.66 (sd=0.009). We then trained the model on the complete training data and applied it to the held-out validation set, yielding an AUC of 0.67. This indicates comparable accuracy to COMPAS [47], the Public Safety Assessment [16], and other risk assessments used in practice [17].

We selected a sample of 300 defendants from the validation set whose profiles would be shown to participants during the Mechanical Turk experiment. To protect defendant privacy, this sample could include only defendants whose seven displayed attributes were shared with at least two other defendants in the complete dataset. This restriction meant that we could not select a uniform random sample of 300 defendants from the validation set. However, we found in practice that sampling from the validation set with weights based on each defendant's risk score yielded a sample population that resembles the complete set of released defendants across most dimensions (Table A.1).

A.2 Home Improvement Loans

We used a dataset of loans from the peer-to-peer lending company Lending Club to create our loans risk assessment. The data contains records about all 2,004,091 loans that Lending Club issued between 2007 and 2018. Each record includes information such as the purpose of the loan; the loan applicant's job, annual income, and approximate credit score; the loan amount and interest rate; and whether the borrower paid off the loan. The data includes the first three digits of each borrower's zip code but does not include further demographic information (such as the age, race, or gender of applicants).

We cleaned the dataset to remove incomplete entries and classified credit scores into one of five categories (Poor, Fair, Good, Very Good, and Exceptional), as defined by FICO [54]. We restricted our analysis to loans issued for home improvements, which represents 6.7% of the total issued loans. Home improvement loans

represent the third most common purpose for loans in the dataset, following debt consolidation and paying off credit cards. We also limited the data to loans that have been either fully paid or defaulted on.[13]

This yielded a dataset of 45,218 home improvement loans (Table A.2). The average loan was for $14,556.38. The average applicant had an income of $95,262.88 and a credit score of 707.5 (categorized by FICO as "Good"). More than 80% of these loans were fully paid off.

We used this data to train a risk assessment that predicts whether each loan will be defaulted on. We trained the classifier using gradient boosted decision trees [22] with the `xgboost` implementation in `R` [9]. Our model considered seven factors about each loan: three factors about each applicant (annual income, credit score category, and whether they own their home) and four factors about each loan (total value, interest rate, monthly installment, and whether its repayment term is 36 or 60 months).

We evaluated the model using ten-fold cross-validation, following the procedure described above for the pretrial risk assessment. Cross-validation on the final model yielded an average test AUC of 0.70 (sd=0.01). Training the classifier on the complete training data (90% of the samples) and applying it to the held-out validation set (the remaining 10% of the data) yielded an AUC of 0.69. This performance is similar to that of other loan default risk assessments [70].

We selected a sample of 300 loan applicants from the validation set whose profiles would be shown to participants during the Mechanical Turk experiment (Table A.2). These applicants were selected through a uniform random sample from the complete validation set.

B COVID-19 RELIABILITY ANALYSIS

As we prepared to run our experiment in May 2020, we wanted to ensure that our results would not be the product of aberrant behavior prompted by the COVID-19 pandemic. Before running the full experiment, therefore, we conducted a retest of a trial experiment that we had conducted in December 2019.

The December 2019 trial closely resembled the experiment described in the main text. We recruited 240 participants from Mechanical Turk to evaluate a sample of 100 defendants. For the May 2020 trial, we recruited 250 participants to evaluate the same set of 100 defendants. We compared the results of these two trials to determine whether COVID-19 altered the population of Mechanical Turk workers or human interactions with risk assessments. We focused on three results central to our study: the demographics of participants, how participants made risk predictions, and how participants made decisions about whether to release or detain defendants. For all three results, we did not observe any notable differences across the two trials, suggesting that COVID-19 did not have notable impacts on our results.

B.1 Participant Demographics

The demographics of our study participants were similar across the two trials. In both cases, participants were predominantly white (80.5% in 12/2019 vs. 73.4% in 05/2020), male (58.6% vs. 58.0%), and college-educated (73.5% vs. 70.2%). A logistic regression predicting which trial each participant was part of (based on all of the demographic attributes reported during the intro survey) yielded no terms that were statistically significant.

B.2 Predictions

We observed a high degree of consistency between the predictions made across the two trials. The correlation between the average prediction made about each of the 100 defendants was $r(198)=+.94$, $P<.001$. A two-sided t-test yielded no statistically significant difference between participants' prediction quality across the two trials (0.751 vs. 0.753, $P=.820$).

We also estimated the function used by participants to predict the risk of each defendant. We used a mixed-effects linear regression model to measure the average risk prediction about each defendant, grouped by whether the risk assessment was shown and whether the prediction was made in the first or second trial (we refer to this variable as "trial number"). The model included fixed effects for whether the risk assessment was shown, whether the predictions were made in the first or second trial, the attributes of defendants, and the interactions between these three sets of factors (up to three-way). We also included random effects for

[13] Although the data represents unpaid loans as being "charged off," which is more extreme than defaulting on a loan, we refer to charged off loans as being defaulted on because the latter is the more commonly used and understood term.

participant and defendant identities to account for repeated samples. Our goal was to evaluate whether trial number influenced how participants made predictions. We observed minimal differences in the prediction function used across the two trials. The trial number and the interaction between trial number and whether the risk assessment was presented were not statistically significant. Only two of the interactions that included trial number were statistically significant: participants were slightly less responsive to prior failures to appear (P=.025) and prior convictions (P=.039) in the second trial.

B.3 Decisions

Finally, we observed a high degree of consistency between the decisions made across the two trials. The correlation between the average detention rate for each of the 100 defendants was r(198)=+.97, P<.001.

We also estimated the function used by participants to decide whether to release or detain each defendant. We used a mixed-effects logistic regression model on all 8,070 decisions made across the two trials. The model included fixed effects for whether the risk assessment was shown, the trial number, and the perceived risk about each defendant, with up to three-way interactions between these factors. We included random effects for participants, defendants, and status in the experiment to account for repeated measurements. None of the coefficients that included trial number were statistically significant, indicating that the decision-making function did not notably differ across the December 2019 or the May 2020 trials.

B.4 Summary

In sum, we found high levels of test-retest reliability. The results found in May 2020 (in the early stages of the COVID-19 pandemic) closely resemble the results found in December 2019. This suggests that the results presented in this paper were not notably influenced by aberrant behaviors that arose in response to COVD-19. More broadly, this also indicates the reliability of our results as being reproducible upon repeated experimentation.

C ANALYSIS

C.1 Predictions

This section provides further detail on the results provided in Section 5.1. We estimated the risk-prediction process of participants using Bayesian linear regression. We used a Bayesian approach for consistency with the next section, where Bayesian regression enabled analysis based on posteriors. For all results throughout the paper, the inferences made from Bayesian and non-Bayesian regressions were almost identical. We implemented models with the `brms` package in `R` [7], which provides a high-level interface to Markov Chain Monte Carlo (MCMC) sampling for Bayesian inference using `Stan` [8].

In both settings, we regressed the average prediction about each subject (both with and without the risk assessment) on the subject attributes presented to participants and a binary variable (*show.RA*) reflecting whether the risk assessment was shown. We included interactions between *show.RA* and each subject attribute. To account for repeated samples of subjects, the model also included random effects for subject identity. Equation A.1 is the regression in the pretrial setting and Equation A.2 is the regression in the loans setting.

$$\begin{aligned} \text{perceived.risk} \sim\ & \text{race} + \text{gender} + \text{age} + \text{offense.type} + \text{number.prior.arrests} \\ & + \text{prior.failure.to.appear} + \text{show.RA} + \text{race} * \text{show.RA} \\ & + \text{gender} * \text{show.RA} + \text{age} * \text{show.RA} + \text{offense.type} * \text{show.RA} \\ & + \text{number.prior.arrests} * \text{show.RA} + \text{number.prior.convictions} * \text{show.RA} \\ & + \text{prior.failure.to.appear} * \text{show.RA} + (1|\text{subject}) \end{aligned} \tag{A.1}$$

$$\begin{aligned} \text{perceived.risk} \sim\ & \text{income} + \text{fico.category} + \text{own.home} + \text{monthly.installment} \\ & + \text{interest.rate} + \text{loan.amount} + \text{loan.term} + \text{show.RA} \\ & + \text{income} * \text{show.RA} + \text{fico.category} * \text{show.RA} + \text{own.home} * \text{show.RA} \\ & + \text{monthly.installment} * \text{show.RA} + \text{interest.rate} * \text{show.RA} \\ & + \text{loan.amount} * \text{show.RA} + \text{loan.term} * \text{show.RA} + (1|\text{subject}) \end{aligned} \tag{A.2}$$

We initialized models with uninformative priors and implemented sampling using four chains with 1,000 iterations, following 1,000 burn-in iterations on each chain. All coefficients in both models returned $\hat{R} = 1.00$, indicating that the chains were well-mixed and converged to a common distribution. We estimated statistical significance from the samples by using the probability of direction measure and obtaining the equivalent frequentist p-value [52, 53]. The results are summarized in Table A.3.

C.2 Decisions

This section provides further detail on the results provided in Section 5.2.2. We estimated the decision-making process using Bayesian mixed-effects logistic regression, implemented in `brms` [7]. In both settings, we regressed each decision on the perceived risk about the subject in question, whether the risk assessment was shown, and the interaction between these two factors (Equation 1).[14] To account for repeated samples, the model also included random effects for the participant identity, the subject identity, and the progress index (1–30) marking the participant's progress in the experiment.

We initialized models with uninformative priors and implemented sampling using four chains with 1,000 iterations, following 1,000 burn-in iterations on each chain. In both settings, all fixed effect coefficients returned $\hat{R} = 1.00$ and all random effect coefficients returned $\hat{R} \leq 1.01$, indicating that the chains were well-mixed and converged to a common distribution. We estimated statistical significance from the samples by using the probability of direction measure and obtaining the equivalent frequentist p-value [52, 53]. The results are summarized in Table 3. The standard deviations for the random effects in the pretrial setting are 1.03 for worker, 0.90 for subject, and 0.07 for experiment progress index. The standard deviations for the random effects in the loans setting are 1.19 for worker, 0.90 for subject, and 0.29 for experiment progress index.

We then studied the characteristics of these fitted decision-making process functions (Figure 5 and Table A.4). We did this using all 4,000 posterior samples of the fixed effect coefficients from the fitted model. First, we used these samples to calculate the fitted negative decision rate at each level of risk from 0% to 100% (in intervals of 0.1%), both with and without the risk assessment. Second, we used these posterior estimates to calculate the shifts in negative decision rates caused by the risk assessment at each level of risk.

C.3 Simulations

This section provides further detail on the results provided in Section 5.3. We used simulations to isolate the effects of the changes in the DMP due to the risk assessments. This required simulating outcomes in the four scenarios described in Table 1 and comparing the results of Scenario 3 and Scenario 4. We did this in two stages. First, we used data from the experiment to learn participant prediction and decision functions both with and without the presence of a risk assessment. Second, we applied those functions to a large sample of defendants and loan applicants to simulate the outcomes of the four Table 1 scenarios.

C.3.1 Fitting Prediction and Decision Models. We began by learning the functions that explain the average risk prediction and negative decision rate for each subject. For predictions, we used Equations A.1 and A.2. For decisions, our formulas included the same factors as the predictions models, plus the perceived risk about that subject and the interaction between the perceived risk and whether the risk assessment was shown. Equation A.3 is the regression in the pretrial setting and Equation A.4 is the regression in the loans setting.

$$\begin{aligned} \text{detention.rate} \sim\ & \text{perceived.risk} + \text{race} + \text{gender} + \text{age} + \text{offense.type} \\ & + \text{number.prior.arrests} + \text{number.prior.convictions} \\ & + \text{prior.failure.to.appear} + \text{show.RA} + \text{perceived.risk} * \text{show.RA} \\ & + \text{race} * \text{show.RA} + \text{gender} * \text{show.RA} + \text{age} * \text{show.RA} \\ & + \text{offense.type} * \text{show.RA} + \text{number.prior.arrests} * \text{show.RA} \\ & + \text{number.prior.convictions} * \text{show.RA} + \text{prior.failure.to.appear} * \text{show.RA} \end{aligned} \tag{A.3}$$

[14] The *perceived.risk* measurements are based on an average of 18.13±4.00 participant predictions about each subject in each treatment (RA or no RA), with an average standard deviation in risk predictions of 21.85±6.64 and an average standard error of 5.21±1.70. These values are almost identical across the two settings.

$$\begin{aligned} \text{rejection.rate} \sim\ & \text{perceived.risk} + \text{income} + \text{fico.category} + \text{own.home} \\ & + \text{monthly.installment} + \text{interest.rate} + \text{loan.amount} + \text{loan.term} + \text{show.RA} \\ & + \text{perceived.risk} * \text{show.RA} + \text{income} * \text{show.RA} + \text{fico.category} * \text{show.RA} \\ & + \text{own.home} * \text{show.RA} + \text{monthly.installment} * \text{show.RA} \\ & + \text{interest.rate} * \text{show.RA} + \text{loan.amount} * \text{show.RA} + \text{loan.term} * \text{show.RA} \end{aligned} \tag{A.4}$$

We fit all models using generalized linear regression with a logit link function from the quasibinomial family. We used this quasibinomial approach because the fitted values of these regressions are bounded probabilities (either risk predictions or negative decision rates, which both range from 0%-100%). Although linear regression yields very similar results, it does not guarantee that predicted values will be bounded between 0 and 1.

Before applying these models to new defendants, we used leave-one-out cross-validation to test the effectiveness of this approach on the data from our experiment. For each model in each setting, we removed one subject at a time, trained the model on the predictions or decisions about the other 299 subjects, and estimated the prediction or decision that would be made about the held-out subject both with and without the risk assessment. We evaluated these models by applying the risk prediction model and then using the output of that model as input to the decision model. The mean average error (MAE) of the entire pipeline for negative decisions rates is 5.92 (RMSE=7.46) in the pretrial setting and 7.33 (RMSE=9.95) in the loans setting. All the models are unbiased estimators, with mean errors close to 0.

We then fit prediction and decision models for both settings on the complete set of 300 subjects for use in our simulations.

C.3.2 Simulating Predictions and Decisions on New Subjects. We applied these models to the held-out validation sets from both settings (not including the 300 subjects sampled from those datasets for inclusion in our experiment). These samples represent approximately 10% of the complete data in each setting. They contain 4,375 defendants and 4,231 loan applicants drawn from the populations described in Tables A.1 and A.2.

Our simulations proceeded as follows:

(1) Apply the predictions and decisions models to every subject to estimate the negative decision probabilities in the four scenarios from Table 1. The predictions and decisions models enable us to simulate outcomes both with and without a risk assessment's advice. Using the outputs of the predictions model as the perceived risk, we applied these models in all four possible combinations of whether the risk assessment affected predictions and decisions. This process yields four estimated negative decision probabilities for each subject: predictions and decisions are both unaffected by the risk assessment (Scenario 1), predictions are unaffected by the risk assessment but decisions are affected by the risk assessment (Scenario 2), predictions are affected by the risk assessment but decisions are unaffected by the risk assessment (Scenario 3), and predictions and decisions are both affected by the risk assessment (Scenario 4).
(2) Run 1,000 trials simulating the outcome for each subject in each scenario, based on the negative decision probabilities found in the prior step. Doing this allowed us to estimate the distribution of outcomes for all four scenarios from Table 1.

D ALTERNATIVE EXPLANATIONS

In this section, we discuss potential alternative explanations for our conclusion that showing the risk assessment altered the DMP and describe why they are inconsistent with our results.

D.1 Participants Have Greater Confidence in Risk Predictions

One alternative explanation is that the risk assessment makes people more confident in their risk prediction rather than more concerned about avoiding risk in decision-making. In other words, people may place a greater weight on their risk prediction because they are more certain about this prediction rather than because they are more concerned about risk as a consideration. If this were the case, we would expect to see risk become a more "extreme" distinguishing factor in decisions: low levels of perceived risk lead to lower negative decision rates, while high levels of perceived risk lead to higher rates. Although that is what we observe in

the pretrial setting, we observe a very different pattern in the loans setting: rejection rates go up at all levels of risk (Figure 5). The loans setting results are consistent with our explanation that the risk assessment makes people more risk-averse, yet inconsistent with people becoming more confident in their risk predictions. For instance, it is relatively implausible that becoming more confident that a loan applicant has a 0% likelihood to default on the loan would more than double the likelihood of rejecting that loan application (Table A.4).

This pattern in the loans setting suggests that the pretrial setting results are also caused by greater attentiveness to risk rather than greater confidence in estimates of risk. Furthermore, even if the pretrial setting does involve greater confidence in risk predictions, the effect would be equivalent to increasing the salience of risk: in both cases, the risk assessment would be causing perceived risk to become a stronger determinant of whether defendants are released or detained.

We can further investigate the role of confidence in decision-making by looking at participant self-reports of confidence. In the exit survey at the end of the experiment, we asked participants how confident they were in their decisions on a Likert scale from 1 (least confident) to 7 (most confident). We found that the risk assessment had no significant effects on participant confidence. In the pretrial setting, the risk assessment did not alter confidence among participants making predictions (P=.978, d=0.00) or decisions (P=.246, d=0.08). Similarly, in the loans setting, the risk assessment did not alter confidence among participants making predictions (P=.580, d=0.07) or decisions (P=.213, d=0.09). Given that the risk assessments did not significantly impact participant self-reports of confidence, it is unlikely that the effects of the risk assessments can be attributed to them making participants more confident in their estimates of risk.

D.2 Prediction-Makers and Decision-Makers Have Different Predictions of Risk

Another alternative explanation is that perceived risk differs between participants making predictions and participants making decisions. In particular, the risk assessment might exert a stronger influence on participants making predictions than on participants making decisions. Our results directly contradict this explanation, however. Most notable is the contrast between the effects of the risk assessment in the loans setting, reducing predictions of risk without reducing loan rejections. For instance, among the 92.3% of loan applicants for whom the risk assessment reduced perceived risk, almost half received a higher likelihood of rejection when the risk assessment was shown. For this explanation to apply here, it would have to be the case that for almost half of the loan applicants, the risk assessment reduced risk estimates for prediction-makers yet increased risk estimates for decision-makers. Although it is plausible that the risk assessment's effects on predictions could be attenuated for decision-makers, it is not plausible that prediction-makers and decision-makers would have their risk estimates influenced in opposite directions.

D.3 The Risk Assessment Provides a Random Shock to Decisions

A third alternative explanation is that the risk assessments provide a random shock to decision-making, adding "noise" to decisions in a manner that is not connected to perceived risk. Two results clearly rule out this explanation. First, we observed that the reduction in pretrial detention was statistically significant, indicating that risk assessments can influence decisions in specific directions. Second, in both settings there was a positive and statistically significant relationship between changes in perceived risk and changes in negative decision rates for each subject (Figure 4). These correlations indicate that the risk assessments' effect on decisions is (at least loosely) connected to the risk assessments' effect on perceived risk.

D.4 The Risk Assessment Alters the "Other Factors" Rather than the DMP

Another potential explanation is that the risk assessment alters the calculation of the "other factors" that are incorporated into the DMP (Figure 1) rather than (or in addition to) altering the DMP itself. In the loans setting, for instance, the risk assessment could cause people to reduce their evaluation of the benefits of granting home improvement loans rather than cause people to become more risk-averse. However, there is little reason to believe that receiving an algorithmic risk estimate would prompt a large enough reduction in perceived benefit to fully offset the large observed reductions in perceived risk. Moreover, although this alternative explanation would place the change at a different place in Figure 1, the overall effect would be similar: the risk assessment would be altering decision-making in unexpected ways that can have significant negative impacts.

E TABLES

Table A.1. Attributes of the full sample of defendants released before trial and the 300-defendant sample presented to participants in the experiment, by race. A violation means that the defendant was rearrested before trial, failed to appear for trial, or both.

	All N=47,141	**Black** N=26,246	**White** N=20,895	**Sample** N=300	**Black** N=189	**White** N=111
Attributes						
Male	76.7%	77.3%	75.5%	86.7%	88.4%	83.8%
Black	55.7%	100.0%	0.0%	63.0%	100.0%	0.0%
Mean age at arrest	30.8	30.1	31.8	28.1	27.1	29.8
Drug crime	36.9%	39.2%	34.0%	49.3%	50.8%	46.8%
Property crime	32.7%	30.7%	35.3%	30.3%	28.0%	34.2%
Violent crime	20.4%	20.9%	19.8%	14.0%	14.3%	13.5%
Public order crime	10.0%	9.3%	10.8%	6.3%	6.9%	5.4%
Has prior arrest(s)	63.4%	68.4%	57.0%	64.7%	73.5%	49.5%
Mean number of prior arrests	3.8	4.3	3.1	4.3	5.0	3.1
Has prior conviction(s)	46.5%	51.2%	40.7%	50.0%	57.7%	36.9%
Mean number of prior convictions	1.9	2.2	1.6	2.4	2.9	1.7
Has prior failure(s) to appear	25.1%	28.8	20.4%	31.7%	34.4%	27.0%
Outcome						
Rearrest	15.0%	16.9%	12.6%	19.0%	20.1%	17.1%
Failure to appear	20.3%	22.6%	17.5%	25.3%	29.6%	18.0%
Violation	29.8%	33.1%	25.6%	36.0%	39.2%	30.6%

Table A.2. Attributes of the full sample of approved home improvement loans and the 300-loan sample presented to participants in the experiment.

	All N=45,218	**Sample** N=300
Applicant		
Mean annual income	$95,262.88	$93,349.22
Mean credit score	707.5	705.9
Has a mortgage	83.9%	83.0%
Loan		
Mean loan amount	$14,556.38	$14,076.00
Mean months to pay off loan	42.4	42.6
Mean monthly payment	$435.75	$419.49
Mean interest rate	13.0%	13.2%
Outcome		
Loan paid off	83.2%	84.7%
Loan defaulted on	16.8%	15.3%

Table A.3. Bayesian linear regression results estimating the risk-prediction process in both settings, following Equations A.1 and A.2. The first column presents the coefficient of each factor; the second column presents the coefficient of the interaction between that factor and the risk assessment being shown. The second column thus describes how showing the risk assessment altered each factor. In the loans regression, annual income, loan amount, and monthly installment are measured in units of $1,000. Parenthetical terms represent standard errors. . P<0.1; * P<0.05; ** P<0.01; *** P<0.001

	Not Shown RA	Shown RA (interaction)
Pretrial		
Intercept	27.88 (1.50) ***	+6.98 (2.03) ***
White	−0.03 (0.72)	−0.98 (0.98)
Male	0.04 (0.91)	−0.42 (1.25)
Age	0.03 (0.04)	−0.20 (0.05) ***
Property crime	−2.29 (0.74) ***	+0.43 (1.04)
Public order crime	−0.28 (1.59)	−3.50 (2.21)
Violent crime	3.00 (0.95) ***	−7.45 (1.27) ***
Number of prior arrests	0.72 (0.17) ***	+0.20 (0.23)
Number of prior convictions	0.31 (0.17) .	+0.09 (0.22)
Prior failure to appear	27.82 (1.33) ***	−7.43 (1.76) ***
Loans		
Intercept	39.37 (1.93) ***	−24.02 (2.45) ***
Annual income	−0.03 (0.01) ***	−0.02 (0.01) *
Good FICO score	−5.81 (1.04) ***	+2.23 (1.31) .
Very good FICO score	−7.91 (1.46) ***	+1.47 (1.83)
Exceptional FICO score	−9.29 (2.52) ***	−0.51 (3.24)
Fully own home	−0.30 (0.99)	+2.13 (1.26) .
Loan amount	0.27 (0.28)	−0.45 (0.37)
Monthly installment	−0.74 (8.96)	+16.81 (11.50)
Interest rate	0.33 (0.12) **	+0.51 (0.15) ***
60-month term	−2.21 (1.91)	+7.41 (2.49) **

Table A.4. Modeled probability of negative decisions at a range of perceived risk levels, by setting and risk assessment treatment. The negative decision in the pretrial setting is detaining the defendant; the negative decision in the loans setting is rejecting the loan application. No RA indicates the probability of negative decisions when the risk assessment is not shown, Shown RA indicates the probability of negative decisions when the risk assessment is shown, and Difference indicates the difference between these values (numbers in brackets indicate the effect size of this difference). All differences in both settings are statistically significant with P<.001. These results are plotted in Figure 5.

	Pretrial			**Loans**		
Perceived Risk	No RA	Shown RA	Difference	No RA	Shown RA	Difference
0%	6.15%	2.06%	−4.09% [5.38]	1.60%	3.24%	+1.64% [3.45]
10%	10.62%	4.74%	−5.88% [5.93]	2.84%	5.98%	+3.15% [4.65]
20%	17.73%	10.52%	−7.21% [5.64]	5.00%	10.82%	+5.82% [6.10]
30%	28.13%	21.80%	−6.33% [3.84]	8.70%	18.81%	+10.11% [7.17]
40%	41.58%	39.83%	−1.75% [0.88]	14.73%	30.68%	+15.95% [7.34]
50%	56.41%	61.11%	+4.70% [2.20]	23.89%	45.78%	+21.89% [7.07]
60%	70.16%	78.83%	+8.67% [4.49]	36.31%	61.63%	+25.32% [6.49]
70%	81.01%	89.80%	+8.79% [5.52]	50.80%	75.27%	+24.47% [5.39]
80%	88.54%	95.41%	+6.87% [5.35]	65.04%	85.19%	+20.14% [4.14]
90%	93.32%	98.00%	+4.68% [4.71]	76.93%	91.55%	+14.63% [3.19]
100%	96.19%	99.14%	+2.95% [4.07]	85.60%	95.32%	+9.72% [2.54]

Table A.5. Participant beliefs about how decision-makers should balance priorities. After making decisions, participants were asked to what extent a decision-maker (i.e., a judge or government loan agent) should value four salient considerations when making decisions. Participants had to assign a total of 100 points (in increments of 5) across the four considerations. None of the average values assigned to these considerations differ significantly across the control (Not Shown RA) and treatment (Shown RA) groups.

	Not Shown RA	Shown RA	P-value	Effect size
Pretrial				
Incapacitation	30.86	29.89	.341	0.07
Freedom	25.76	26.68	.372	0.07
Deterrence	20.04	19.05	.245	0.08
Rehabilitation	23.35	24.38	.289	0.08
Loans				
Likelihood to pay	40.98	39.28	.211	0.09
Equity	21.51	22.59	.124	0.11
Economic development	19.63	19.29	.622	0.03
Neighborhood stability	17.89	18.84	.200	0.09