NBER WORKING PAPER SERIES

ALGORITHMIC RECOMMENDATIONS AND HUMAN DISCRETION

Victoria Angelova
Will S. Dobbie
Crystal Yang

Working Paper 31747
http://www.nber.org/papers/w31747

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023

We thank Alex Albright, David Arnold, Ian Ayres, Peter Bergman, David Chan, Raj Chetty, Dhruv Gaur, Benjamin Goldman, Nathan Hendren, Mitch Hoffman, Peter Hull, Louis Kaplow, Lawrence Katz, Bentley MacLeod, Jack Mountjoy, Sendhil Mullainathan, Isaac Opper, Emma Rackstraw, Manish Raghavan, Ashesh Rambachan, Steven Shavell, Andrei Shleifer, Sonja Starr, and numerous seminar participants for helpful comments and discussions. We are indebted to Kenneth Gu, Sara Kao, Qing Liu, Stephanie Lukins, Dan Ma, Antonn Park, Miguel Purroy, Nada Shalash, and Michelle Wu for their outstanding contributions to this work. This research was funded by the Russell Sage Foundation and Harvard University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2023 by Victoria Angelova, Will S. Dobbie, and Crystal Yang. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Algorithmic Recommendations and Human Discretion
Victoria Angelova, Will S. Dobbie, and Crystal Yang
NBER Working Paper No. 31747
September 2023
JEL No. C01,D8,K40

ABSTRACT

Human decision-makers frequently override the recommendations generated by predictive algorithms, but it is unclear whether these discretionary overrides add valuable private information or reintroduce human biases and mistakes. We develop new quasi-experimental tools to measure the impact of human discretion over an algorithm on the accuracy of decisions, even when the outcome of interest is only selectively observed, in the context of bail decisions. We find that 90% of the judges in our setting underperform the algorithm when they make a discretionary override, with most making override decisions that are no better than random. Yet the remaining 10% of judges outperform the algorithm in terms of both accuracy and fairness when they make a discretionary override. We provide suggestive evidence on the behavior underlying these differences in judge performance, showing that the high-performing judges are more likely to use relevant private information and are less likely to overreact to highly salient events compared to the low-performing judges.

Victoria Angelova
Harvard University
vangelova@g.harvard.edu

Will S. Dobbie
Harvard Kennedy School
79 John F. Kennedy St.
Cambridge, MA 02138
and NBER
willdobbie@gmail.com

Crystal Yang
Harvard Law School
Griswold 301
Cambridge, MA 02138
and NBER
cyang@law.harvard.edu

I Introduction

Human decisions are often mistaken, noisy, and biased (e.g., Tversky and Kahneman 1974; Mullainathan 2002; Bordalo, Gennaioli, and Shleifer 2012; Kahneman, Sibony, and Sunstein 2021). These seemingly intractable problems have contributed to the rapid adoption of predictive algorithms in a range of high-stakes settings, from job screening to medical diagnoses to pretrial release decisions. Yet these same settings still require that a human oversees the algorithm and makes the final decision. The hope is that by retaining human oversight, the human decision-maker can add valuable private information and correct inaccurate algorithmic predictions. However, allowing for human discretion can also reintroduce the same human biases and mistakes that motivated the introduction of the algorithms in the first place. Distinguishing between these possibilities and measuring the impact of human discretion on the accuracy of decisions remain difficult, complicating efforts to develop optimal human oversight policies.

This paper develops new quasi-experimental tools to measure the impact of human discretion over an algorithm on the accuracy of decisions. We develop these tools in the context of bail decisions, where judges are directed to release most defendants before trial while minimizing the risk of pretrial misconduct (such as failing to appear in court or being arrested for a new crime). By law, bail judges can only consider case and defendant characteristics if that information is relevant to a defendant's risk of pretrial misconduct. Judges are explicitly told that pretrial release decisions cannot be based on other objectives, such as an individual's decision to not admit culpability. To help guide these decisions, judges are often given an algorithmic risk assessment that predicts the likelihood of misconduct and recommends whether to release or detain the defendant. They frequently override these algorithmic recommendations, however, despite influential work showing that such algorithms can substantially outperform a bail judge working alone (e.g., Kleinberg et al. 2018). The key open question is whether allowing for such discretion can yield more accurate decisions such that a human and an algorithm working together can be better than an algorithm working alone.

We measure the impact of human discretion over an algorithm on the accuracy of decisions by comparing each judge's observed pretrial misconduct rate to the counterfactual misconduct rate under the algorithm at the same release rate, thereby avoiding the challenges associated with jointly identifying variation in performance and preferences (e.g., Chan, Gentzkow, and Yu 2022). The intuition for our approach is simple: a judge who only overrides the algorithmic recommendation because she prefers a different release rate will have the same misconduct rate as the algorithm, after holding the release rate fixed. But a judge who overrides the algorithm because she disagrees with the algorithm's misconduct predictions will have a lower (higher) misconduct rate if she is more (less) skilled at predicting misconduct than the algorithm, again after holding the release rate fixed. We can therefore say that human discretion leads to more (less) accurate decisions on average if the judges can, on average, achieve a lower (higher) pretrial misconduct rate than the algorithm alone, at the judges' existing release rates. Our approach of comparing judges to a counterfactual at the same release rate builds on approaches used by Lakkaraju et al. (2017) and Kleinberg et al. (2018).

Estimating the impact of human discretion over an algorithm on the accuracy of decisions in this way is complicated by an important selection challenge. We only observe pretrial misconduct among the selected subset of defendants that the judges choose to release before trial. We therefore cannot directly measure the counterfactual misconduct rate under the algorithm at the judges' existing release rates, as we are missing

the required misconduct outcomes among the defendants whom the algorithm would have released but the judges chose to detain. This selection challenge can be understood as a kind of missing data problem that economists and statisticians have developed methods to overcome in a variety of contexts.

In the first part of the paper, we overcome this selection challenge and measure the impact of human discretion on the accuracy of decisions by using quasi-experimental tools that leverage the quasi-random assignment of decision-makers (such as bail judges) to individuals (such as defendants). Our approach can be illustrated in three steps. First, we estimate the average misconduct potential of defendants with risk scores at or below the relevant risk score cutoff to solve the selection problem at a given release rate. We estimate the required average misconduct parameter by extrapolating observed misconduct rates across quasi-randomly assigned judges among the relevant subset of defendants. Second, we compare each judge to the algorithmic counterfactual at their existing release rate by repeating these extrapolations for a wide range of risk score cutoffs that span the judges' existing release rates. Third, we calculate the share of judges with conditional misconduct rates that are higher and lower than the algorithmic counterfactual after accounting for sampling error in our judge-level estimates.[1]

The second part of the paper uses our quasi-experimental approach to measure the impact of human discretion on the accuracy of decisions in a large, mid-Atlantic city that was one of the first places in the country to introduce a pretrial risk assessment. The judges in our setting respond to the algorithmic recommendations, with release rates sharply falling when the algorithmic recommendation discontinuously changes from release to detain. But they also frequently override these recommendations for both observably low- and high-risk defendants, indicating substantial disagreement with the algorithm's risk score predictions. We also observe considerable variation in the judges' misconduct rates at the same release rate, indicating important differences in their skill. The combination of a long-standing risk assessment algorithm, frequent overrides for both observably low- and high-risk defendants, and variation in judge skill makes this an ideal setting to study the impact of human discretion over an algorithm on the accuracy of decisions.

We find that the judges in our setting underperform the algorithm when they make discretionary overrides, increasing pretrial misconduct by an average of 2.4 percentage points at the judges' existing release rates (a 15% increase from the mean). This finding indicates that the typical judge in our setting is less skilled at predicting misconduct than the algorithm and that we could substantially decrease misconduct rates by automating release decisions. But this average impact masks substantial variation in the judges' performance compared to the algorithm. The negative average impact of human discretion on the accuracy of decisions is explained by the 90% of judges who underperform the algorithm when they make a discretionary override. In fact, nearly 70% of the judges make override decisions that are no better than random—that is, they could achieve a lower pretrial misconduct rate by flipping a coin or using a random number generator. At the same time, we also find that 10% of the judges outperform the algorithm when

[1] Our quasi-experimental approach to estimating the algorithmic counterfactual builds on a recent literature estimating average treatment effects with multiple discrete instruments (Brinch, Mogstad, and Wiswall, 2017; Mogstad, Santos, and Torgovitsky, 2018; Hull, 2020; Arnold, Dobbie, and Hull, 2022). Our approach is most closely related to Hull (2020) and Arnold, Dobbie, and Hull (2022), who consider different extrapolations of quasi-experimental moments in the spirit of "identification at infinity" in sample selection models (Chamberlain, 1986; Heckman, 1990; Andrews and Schafgans, 1998). An important advantage of our approach is that it does not require a conventional first-stage monotonicity assumption (Imbens and Angrist, 1994; Heckman and Vytlacil, 2005), which, in our setting, effectively requires that the judges are equally skilled at predicting misconduct outcomes.

they make a discretionary override, suggesting that a human and algorithm working together can potentially outperform automated release decisions. These high-skill judges are evenly distributed across the range of release rates and have similar demographics, political affiliations, and years of experience as the low-skill judges, despite the large differences in performance. The one notable difference is that the high-skill judges are much less likely to have previously worked in law enforcement compared to the low-skill judges.

Our approach to measuring the impact of human discretion on the accuracy of pretrial decisions only requires that the average misconduct parameters can be accurately extrapolated from the data and that the judges' objectives are well specified. We consider several extensions that test or relax these assumptions. We start by considering a range of alternative specifications when estimating the average misconduct risk parameters, finding similar results when we use linear extrapolations, local linear extrapolations, and a modified estimation approach where we only extrapolate to the most lenient judge at a given risk score cutoff and then calculate bounds for the remaining defendants. We then consider a range of pretrial misconduct definitions, finding similar results when we use only failures to appear, only new criminal activity, only new violent criminal activity, or a modified measure that captures the social cost of different misconduct types. We finally consider the importance of extra-legal objectives, finding similar results by defendant subgroups such as race, age, education, and charge severity. All of these findings support our interpretation of the results and highlight the considerable variation in judge performance in our setting.

The final part of the paper provides more suggestive evidence of the behavior underlying these differences in judge performance. We start by showing that the high- and low-skill judges use the observable information that is available to both the judges and the algorithm in a remarkably similar way. The two sets of judges override the algorithm at similar rates, at similar parts of the risk score distribution, and for observably similar defendants. These findings suggest that the differences in judge performance are not driven by the use of such observable information but rather by the use of private information that is not available to the algorithm. There is also more direct support for the idea that the high- and low-skill judges differ in how they use private information. The high-skill judges meaningfully outperform a new algorithm that predicts their release decisions using the observable information that is available to both them and the original algorithm, suggesting they are using relevant private information to improve the accuracy of their decisions. In contrast, the low-skill judges meaningfully underperform the new algorithm, suggesting they are instead adding noise and inconsistency to their decisions when they attempt to use such private information.

We then document three specific ways that private information leads to noise and inconsistency for low-skill judges. The first is that the low-skill judges consistently overweight factors that are not particularly predictive of misconduct risk, such as having an out-of-state address, and underweight factors that are particularly predictive of misconduct risk, such as having an aggravating risk factor, leading to a deterioration in the low-skill judges' performance relative to both the high-skill judges and the algorithm. The second is that the low-skill judges are more likely to assign monetary bail than the high-skill judges and perform worse when we restrict to the subset of defendants who are released on monetary bail, suggesting that the low-skill judges may mistakenly release some high-risk defendants and mistakenly detain some low-risk ones when setting monetary bail. In contrast, the high-skill judges are much more likely to impose non-financial conditions such as treatment for substance abuse disorders or counseling for mental

health issues that are meant to more directly address the defendants' underlying needs. The third is that the low-skill judges are more likely to overreact to highly salient but largely uninformative events like hearing a case where a different defendant is arrested for a serious violent offense while on pretrial release. We find that the low-skill judges are much more likely to detain observably low-risk defendants after such an uninformative event, with no detectable changes in conditional misconduct rates. We also see that the effects are concentrated among defendants who are particularly representative of those arrested for serious violent crimes (Kahneman and Tversky, 1972; Bordalo et al., 2016) and who have observable characteristics that are particularly overweighted by judges (Bordalo, Gennaioli, and Shleifer, 2015; Sunstein, 2022). These findings suggest that the low-skill judges are worse at filtering out noise than the high-skill judges.[2]

We conclude by providing suggestive evidence on the factors the high- and low-skill judges consider when making release decisions, yielding new insights into the types of private information that add valuable signal versus noise. We asked the judges in our sample to rank the importance of different factors when making the decision to impose monetary bail in an original survey. The high- and low-skill judges report using the observable information that is available to both the judges and the algorithm in a remarkably similar way, consistent with the patterns seen in the administrative court data. But there are striking differences in the importance attached to the private information that is not available to the algorithm. The low-skill judges place far more importance on demographic factors such as race, while the high-skill judges place far greater importance on non-demographic factors such as mental health, substance abuse, and financial resources. These findings help confirm the patterns seen in the administrative data, while suggesting that we may be able to improve the accuracy of release decisions by teaching the judges to focus only on the most relevant factors not included in the algorithm. For example, our results suggest that teaching the judges to consider a defendant's financial resources when setting monetary bail may help improve their performance relative to the algorithm.

Our results are an important proof of concept that the most skilled human decision-makers can still add value to the decision-making process by distinguishing between valuable private information and noise. One insight from our work is that, to the extent that humans still make the final decision in a setting, there will not necessarily be a single correct human oversight policy since the impact of such policies depends on the predictive abilities of the human decision-makers. The most skilled human decision-makers can potentially improve the accuracy and fairness of decisions compared to an algorithm working alone, even though the majority of human decision-makers may be better off strictly following the algorithmic recommendations (e.g., Berk 2017; Kleinberg et al. 2018; Jung et al. 2020; Mullainathan and Obermeyer 2022). These findings are consistent with recent work showing that strict guidelines can reduce welfare when there is variation in human ability and that more nuanced policies are needed to improve decision-making in such settings (e.g., Currie and MacLeod 2020; Rambachan 2022; Chan, Gentzkow, and Yu 2022). An important question for future work is how to improve the predictive abilities of human decision-makers by learning from the most skilled and, when that is not possible, how to constrain the least skilled decision-makers.

[2]These findings are consistent with a body of work studying high-performing forecasters in a large, government-funded tournament (e.g., Tetlock and Gardner 2015). Mellers et al. (2015) find that high-performing forecasters have a greater ability to accurately distinguish signals from noise compared to typical forecasters. More recent work by Satopää et al. (2021) shows that interventions to improve forecasts in this setting work primarily by reducing noise versus increasing information or reducing bias.

Our paper adds to a small but important literature studying the impact of human discretion over algorithmic or evidence-based guidelines. Hoffman, Kahn, and Li (2018) find that hiring managers with high override rates end up with worse overall hires, suggesting that discretion may decrease the accuracy of decisions in this setting. Abaluck et al. (2021) similarly find that most departures from the medical guidelines for atrial fibrillation patients are not justified by measurable treatment effect heterogeneity, with many physicians making decisions that are no better than random. Conversely, Finkelstein et al. (2022) find that physicians and their close relatives are less likely to adhere to medication-related guidelines, perhaps because of greater expertise or knowledge. Most recently, Agarwal et al. (2023) find that providing algorithmic predictions to professional radiologists does not increase the probability of making a correct decision on average unless contextual information is also included. We contribute to this literature by developing new tools to measure the impact of human discretion on the accuracy of decisions, showing that we can identify the required counterfactual outcomes using the quasi-random assignment of decision-makers to individuals. These tools are broadly applicable in settings where there is quasi-random variation in human decision-makers and the objective of these decision-makers is both known and well measured among the subset of individuals that the decision-maker selects for treatment.

The remainder of this paper proceeds as follows. Section II outlines the conceptual framework underlying our analysis. Section III describes the setting and data. Section IV develops and implements our quasi-experimental approach to estimating the impact of human discretion over an algorithm on the accuracy of decisions. Section V explores potential mechanisms, and Section VI concludes. The Online Appendix provides additional results and the details of our judge survey.

II Conceptual Framework

II.A Model Setup

We start by developing a general framework to study the impact of human discretion over an algorithm on the accuracy of decisions. We consider a setting where a set of human decision-makers indexed by j make binary decisions $D_{i,j} \in \{0,1\}$ across a population of individuals i who are differentiated by a latent indicator variable $Y_i^* \in \{0,1\}$. For each individual, there is a vector of characteristics that is available to both the algorithm and the human decision-maker $\mathbf{X}_i \in \mathscr{X}$ ("observable information") and another vector of characteristics that is not available to the algorithm but is available to the human decision-maker $\mathbf{V}_{i,j} \in \mathscr{V}$ ("private information"). We explain below that the observable information $\mathbf{X}_i$ and a meaningful subset of the private information $\mathbf{V}_{i,j}$ are observable to the econometrician.

Each decision-maker's goal is to align $D_{i,j}$ with Y_i^*, which captures the legitimate justification for setting $D_{i,j} = 1$. In the context of bail decisions, which we focus on in the remainder of this section, $D_{i,j} = 1$ indicates that judge j would release defendant i if assigned to her case (with $D_{i,j} = 0$ otherwise), while $Y_i^* = 1$ indicates that the defendant would subsequently fail to appear in court or be rearrested for a new crime if released (with $Y_i^* = 0$ otherwise). Each judge's legal objective is to release individuals without misconduct potential (set $D_{i,j} = 1$ when $Y_i^* = 0$) and detain individuals with misconduct potential (set $D_{i,j} = 0$ when $Y_i^* = 1$), but judges may differ in their predictions of which individuals fall into which category. We note

that $D_{i,j}$ is defined as the potential decision of judge j for defendant i, setting aside, for now, the judge decision rule that yields actual release decisions from these latent variables.

We define an algorithm by a mapping $a(\cdot) : \mathscr{X} \to [0,1]$ of the observable information $\mathbf{X}_i$. We similarly define an algorithmic recommendation as a suggested decision based on this mapping, such as $D_{i,s} = \mathbf{1}[a(\mathbf{X}_i) \leq s]$, where s is a threshold set by the algorithmic designer and represents the designer's preference for release. In our setting, $a(\mathbf{X}_i)$ is an algorithmic risk score that is meant to predict an individual's misconduct potential Y_i^* given observable case and defendant characteristics $\mathbf{X}_i$. Higher algorithmic risk scores are associated with a higher predicted misconduct potential such that the algorithm recommends releasing individuals with low-risk scores and detaining individuals with high-risk scores.

Each judge j observes the algorithmic risk score $a(\mathbf{X}_i)$, the algorithmic recommendation $D_{i,s}$, the observable information $\mathbf{X}_i$, and the private information $\mathbf{V}_{i,j}$. The judge uses this information to form a subjective prediction of misconduct potential, $h_{i,j}(a(\mathbf{X}_i), D_{i,s}, \mathbf{X}_i, \mathbf{V}_{i,j})$, from a mapping $h_j(\cdot) : \bar{a} \times \bar{D} \times \mathscr{X} \times \mathscr{V} \to [0,1]$, where $\bar{a} = [0,1]$ and $\bar{D} \in \{0,1\}$.[3] We assume that each judge releases individuals in order of her subjective prediction, implying that the judge's decision rule can be represented by a threshold τ_j with $D_{i,j} = \mathbf{1}[h_{i,j}(a(\mathbf{X}_i), D_{i,s}, \mathbf{X}_i, \mathbf{V}_{i,j}) \leq \tau_j]$, where she releases defendants with low perceived misconduct potential and detains defendants with high perceived misconduct potential. The release threshold τ_j can be interpreted as judge j's preference for release under a simple model where the judge weighs the expected perceived cost of pretrial misconduct relative to the perceived social benefit of release (e.g., Kleinberg et al. 2018). This decision rule results in a judge-specific release rate $R_j = E[D_{i,j}]$ and a judge-specific misconduct rate among released defendants $M_j = E[Y_i^* | D_{i,j} = 1]$.

One important feature of our model is that we do not assume that the judge agrees with the algorithm's release threshold. The judge may therefore override the algorithmic recommendations (i.e., $\exists\ i$ s.t. $D_{i,s} \neq D_{i,j}$) either because she prefers a different release rate or because she disagrees with the algorithm's misconduct predictions and rankings for some or all individuals. This issue has complicated efforts to measure the relative skill of one human decision-maker compared to another human decision-maker, with recent work using a combination of quasi-experimental variation and structural assumptions to overcome this identification challenge and to jointly identify predictive skill and preferences (e.g., Arnold, Dobbie, and Hull 2022; Chan, Gentzkow, and Yu 2022).

We measure the impact of human discretion over an algorithm on the accuracy of decisions by comparing each judge's observed misconduct rate to the counterfactual misconduct rate under the algorithm at the same release rate, thereby avoiding this identification challenge and isolating predictive skill at the judge's observed release rate. To build up to this measure, let the algorithmic release rule at judge j's existing release rate be

$$D_{i,s(j)} = \mathbf{1}[a(\mathbf{X}_i) \leq s(j)], \tag{1}$$

where $s(j)$ is the risk score threshold that results in the same release rate as judge j. Formally, let $s(j) = F^{-1}(G(\tau_j))$, where $G(\cdot)$ is the cumulative distribution function of $h_{i,j}$ and $F(\cdot)$ is the cumulative distribution function of $a(\mathbf{X}_i)$ such that $R_j = E[D_{i,s(j)}] = R_{s(j)}$.

[3] We include $a(\mathbf{X}_i)$, $D_{i,s}$, and $\mathbf{X}_i$ as separate inputs to the subjective misconduct prediction since the judge may not know the precise mapping from $\mathbf{X}_i$ to $a(\mathbf{X}_i)$ or may be more or less attentive to $a(\mathbf{X}_i)$ and $D_{i,s}$.

The counterfactual misconduct rate of the algorithm at the judge's existing release rate is then

$$M_{s(j)} = E[Y_i^* | D_{i,s(j)} = 1], \tag{2}$$

where, by design, $M_{s(j)}$ will only differ from judge j's conditional misconduct rate M_j if she disagrees with the algorithm's misconduct predictions and rankings for some or all individuals.[4]

The impact of human discretion on the accuracy of decisions can therefore be measured by comparing a judge's observed misconduct rate M_j to the counterfactual misconduct rate of the algorithm at the judge's existing release rate $M_{s(j)}$:

$$\Delta M_{j,s(j)} = M_j - M_{s(j)}, \tag{3}$$

where we say that judge j's discretion leads to less accurate decisions on average when $\Delta M_{j,s(j)} > 0$, more accurate decisions on average when $\Delta M_{j,s(j)} < 0$, and equally accurate decisions on average when $\Delta M_{j,s(j)} = 0$. The system-wide impact of human discretion on the accuracy of decisions is given by the case-weighted average of $\Delta M_{j,s(j)}$ across all judges. We note that $\Delta M_{j,s(j)}$ is not the only possible measure of human discretion on the accuracy of decisions. One may, for example, be interested in a counterfactual where the judges are allowed to reoptimize and choose a different release rate under the algorithm. We nevertheless believe that $\Delta M_{j,s(j)}$ provides an important benchmark to understand whether a human and algorithm working together can outperform an algorithm working alone.

There are two main reasons why the judge's subjective ranking of individuals may differ from the algorithm's ranking such that $\Delta M_{j,s(j)} \neq 0$. First, the judge may systematically over- or underweight the observable information $\mathbf{X}_i$ that is available to both the judge and the algorithm. For example, she may believe that a defendant's parole or probation status is a stronger predictor of misconduct potential than captured by the algorithm. The effect of such over- or underweighting on accuracy is theoretically ambiguous as many existing pretrial algorithms (including the one we study) are deliberately simple and may only roughly approximate $E[Y_i^* \mid \mathbf{X}_i]$. Second, the judge may use private information $\mathbf{V}_{i,j}$ that is not available to the algorithm. The effect of such private information on accuracy is also theoretically ambiguous since this information may either be a predictive or non-predictive signal of misconduct potential. Examples of predictive private information include relevant information such as the aggravating risk factors highlighted in the pretrial risk report, while examples of non-predictive private information include extraneous factors like whether a local football team won or lost that week (Eren and Mocan, 2018) and specific features of the defendant's appearance (Ludwig and Mullainathan, 2023). In Section V, we will provide suggestive evidence on the most likely reasons that the judges' subjective rankings differ from the algorithm's ranking.

Our framework relies on the assumption that the judges' objectives are well specified and, as a result, that the judges release individuals in order of their subjective prediction of misconduct risk. We view

[4] We can show this by first applying a probability integral transform to the rankings of both the judge and the algorithm to form random variables $U_i^j = F(h_{i,j})$ and $U_i^a = G(a(\mathbf{X}_i))$ such that $U_i^j \sim U[0,1]$ and $U_i^a \sim U[0,1]$. We can then rewrite the judge's decision rule as $D_{i,j} = \mathbf{1}[U_i^j \leq \bar{u}_j]$ and the algorithm's decision rule at the same release rate as $D_{i,s(j)} = \mathbf{1}[U_i^a \leq \bar{u}_j]$, for some judge-specific threshold $\bar{u}_j$ such that $E[D_{i,j}] = R_j$. The rewritten decision rules allow us to compare the judge's conditional misconduct rate $E[Y_i | U_i^j \leq \bar{u}_j]$ and the algorithm's conditional misconduct rate $E[Y_i | U_i^a \leq \bar{u}_j]$ using the same threshold $\bar{u}_j$, as both U_i^a and U_i^j share the same distribution. Any differences in the two conditional misconduct rates will therefore only come from differences in the two sets of rankings.

this assumption as reasonable in our setting as, by law, the judges are only permitted to consider the risk of pretrial misconduct when making release decisions. The judges are explicitly told that their release decisions cannot be based on other objectives and that they can only consider case and defendant characteristics if they are relevant to the risk of pretrial misconduct. However, there are at least two potential ways that this assumption could be violated in practice. First, the judges could care about certain types of pretrial misconduct like being rearrested for a new crime more than other types of misconduct like failing to appear. This issue could complicate the interpretation of our estimates if, for example, the judges we identify as low-skill perform relatively well for the types of misconduct that they care most about. In such a scenario, these judges may simply have different preferences and not lower skill. Second, the judges could care about objectives that are not permitted under the law like releasing certain types of defendants more than other types of defendants, even after accounting for the risk of misconduct, i.e. if τ_j is allowed to vary by $\mathbf{X}_i$ and $\mathbf{V}_{i,j}$. This issue could also complicate the interpretation of our estimates if, for example, the judges we identify as low-skill underperform the algorithm due to extra-legal objectives that the algorithm does not consider. In such a scenario, we can still say that these judges underperform the algorithm based on their legal mandate, but we cannot necessarily say that the reason for the judges' underperformance is solely due to misprediction of misconduct risk. In Section IV, we will explore these possibilities and show that our results are unlikely to be explained by the judges having different objective functions or preferences.[5]

II.B Graphical Intuition

Figure 1 illustrates the intuition for our approach using hypothetical variation in release and conditional misconduct rates. The solid curved line represents the counterfactual misconduct rate of the algorithm at each possible release rate, with the dashed vertical line at R_s denoting the release preference of the algorithmic designer at risk score threshold s. We also plot different hypothetical release rates and conditional misconduct rates for a judge j to illustrate different scenarios that highlight the logic underlying our approach. Panel A depicts a scenario where judge j has a lower release threshold than the algorithmic designers ($R_j < R_s$) and thus overrides the algorithmic recommendations. However, the judge overrides the algorithm solely because of differences in release preferences and thus follows the algorithm's ranking of individuals. As a result, her conditional misconduct rate is equal to the algorithm's conditional misconduct rate at her release rate such that $\Delta M_{j,s(j)} = 0$. We can therefore say that there is no impact of human discretion on the accuracy of decisions in this first example.

Panel B illustrates an alternative scenario where judge j has a higher release threshold than the algorithmic designers ($R_j > R_s$) and thus overrides some of the algorithmic recommendations. But now judge j overrides the algorithm because she has a different ranking of individuals compared to the algorithm due to the different use of observable information or reliance on private information. In this particular case, she can make more accurate predictions than the algorithm and thus achieve a lower conditional misconduct rate

[5]The single permissible objective of judges at the pretrial stage differs from later stages of the criminal justice system such as sentencing, where judges are permitted to consider multiple objectives at the same time (e.g., Stevenson and Doleac 2022). In these later stages of the criminal justice system, predictive algorithms are often based on just a subset of the objectives that judges are permitted to consider, making it challenging to estimate the impact of human discretion on the accuracy of decisions in such settings. Similar issues arise in other settings with multiple permissible objectives such as college admissions.

such that $\Delta M_{j,s(j)} < 0$. We can therefore say that human discretion increases the accuracy of decisions in this example. The same underlying logic applies to a scenario where the judge makes less accurate predictions than the algorithm and thus achieves a higher conditional misconduct rate, again due to the use of either observable or private information.

To summarize, we measure the impact of human discretion on the accuracy of decisions by comparing a judge's observed misconduct rate to the counterfactual misconduct rate under the algorithm at the judge's existing release rate. A judge who only overrides the algorithmic recommendation because she prefers a different release threshold will have the same conditional misconduct rate as the algorithm at the same release rate. In contrast, a judge who overrides the algorithm because she ranks individuals differently than the algorithm will have a lower (higher) conditional misconduct rate if she is more (less) skilled at predicting misconduct than the algorithm, again at the same release rate. The more (less) skilled judges may be better (worse) than the algorithm either because they are more (less) effective at using the observable information available to both the judges and the algorithm or because they are using predictive (non-predictive) private information that is not available to the algorithm.

II.C Empirical Challenges

Estimating the impact of human discretion over an algorithm on the accuracy of decisions is complicated by an important selection challenge. The available data suffer a missing data problem as we only observe misconduct outcomes among the selected subset of defendants that a judge chooses to release before trial. The selected nature of the data means that we cannot directly measure the conditional misconduct rate under an algorithmic counterfactual $M_{s(j)}$.

We formalize this econometric challenge in an idealized version of our setting with continuous algorithmic release thresholds s and unconditional random assignment of J total judges to defendants. Let $Z_{i,j} = 1$ if defendant i is assigned to judge j, let $D_i = \sum_j Z_{i,j} D_{i,j}$ indicate defendant i's release status, and let $Y_i = D_i Y_i^*$ indicate the observed pretrial misconduct outcome for the defendant. Importantly, $Y_i = 0$ when $D_i = 0$ regardless of individual i's misconduct potential Y_i^*. The econometrician observes $(\mathbf{X}_i, a(\mathbf{X}_i), D_{i,s}, Z_{i,1}, ..., Z_{i,J}, D_i, Y_i)$ for each defendant as well as some elements of $\mathbf{V}_{i,j}$ such as race and gender. With unconditional random assignment, $Z_{i,j}$ is independent of $(\mathbf{X}_i, a(\mathbf{X}_i), D_{i,s}, D_{i,j}, \mathbf{V}_{i,j}, Y_i^*)$.

We observe the judge's misconduct rate among released defendants M_j directly, as we observe misconduct potential Y_i^* among the defendants who the judge chooses to release before trial. However, we are unable to directly measure the misconduct rate under the algorithmic counterfactual, $M_{s(j)} = E[Y_i^* | D_{i,s(j)} = 1] = E[Y_i^* | a(\mathbf{X}_i) \leq s(j)]$. This is because there are generally some defendants whom the algorithm would release ($D_{i,s(j)} = 1$) but the judge does not ($D_{i,j} = 0$). Individuals who are detained by the judge ($D_{i,j} = 0$) cannot engage in misconduct, and so $Y_i = 0$ regardless of true misconduct potential Y_i^*.

We illustrate the importance of this selection challenge by considering a simple comparison of judge j's misconduct rate to the misconduct rate of the algorithmic counterfactual at score cutoff $s(j)$ using the

observed misconduct outcomes among released defendants:

$$\begin{aligned} & E[Y_i^*|D_{i,j}=1] - E[Y_i^*|D_i=1, a(\mathbf{X}_i) \leq s(j)] \\ & \quad = E[Y_i^*|D_{i,j}=1] - E[Y_i^*|a(\mathbf{X}_i) \leq s(j)] + E[Y_i^*|a(\mathbf{X}_i) \leq s(j)] - E[Y_i^*|D_i=1, a(\mathbf{X}_i) \leq s(j)] \\ & \quad = M_j - M_{s(j)} + \underbrace{E[Y_i^*|a(\mathbf{X}_i) \leq s(j)] - E[Y_i^*|D_i=1, a(\mathbf{X}_i) \leq s(j)]}_{\text{= Selection Bias}}, \end{aligned} \tag{4}$$

where the final line follows from the definitions of M_j and $M_{s(j)}$. The most important takeaway from Equation (4) is that a simple comparison based on the outcomes among released defendants will generally yield biased estimates of $M_{s(j)}$ and, as a result, biased estimates of $\Delta M_{j,s(j)}$. The exception is when judge release decisions are uncorrelated with misconduct potential among the relevant set of cases so that $E[Y_i^*|a(\mathbf{X}_i) \leq s(j)] = E[Y_i^*|D_i=1, a(\mathbf{X}_i) \leq s(j)]$.

We show in Section IV that we can recover unbiased estimates of $M_{s(j)}$ using the as-good-as-random assignment of judges to defendants. The unbiased estimates of $M_{s(j)}$ then allow us to calculate unbiased estimates of $\Delta M_{j,s(j)}$ and other statistics of interest, such as the share of low-skill judges, which we define as judges with $\Delta M_{j,s(j)} > 0$. Our quasi-experimental approach provides an alternative to other approaches such as imposing strong selection-on-observables assumptions or imposing worst- and best-case bounds on the algorithmic counterfactual.[6]

III Setting and Data

III.A Our Setting

We study the impact of human discretion on the accuracy of decisions in the context of a large, mid-Atlantic city that was one of the first jurisdictions in the country to introduce a pretrial risk assessment tool. The pretrial system is meant to allow the vast majority of criminal defendants to be released pending case disposition while minimizing pretrial misconduct. Bail judges are not meant to assess guilt or punishment when determining which individuals should be released from custody. In our setting, bail judges are directed by law to consider case and defendant characteristics only as that information is relevant to minimizing pretrial misconduct, defined as either failing to appear for a required court appearance (FTA) or being arrested for new criminal activity prior to case disposition (NCA), as measured by a new arrest. The bail judges in our setting are also explicitly told that they must consider a range of criteria when making release decisions and cannot simply make a decision on the basis of, for example, the offense or the defendant's residency status.

The pretrial services agency in our setting first started providing algorithmic risk scores and release recommendations in the mid-2000s. The risk assessment tool underlying the algorithmic risk scores and release recommendations was developed by the jurisdiction and tested on defendants arraigned and released

[6] Our quasi-experimental approach to measuring judge skill is much more informative than simple bounding approaches that do not exploit the quasi-random assignment of judges. Appendix Figure A.1 reports our main estimates using both best-case bounds, where we impute $Y_i^* = 0$ for all detained defendants, and worst-case bounds, where we impute $Y_i^* = 1$ for all detained defendants. These best- and worst-case bounds yield extremely wide estimates for the share of judges underperforming the algorithmic counterfactual, with anywhere from 0% to 98% of them being categorized as low skill. We explore more sophisticated bounding approaches that use the quasi-random assignment of judges below.

in the jurisdiction, and has been updated several times to reflect the most recent population of defendants. We study one of the most recent iterations of this locally-validated risk assessment tool, first implemented in 2016 and used until early 2020 when the jurisdiction stopped conducting in-person interviews due to the pandemic.

Figure 2 shows a redacted example of the pretrial risk assessment report provided to the bail judges in our setting. The report details the algorithmic release recommendation ($D_{i,s}$ in our model) and aggregate FTA and NCA scores ($a(\mathbf{X}_i)$) that range from 1 to 6. The report also includes information on the defendant's arrest date and date of birth and all of the risk factors used by the algorithm, such as the defendant's age at first arrest and number of prior arrests ($\mathbf{X}_i$). The report also includes private information that is not used by the algorithm, such as the defendant's race and gender, a detailed description of the charges, the defendant's residence, and a telephone number if available ($\mathbf{V}_{i,j}$). Additional private information comes from the pretrial services officer, who can recommend an override of the algorithm in specific situations with a supervisor's approval. These override recommendations occur in about 8% of the cases in our sample, with nearly all being harsher recommendations to detain. The pretrial services officer can also indicate additional factors not considered by the algorithm that support a potential override. These aggravating factors are noted in about 10% of cases, with nearly 8% due to the defendant's perceived threat of harm to the community.

The bail judge uses the information in this report to help decide whether to release on recognizance (ROR), release with non-monetary conditions, impose monetary bail, or detain the defendant. Release on non-monetary conditions includes release with the condition of reporting to a pretrial services officer; release on unsecured bail bond, which requires no money or deposit to secure release; and release on nominal bail, which requires the defendant to typically post $1.00 as a deposit and have a designated person or organization act as a surety. Release with monetary bail typically requires the defendant (or a third-party surety such as a bail bondsman) to pay 10% of the bail amount as a deposit. Figure 3 presents a timeline of the process in our setting.

The two most important features of the pretrial risk assessment report shown in Figure 2 are the risk scores and algorithmic release recommendations, both circled in red. This particular defendant has an NCA score of 4 and an FTA score of 5, with a corresponding algorithmic recommendation to release the individual with the condition to report to pretrial services in person. The risk scores are based on defendant demographics (e.g., age at the current arrest), defendant criminal history (e.g., age at first arrest, number of prior felony and prior misdemeanor convictions), case characteristics (e.g., number of pending charges and charge types), and current criminal justice status (e.g., parole/probation status and pretrial release status). The specific characteristics included in each risk score are based on whether there was a statistically significant association with the relevant misconduct outcome in the data used to build the algorithm, resulting in slightly different inputs for the two risk scores. The included characteristics are each associated with a certain number of points, which are aggregated to yield a detailed risk score that ranges from 0 to 73 for the FTA score and 0 to 30 for the NCA score. The scores are then binned into aggregate scores that range from 1 to 6, with lower scores indicating a lower probability of misconduct and higher scores indicating a higher probability of misconduct. The judges are only shown these aggregate NCA and FTA risk scores, but they were all trained on the risk scoring system and are provided with the complete set of factors used

in the risk scores if they want to calculate the more detailed FTA and NCA scores. The FTA and NCA risk scores are also highly predictive of pretrial misconduct among released defendants, closely tracking the performance of more sophisticated algorithms such as the gradient-boosted decision tree algorithm developed by Kleinberg et al. (2018) (see Appendix Figure A.2).

The second important feature of the risk assessment report is the algorithmic release recommendation, which is automatically generated using a combination of the FTA and NCA risk scores (see Appendix Table A.1). The lowest FTA and NCA risk scores yield a recommendation of release with no conditions, generally known as a release on own recognizance (ROR). These observably low-risk cases make up approximately 25% of the cases in our data. More moderate FTA and NCA risk scores yield a recommendation of release with regular phone or in-person check-ins, with these again relatively low-risk cases making up approximately 11% and 48% of the cases in our data, respectively. The highest NCA risk score of 6 yields a detention recommendation regardless of the FTA risk score, with these observably high-risk cases making up approximately 16% of the cases in our data. We focus on the NCA risk score throughout our analysis, as it alone generates the detention recommendation. We also see that release decisions are much more responsive to the NCA score in practice, with sharp changes in release rates as the NCA score changes from 5 to 6 (see Appendix Figure A.3).

There are also several pieces of private information that are not used by the algorithm but which could be salient to the bail judge in Figure 2, with some examples circled in blue. For example, the defendant is described as homeless and charged with endangering the welfare of children. The pretrial services officer further recommended an harsh override of the algorithm, noting three aggravating factors: that the defendant poses a threat to victim, witness or the community; that there is evidence of mental illness which may prove harmful to self or others; and that the defendant is currently out on pretrial release for similar charges. This private information can be used by the judge to decide the appropriate pretrial conditions.

The bail judges are free to override the algorithmic recommendation (and pretrial recommendation) for any reason, including but not limited to the private information discussed above. We see that judges in our sample override the default algorithmic recommendation in approximately 12% of cases where the algorithm recommends release ("low-risk cases") and in over half of cases where the algorithm recommends detention ("high-risk cases"), with an overall override rate of 18%. But at the same time, the judges in our setting are responsive to the algorithmic recommendations, with pretrial release rates sharply falling by nearly 14 percentage points (a 16% decrease from the mean release rate) when the algorithmic recommendation discontinuously changes from release to detain (see Appendix Figure A.4). These patterns indicate that judges do consider and respond to the algorithmic recommendation but nevertheless choose to override the recommendation in many cases. These patterns are consistent with recent descriptive work showing that judges frequently override the recommendations generated by predictive algorithms in criminal justice settings (e.g., Stevenson 2018; Stevenson and Doleac 2022; Albright 2023; Anwar, Bushway, and Engberg 2023). The combination of a long-standing risk assessment algorithm and frequent overrides for both observably low- and high-risk defendants makes this jurisdiction an ideal setting to study the impact of human discretion on the accuracy of decisions.

We exploit four additional features of the pretrial system in our setting for our analysis. First, the bail

judges generally make their pretrial decision before the bail hearing based on the pretrial risk assessment report described above (see Figure 2) and before speaking with the defendant, the defense attorney, or the prosecutor. The bail judges then announce these decisions at the bail hearing, where the defendant appears via videoconference from the local jail. The bail judge typically does not ask any questions of the defendant, with the hearings generally lasting less than five minutes. The prosecutor is almost never present at the bail hearing and public defenders only provide perfunctory information to the defendant. This unique feature of our setting where judges generally make their decision before the hearing means that the information contained in the pretrial risk assessment report largely captures the judges' information set when making a decision. We exploit this institutional feature in Section V when we explore the potential drivers of our results using data from the pretrial risk assessment reports for each defendant and coding the information contained in each.

Second, the legal objective of bail judges in our setting—to release the vast majority of individuals while minimizing pretrial misconduct—is both narrow and measurable. This narrow and measurable legal objective yields a natural approach to measuring the accuracy of decisions at a given release rate, with lower pretrial misconduct rates indicating more accurate decisions and higher pretrial misconduct rates indicating less accurate decisions. We also explore the importance of potential objectives such as racial fairness and specific forms of misconduct in robustness checks.

Third, we follow the prior literature in treating bail judges as effectively making binary decisions: releasing low-risk defendants (generally by ROR, non-monetary bail, or setting a low monetary bail amount) and detaining high-risk defendants (generally by setting a high monetary bail amount or outright detaining them). This binary classification is particularly well suited to our setting, as the judges are legally allowed to detain defendants when no other set of conditions could reasonably ensure the public's safety. Moreover, the predictive algorithm only makes release versus detain recommendations and recommends detaining observably high-risk defendants even when individuals are charged with less serious offenses such as misdemeanors. Recent work also finds that monetary bail does not change an individual's risk of misconduct conditional on release (Ouss and Stevenson, Forthcoming), suggesting that the use of monetary bail effectively serves as a de facto decision to release or detain. We return to this issue in Section V when considering the role of monetary bail in explaining our results.

Finally, the case assignment procedures used in our setting (and many other jurisdictions) generate quasi-random variation in bail judge assignment for defendants arrested at the same time and place. The quasi-random variation in judge assignment, in turn, generates quasi-experimental variation in the probability that a defendant is released before trial, which we exploit in our analysis. The specific court in our setting operates seven days a week and 24 hours a day and is staffed by approximately 60 judges on a rotating basis during our sample period. Daytime shifts are heard by a group of core judges whose full-time assignment is to the court, while nighttime shifts and weekend/holiday shifts are covered by a group of nearby judges and/or senior judges. Appendix Table A.2 confirms that judge assignment to cases is balanced on all characteristics observed in our data conditional on shift-by-time fixed effects, while Appendix Table A.3 shows that judge assignment has a strong first-stage effect on the probability that a defendant is released pretrial.

III.B Data and Summary Statistics

Our study is based on the universe of arraignments made in the jurisdiction's main jail between October 16, 2016, and March 16, 2020. The start of the period corresponds to when the jurisdiction adopted a recent iteration of its locally validated algorithm. The end of the period corresponds to when the jurisdiction temporarily stopped using the algorithm due to the pandemic.

The data contain information on offense type and each defendant's age at arrest, gender, race, prior criminal history, and prior pretrial misconduct. They also contain information on all of the factors used to calculate the FTA and NCA risk scores, the automatic algorithmic recommendation, the bail judge assigned to the case, whether the defendant was ultimately released before trial, and whether this release was due to ROR, release with non-monetary conditions, or release conditional on paying money bail. We categorize defendants as either released (including ROR, release with non-monetary conditions, and payment of money bail) or detained (including non-payment of money bail or outright detention). Among the subset of defendants released by the judge, we also observe whether a defendant subsequently failed to appear for a required court appearance or was arrested for new criminal activity before case disposition. We take either form of pretrial misconduct as the primary outcome of our analysis.[7] Finally, we collected copies of the pretrial risk assessment reports given to the bail judges for defendants arraigned during our sample period. We use these reports to capture all of the observable information that is used by the algorithm and all of the private information that is included in the reports.

We make the following restrictions to arrive at our estimation sample. First, we omit cases where we are missing risk scores or important demographic or case information (dropping 646 cases). Second, we focus on the first bail hearing for each case by dropping observations where the risk score was not recorded in the seven days before or after the bail date, following the guidance of the jurisdiction's pretrial services on how these cases are recorded in the data (dropping 12,848 cases). Third, we omit cases where there was a detainer hold on the defendant that would have prevented the judge from releasing the individual, even if the algorithm recommended release (dropping 2,144 cases). Finally, we omit observations where the case is assigned to a judge with fewer than 100 observations in our sample period (dropping 230 cases). These restrictions leave us with 37,855 cases among 27,503 unique defendants assigned to 62 unique bail judges.

Table 1 summarizes our estimation sample, both overall and by the algorithm's automatic recommendation. Panel A, column 1 shows that 83% of all defendants are released at some point before trial. Relatively few of these releases are without conditions, with 52% and 37% of released defendants having been assigned non-monetary and monetary conditions at the first bail hearing, respectively. There are also a small handful of defendants who are initially remanded without bail but are later released. Columns 2 and 3 report summary statistics for observably high-risk defendants whom the algorithm recommends detaining, separately by whether the judge overrides the recommendation ("lenient override") or follows the recommendation. Columns 4 and 5 present summary statistics for observably low-risk defendants whom the algorithm recommends releasing, separately by whether the judge overrides the recommendation ("harsh override") or follows the recommendation. Importantly, these release statistics indicate that judges override the algo-

[7] We observe that 1.4% of detained defendants (0.2% of all cases) commit pretrial misconduct in our data, likely due to miscodings in the court data. Our results are unchanged if we drop these cases.

rithm's recommendation in 54% of observably high-risk cases and 12% of observably low-risk cases. The total override rate is thus 18% after accounting for the distribution of observably low- and high-risk cases. On net, these overrides mean that observably high-risk defendants are less likely to be released than observably low-risk defendants, with a 54% release rate relative to an 88% release rate.

Panel B shows that the observably high-risk defendants who the algorithm recommends detaining are slightly younger at both the current and first arrest, have greater prior arrests and convictions, and are generally charged with greater and more serious crimes compared to the observably low-risk defendants who the algorithm recommends releasing. Observably high-risk defendants are also overall more likely to be on parole or probation and are more likely to be on an existing pretrial release at the time of the current arrest compared to observably low-risk defendants. Panel C further shows that observably high-risk defendants are also more likely to be male, less likely to be white, and more likely to be homeless and have a missing telephone.

Panels B and C also indicate that lenient overrides among observably high-risk cases and harsh overrides among observably low-risk cases are not random. Defendants for whom judges issue a lenient override have fewer prior arrests and convictions, are less likely to be on parole or probation, and are more likely to be charged with drug or traffic offenses. Defendants who receive lenient overrides are also less likely to be male, homeless, and missing a telephone, less likely to be charged with violent offenses, and more likely to have a pretrial services override recommendation. The pattern is generally reversed for harsh overrides, with these defendants having more prior arrests and convictions, being more likely to be on parole or probation, and being more likely to be charged with property and public order charges. These defendants are also more likely to be male, non-white, homeless, missing a telephone, more likely to reside out of state, and more likely to have an aggravating condition and a pretrial services override recommendation.

Panel D shows that observably high-risk defendants who are released despite the algorithm's detention recommendation are 14.8 percentage points more likely to be rearrested or have an FTA than low-risk defendants who are released in compliance with the algorithm's release recommendation. Among released defendants who commit pretrial misconduct, the vast majority are only rearrested for a new offense. Importantly, and in contrast to the other statistics in Table 1, the risk statistics in Panel D are only measured among released defendants, which are a selected sample of all defendants. Pretrial misconduct potential is, by definition, not observed among detained individuals.

IV Effects of Human Discretion on Pretrial Release Decisions

IV.A Methods

We measure the impact of human discretion on the accuracy of pretrial decisions by comparing each judge's observed misconduct rate, M_j, to our quasi-experimental estimate of the algorithmic counterfactual at the same release rate, $M_{s(j)}$. Our approach, as described below, requires that the judges' legal objectives are well specified such that judges release individuals in order of their subjective predictions of misconduct risk and that the threshold-specific average misconduct parameters used to construct the algorithmic counterfactual can be accurately extrapolated from the data.

Our approach for estimating the impact of human discretion over the algorithm proceeds in three main steps. First, we estimate the average misconduct potential among the subset of defendants with risk scores at or below a given risk score cutoff to solve the selection problem at a given algorithmic release threshold. The key insight underlying our approach is that the problem of measuring the algorithmic counterfactual at a given judge's release rate, $M_{s(j)} = E[Y_i^* | a(\mathbf{X}_i) \leq s(j)]$, is equivalent to the problem of estimating the average misconduct risk among the subset of defendants with risk scores at or below the relevant algorithmic release threshold, $s(j)$. When judges are as-good-as-randomly assigned, the average misconduct risk among the relevant subset of defendants at a given release threshold is common to all judges. As a result, the algorithmic counterfactual for all of the judges at that release threshold is captured by the threshold-specific misconduct risk.

The required threshold-specific misconduct risk parameters can be estimated from quasi-experimental variation in pretrial release and misconduct rates. To build intuition for our approach, consider a setting with as-good-as-random judge assignment and a hypothetical bail judge j^* who releases nearly all of the defendants with risk scores at or below a given algorithmic release threshold $s(j)$ for $j \neq j^*$, regardless of their potential for pretrial misconduct. This hypothetical judge may release all or nearly all defendants with risk scores below a given algorithmic release threshold for two different reasons. The first is that she is supremely lenient and releases all or nearly all defendants in the full population, not just at the algorithmic release threshold $s(j)$. The second is that she is supremely compliant with the algorithm at this particular risk score and releases all or nearly all defendants with risk scores at or below the release threshold $s(j)$ but detains some defendants with higher risk scores. In either case, this hypothetical judge's release rate among the relevant subset of defendants with risk scores at or below the release threshold $s(j)$ is close to 100%:

$$E[D_i \mid Z_{i,j^*} = 1, a(\mathbf{X}_i) \leq s(j)] = E[D_{i,j^*} \mid a(\mathbf{X}_i) \leq s(j)] \approx 100\%, \tag{5}$$

making the threshold-specific misconduct rate among defendants that the hypothetical judge releases close to the average misconduct risk in the relevant subset of defendants with risk scores at or below the release threshold $s(j)$:

$$E[Y_i \mid D_i = 1, Z_{i,j^*} = 1, a(\mathbf{X}_i) \leq s(j)] = E[Y_i^* \mid D_{i,j^*} = 1, a(\mathbf{X}_i) \leq s(j)] \approx E[Y_i^* \mid a(\mathbf{X}_i) \leq s(j)], \tag{6}$$

where the first equality in both expressions follows from the as-good-as-random assignment of judges. Without further assumptions, the decisions of such a judge that is as-good-as-randomly assigned can be used to estimate the threshold-specific average misconduct risk parameters needed for our analysis (i.e., $M_{s(j)}$ for all j).

In the absence of such a hypothetical judge, the required threshold-specific misconduct parameters can be estimated using model-based or statistical extrapolations of release and misconduct rate variation across as-good-as-randomly assigned judges. Our approach to estimating the required threshold-specific misconduct parameters is conceptually similar to how average potential outcomes at a treatment cutoff can be extrapolated from nearby observations in a regression discontinuity design, particularly so-called donut designs where the data in some window of the treatment cutoff are excluded. Here, released misconduct rates

are extrapolated from as-good-as-randomly assigned judges to the release rate cutoff of 100% for defendants with risk scores at or below a given algorithmic release threshold. Mean risk estimates may, for example, come from the vertical intercept at 100% of linear or local linear regressions of estimated misconduct rates among released individuals $E[Y_i^* \mid D_{i,j} = 1, a(\mathbf{X}_i) \leq s(j)]$ on estimated release rates $E[D_{i,j} \mid a(\mathbf{X}_i) \leq s(j)]$ across J judges at a given algorithmic release threshold $s(j)$. Our approach to estimating the algorithmic counterfactual at a given release threshold is closely related to Hull (2020) and Arnold, Dobbie, and Hull (2022), who consider different extrapolations of quasi-experimental moments in the spirit of "identification at infinity" in sample selection models.

Second, we create an algorithmic counterfactual for each judge in our sample by repeating these extrapolations for a wide range of risk score cutoffs that span the judges' existing release rates, i.e., $s(j)$ for all j. Repeating these extrapolations for a wide range of risk score cutoffs allows us to estimate $M_{s(j)}$ for all j in our sample. We can then construct $\Delta M_{j,s(j)}$ for each judge by comparing the judge's observed misconduct rate M_j to the quasi-experimental estimate of the algorithm's performance at the same release rate $M_{s(j)}$.

Finally, we create a summary measure of the impact of human discretion on the accuracy of decisions by calculating the share of judges with observed misconduct rates that are higher and lower than the algorithmic counterfactual, $\Delta M_{j,s(j)} > 0$ and $\Delta M_{j,s(j)} < 0$, respectively. The judge-level estimates of $\Delta M_{j,s(j)}$ are likely to involve substantial sampling error, however, particularly for the judges with relatively few case observations. As a result, the observed share of judges with misconduct rates that are higher and lower than the algorithmic counterfactual is unlikely to be a valid estimate of the true share of judges with such performance. We therefore adjust for the sampling error in our judge-level estimates using the posterior average effect approach of Bonhomme and Weidner (2022), which provides unbiased estimates of the true share of judges with misconduct rates that are higher and lower than the algorithmic counterfactual under relatively weak distributional assumptions.[8] We focus on the share of judges with misconduct rates that are higher or lower than the algorithm throughout the paper, as the judge-specific estimates are generally too imprecise to reliable identify the performance of each individual judge. We obtain standard errors for these estimates using a bootstrap procedure, where we first take independent random draws from the distributions of the estimated judge-specific release rate $\hat{R}_j$ and conditional misconduct rates $\hat{M}_j$ and then recalculate the threshold-specific extrapolations and statistics of interest.

We emphasize that our quasi-experimental approach to estimating the impact of human discretion on the accuracy of pretrial decisions only requires that the judges' legal objectives are well specified and the average misconduct parameters can be accurately extrapolated from the data. An important advantage of our approach is that it can be justified without a conventional first-stage monotonicity assumption (Imbens and Angrist, 1994; Heckman and Vytlacil, 2005), which recent work has questioned in the context

[8] Estimation of the posterior average effects depends on two main distributional assumptions. First, each judge-specific estimate $\Delta M_{j,s(j)}$ can be expressed as the sum of the unknown true judge-specific effect and a judge-specific error term, where the error term follows a known normal distribution, $\Delta\hat{M}_{j,s(j)} = \Delta M_{j,s(j)} + \varepsilon_j$, where $\varepsilon_j \sim N(0, \Sigma_j)$. Second, the unknown true judge-specific effects are drawn from a normal distribution with unknown hyperparameters, $\Delta M_{j,s(j)} \sim N(\Delta\bar{M}, \Lambda)$. Bonhomme and Weidner (2022) show that under these assumptions, the resulting posterior average effects have minimum worst-case specification error under various forms of misspecification. We also show in robustness checks that we obtain very similar estimates if we first estimate a population prior of the judges' conditional misconduct rates relative to the algorithm using a deconvolution approach and then estimate posterior means of each $\Delta M_{j,s(j)}$ estimate.

of quasi-random decision-makers (Chan, Gentzkow, and Yu, 2022; Frandsen, Lefgren, and Leslie, 2023). In our setting, for example, the conventional first-stage monotonicity assumption effectively requires that the judges are equally skilled at predicting misconduct outcomes—an assumption that is in direct conflict with our main results that show considerable variation in judge skill. Our extrapolated average misconduct parameters are valid so long as the *average* relationship between conditional misconduct rates and release rates across judges can be reliably estimated, which is likely to hold given the large number of judges with high release rates at most risk score cutoffs.

There are two practical complications that we must also consider in our setting. The first is that we only observe discrete algorithmic scores corresponding to the pretrial algorithm's scoring system rather than a continuous prediction of pretrial misconduct. Since observed judge release rates rarely coincide exactly with the release rate under the observed discrete risk score thresholds, this poses a challenge for identifying the algorithmic counterfactual at each judge's release rate ($M_{s(j)}$). This problem could be important if there are large changes in release rates across observed risk scores and steep changes in misconduct rates across risk scores. In practice, however, we have relatively closely spaced observed risk scores, and the relationship between misconduct and risk scores is relatively flat, making this issue less important in our setting. We therefore rely on a simple and transparent approach, where we use a linear spline connecting the observed threshold-specific estimates to obtain algorithmic misconduct estimates spanning all judge release rates observed in our setting. We will show that our main estimates are robust to alternative best- and worst-case approaches where we rely only on the assumption that the relationship between misconduct and risk scores is weakly monotonic.

The second practical complication is that the as-good-as-random assignment of judges to defendants is conditional on shift-by-time fixed effects in our setting. We follow Arnold, Dobbie, and Hull (2022) and account for these shift-by-time effects using linear regression adjustment, which tractably incorporates a large number of shift-by-time fixed effects under an auxiliary linearity assumption. Specifically, we estimate judge-specific release and misconduct rates accounting for shift-by-time effects using the following OLS regressions:

$$D_i = \sum_j \zeta_j Z_{ij} + \mathbf{W}_i' \gamma^R + u_i, \tag{7}$$

$$Y_i = \sum_j \rho_j Z_{ij} + \mathbf{W}_i' \gamma^M + v_i, \tag{8}$$

where D_i indicates whether defendant i was released pretrial, Y_i indicates whether defendant i committed pretrial misconduct among released individuals ($D_i = 1$), and $\mathbf{W}_i$ is a set of shift-by-time fixed effects. We then truncate the estimated release parameters so that they lie in $[0,1]$ after adjusting for the shift-by-time fixed effects, impacting 5 of the 558 judge-level moments used in the extrapolations for our main results (1% of the total). We next use the estimates of $R_j = E[D_{i,j}]$ and $M_j = E[Y_i | D_{i,j} = 1]$ to extrapolate the average misconduct parameters for a range of algorithmic score cutoffs, allowing us to construct $\Delta M_{j,s(j)}$ for each judge j, after accounting for shift-by-time effects. We will show in robustness checks that our results are similar when we estimate Equations (7) and (8) without shift-by-time fixed effects.

The linear regression adjustment in Equation (7) is appropriate if release rates are linear in the shift-

by-time effects for each judge, with constant coefficients,i.e., when $E[D_{ij} \mid \mathbf{W}_i] = \phi_j + \mathbf{W}_i'\gamma^R$. Similarly, a sufficient condition for Equation (8) to consistently estimate misconduct rates among released defendants is $E[Y_i^* \mid D_{ij} = 1, \mathbf{W}_i] = \psi_j + \mathbf{W}_i'\gamma^M$. Intuitively, both conditions require the shift-by-time effects to shift judge actions similarly across the judges j. A judge who is lenient in one shift and time period is thus restricted to still be lenient in different shifts and time periods.

IV.B Counterfactual Misconduct Under the Algorithm

Figure 4 illustrates our extrapolation-based estimation of the average misconduct risk for two algorithmic release thresholds. Panel A reports results for the full sample of cases, corresponding to an algorithmic release threshold of 100%. Panel B restricts the sample to cases where the algorithm recommends release, corresponding to an algorithmic release threshold of just over 80%. The horizontal axis in each panel plots estimated release rates among defendants with algorithmic risk scores below the respective threshold ($\hat{R}_j$) for each of the 62 judges in our data after regression adjustment for shift-by-time fixed effects but not for sampling error. We find sizable variation across judges at each risk score threshold, with many judges releasing a high fraction of defendants. The vertical axis plots estimated conditional misconduct rates for each judge among defendants with algorithmic risk scores below the respective threshold ($\hat{M}_j$), again adjusted for shift-by-time fixed effects but not for sampling error.

The vertical intercepts of the lines of best fit, at 100%, provide estimates of the threshold-specific average misconduct rates. The lines of best fit are obtained by OLS regressions of judge-specific conditional misconduct rate estimates on judge-specific release rate estimates, with the judge-level regressions weighted inversely by the variance of misconduct rate estimation error. We obtain standard errors using a bootstrap procedure, where we first take independent random draws from the distributions of the estimated judge-specific release rates ($\hat{R}_j$) and conditional misconduct rates ($\hat{M}_j$) and then recalculate the threshold-specific extrapolations. These estimates and associated standard errors are reported at the bottom of each panel. The simple linear extrapolation yields a precise mean misconduct estimate of 14.7% (SE: 1.0) for the full population of cases, which corresponds to a release rate of 100%. The extrapolation yields a mean misconduct estimate of 13.8% (SE: 0.8) for cases where the algorithm recommends release, which again corresponds to a release rate of just over 80%. The results are similar using a local linear extrapolation, which yields a mean misconduct estimate of 13.5% (SE: 1.3) for the full population of cases and 12.8% (SE: 0.9) for cases where the algorithm recommends release.

We repeat these extrapolations for the algorithmic risk score cutoffs that correspond to release rates ranging from just under 70% to 100%, spanning both the unadjusted and regression-adjusted judge release rates observed in our sample. The results from these extrapolations are plotted in Figure 5 and reported with standard errors in Appendix Table A.4. The extrapolations show that, in practice, we observe risk scores that correspond to closely spaced release rates (column 1, Appendix Table A.4) and that there is a (weakly) monotonically increasing relationship between risk scores and conditional misconduct rates (columns 2–3, Appendix Table A.4). These threshold-specific estimates allow us to measure the counterfactual misconduct rate under the algorithm and to construct $\Delta M_{j,s(j)}$ for each judge j in our sample using the linear spline approach discussed above. We take the linear extrapolation as our baseline specification for estimating

the impact of judge discretion on the accuracy of decisions and explore the robustness of our results to alternative mean risk estimates below.

IV.C Effect of Judicial Discretion on Conditional Misconduct Rates

Figure 5 presents our main findings on the effect of human discretion over an algorithm on the accuracy of pretrial decisions. Each of the 62 judges in our sample is represented by a green dot that shows the judge's estimated conditional misconduct rate $\hat{M}_j$ against the judge's estimated release rate for all cases $\hat{R}_j$, where both are adjusted for shift-by-time fixed effects but not for sampling error. The dashed orange line shows the estimated conditional misconduct rate for the algorithm at different release rates, estimated using linear extrapolations of average misconduct at each discrete risk score threshold and connected using a linear spline. The solid navy line shows the conditional misconduct rate under a random release rule, estimated using a linear extrapolation of average misconduct in the full sample. As our key summary measure, we also report the share of judges with conditional misconduct rates that are higher than the algorithmic counterfactual and the random release rule, estimated using the posterior average effect approach of Bonhomme and Weidner (2022) that accounts for the sampling error in our judge-level estimates. Standard errors for these estimates are constructed using the bootstrap procedure described above.

Three striking patterns emerge from the distribution of estimated judge conditional misconduct rates and release rates, even before we compare the judges to the algorithmic counterfactual and random release rule. First, there is substantial variation in judges' preferences for release, with the regression-adjusted release rates ranging from just above 70% to over 95%. These patterns are consistent with prior work on the pretrial system (Arnold, Dobbie, and Yang, 2018; Arnold, Dobbie, and Hull, 2022) and highlight the importance of comparing each judge's outcomes to the algorithmic counterfactual holding the release rate constant. Second, there is also substantial variation in the judges' conditional misconduct rates at the same release rate, with a judge at the case-weighted 90th percentile of conditional misconduct rates having a misconduct rate that is 6.6 percentage points higher than a judge at the case-weighted 10th percentile. These findings suggest important variation in predictive skill across the judges and provide evidence against the standard monotonicity assumption, which implies that judges with the same release rate will have the same conditional misconduct rate (Chan, Gentzkow, and Yu, 2022; Frandsen, Lefgren, and Leslie, 2023). Third, the judges' conditional misconduct rates do not increase with the release rate, with an OLS regression of the judge-specific misconduct rates on the judge-specific release rates yielding a coefficient of –0.05 (SE: 0.08). This pattern is inconsistent with standard models where the monotonicity assumption holds, which generally imply that the conditional misconduct rate should increase with the release rate (Chan, Gentzkow, and Yu, 2022).

The comparison of the judges and algorithmic counterfactual reveals an even more striking pattern—the vast majority of judges significantly underperform the algorithm, as indicated by a conditional misconduct rate that is above the algorithmic counterfactual at the same release rate. Using the posterior average effect approach to account for sampling error, we estimate that 90% (SE: 6.1) of judges generally underperform the algorithm when they make discretionary overrides, with a remarkable 69% (SE: 14.1) of judges underperforming the random release rule. These findings mean that, incredibly, most judges could achieve a lower

misconduct rate by flipping a coin or using a random number generator. The system-wide impact of human discretion on the accuracy of release decisions is correspondingly negative, with the judges increasing pretrial misconduct by an average of 2.4 percentage points (SE: 0.5) at their existing release rates relative to the algorithm (column 1, Appendix Table A.7). These findings therefore indicate that the typical judge in our setting is less skilled at predicting misconduct than the algorithm and that we could substantially decrease misconduct by automating release decisions.

Our estimates show that there are both statistically and economically significant costs of human discretion over an algorithm in our setting. We can calculate the social costs of human discretion relative to automated release decisions in two ways. First, we can calculate the social costs of the higher pretrial misconduct rate at the judges' existing release rates. Our estimates imply that we could reduce the number of pretrial misconduct offenses by about 300 per year using automated release decisions at the judges' existing release rates. Reducing the number of pretrial misconduct offenses by this amount would generate about $2.8 million in social cost savings per year based on estimates from Dobbie, Goldin, and Yang (2018) and Miller et al. (2021).[9] Second, we can calculate the social costs of the lower pretrial release rates at the judges' existing misconduct rates. The extremely high misconduct rates among the judges in our sample mean that we can achieve a lower misconduct rate by releasing all defendants. This surprising result is due to the fact that the average conditional misconduct rate in our sample is 15.9% (column 1, Table 1), whereas the misconduct rate in the full population is only 14.7% (Panel A, Figure 4). Increasing the release rate to 100% means that we could release about 1,950 additional individuals per year, generating about $58 million in recouped earnings and government benefits annually based on the estimates from Dobbie, Goldin, and Yang (2018).

However, this negative system-wide impact of discretion on the accuracy of decisions masks substantial variation in the judges' performance compared to the algorithm. Importantly, we find that 10% of the judges generally outperform the algorithm when they make discretionary overrides, as indicated by a conditional misconduct rate that is lower than the algorithmic counterfactual at the same release rate. This more positive finding suggests that a human and an algorithm working together can outperform automated release decisions in at least some situations and that a human can still add value to the decision-making process. This finding also suggests that there will not necessarily be a single optimal policy on human oversight of algorithms, as the impact of such policies depends on the ability of the human decision-makers in a particular context.

There is also a relatively consistent set of high- and low-skill judges that drive the variation in performance relative to the algorithm, suggesting that our measure of accuracy captures skill and not simply idiosyncratic luck. We show the out-of-sample predictive power of our judge performance measure by computing separate $\Delta M_{j,s(j)}$ estimates in the first and second half of cases that each judge sees in our sample period. We then estimate OLS regressions of the most recent judge performance estimates on lagged judge

[9] We first calculate the average social cost estimate for FTAs, DUIs, drug offenses, motor vehicle offenses, person offenses, property offenses, public order offenses, and weapons offenses. Within each of these misconduct types, we take the lowest social cost estimate to provide the most conservative estimate possible (e.g., we use the cost estimate for assault instead of murder for persons offenses). We then use the frequency of each misconduct type among released defendants to calculate the average social cost of misconduct.

performance posteriors weighted by the inverse variance of the most recent judge performance estimates. We compute these posteriors using a conventional empirical Bayes "shrinkage" procedure. Appendix Figure A.5 reports these OLS estimates for the 54 judges who see at least 100 cases in each period. We find that the correlation between the most recent performance estimates and lagged performance posteriors is 0.64 after accounting for statistical noise, indicating significant out-of-sample forecasting power even when using only half of the available data. The autoregressive coefficient of 0.69 (SE: 0.27) can also be interpreted directly, indicating that a judge who is predicted to have a conditional misconduct rate that is 1 percentage point higher than the algorithmic counterfactual in the first period will, on average, have a conditional misconduct rate that is 0.69 percentage points higher than the algorithmic counterfactual in the second period.

We next ask whether we can identify high- versus low-skill judges on the basis of judge characteristics. While the judge-specific estimates are statistically noisy, we can still meaningfully divide the judges into two exhaustive and mutually-exclusive groups based on their likely performance. The first group consists of low-skill judges who are likely to underperform the algorithmic counterfactual at the same release rate, defined as those with a posterior probability of $\Delta M_{j,s(j)} > 0$ that is 0.9 or above. The second group consists of high-skill judges who are more likely to outperform the algorithmic counterfactual at the same release rate, defined as those with a posterior probability of $\Delta M_{j,s(j)} > 0$ that is below 0.9.

Table 2 shows results from OLS regressions of an indicator for being a high-skill judge on different judge characteristics to better understand the variation in judge performance. There is no statistically significant relationship between judge performance and propensity to override the algorithm (column 1), ruling out the explanation that high-skill judges are simply more likely to follow the algorithmic recommendations compared to low-skill ones. In addition, there is no statistically significant relationship between judge performance and years of experience, gender, race, political affiliation, and whether the judge has a law degree or is a former prosecutor (columns 2–7). The only statistically significant predictor of judge performance is a history in law enforcement, with former police officers being 24.6 percentage points (SE: 11.8) less likely to be classified as high-skill (column 8). We find a similar pattern using our continuous measure of $\Delta M_{j,s(j)}$ as the dependent variable in Appendix Table A.5, where the only statistically significant predictor of judge performance is a history in law enforcement. The lower performance for former police officers could be associated with the unique cognitive demands of police work, where officers often need to make fast decisions without sufficient consideration of alternative options (Dube, MacArthur, and Shah, 2022). However, none of the estimates from Table 2 or Appendix Table A.5 suggests a clear explanation for the variation in judge performance.

IV.D Robustness and Extensions

Our approach to measuring the impact of human discretion on the accuracy of decisions requires that the average misconduct risk among defendants at different release rates can be accurately extrapolated from the quasi-experimental data and that the judges' objectives are well specified. We now consider several extensions to our main results that test or relax these assumptions. We also confirm the robustness of our results to different sampling error adjustments and algorithmic comparisons.

Extrapolations of Misconduct Risk. Our baseline specification estimates the required average misconduct risk parameters using a series of linear extrapolations that control for shift-by-time fixed effects. We then connect the extrapolation-based estimates at each discrete risk score with a linear spline. We consider a range of alternative specifications when estimating the distribution of judge performance relative to the algorithm, finding qualitatively similar results when using local linear extrapolations in Appendix Figure A.6, linear extrapolations that omit the shift-by-time fixed effects in Appendix Figure A.7, and best- and worst-case step functions that connect the discrete risk scores in Appendix Figure A.8. We also find qualitatively similar results when we use a modified estimation approach where we only extrapolate to the most lenient judge at a given risk score cutoff and then calculate bounds for the remaining share of defendants in Appendix Figure A.9.[10] In this final specification where we only extrapolate to the most lenient judge, for example, we find that at least 87% (SE: 7.7) of judges underperform the algorithm, compared to 90% (SE: 6.1) in our baseline specification. Taken together, these findings suggest that the average misconduct risk among defendants at different release rates can be accurately extrapolated in our data and is robust to a range of alternative specifications.

Judge Objectives. Our approach also requires that the judges' objectives are well specified. One particularly important concern is that the judges in our setting may care about certain types of pretrial misconduct more than other types of misconduct and that the judges we identify as low-skill may prioritize more serious forms of misconduct like NCA or violent NCA over less serious forms of misconduct like FTA. We explore this concern by considering several alternative definitions of pretrial misconduct, finding qualitatively similar results when using only FTA in Appendix Figure A.10, only NCA in Appendix Figure A.11, and only violent NCA in Appendix Figure A.12. We also find very similar results when we use a modified misconduct measure that captures the social cost of different types of misconduct in Appendix Figure A.13.[11] In this last specification where we use a misconduct measure that captures the social cost of different misconduct types, for example, we find that 91% (SE: 6.1) of judges underperform the algorithm, again compared to 90% (SE: 6.1) in our baseline specification. We also find that more than 90% of the judges we identify as likely low-skill in our baseline specification continue to underperform the algorithm when we use this new social cost measure. These findings indicate that our main results are unlikely to be explained by the low- and high-skill judges having different preferences across the more and less serious forms of pretrial

[10] We extrapolate to the most lenient judge at a given risk score before calculating bounds to reduce sampling variation and leverage all the available data. The modified estimation approach proceeds in four steps. First, we identify the release rate of the most lenient judge for each risk score cutoff in the data. Second, we estimate the conditional misconduct rate at the most lenient judge's release rate using our extrapolation-based procedure. Third, we construct worst- and best-case bounds for the remaining share of defendants. Finally, we calculate the counterfactual misconduct rate using a weighted average of the misconduct rate from the extrapolation at the most lenient judge's release rate and the bounding procedure for the remaining share of defendants. Our results are very similar using this modified approach as we generally observe at least one judge that releases 100% of defendants at the lower risk scores.

[11] We first calculate the social cost for FTAs, DUIs, drug offenses, motor vehicle offenses, persons offenses, property offenses, public order offenses, and weapons offenses using the estimates from Dobbie, Goldin, and Yang (2018) and Miller et al. (2021). Within each of these misconduct types, we use the lowest social cost estimate available (e.g., we use the cost estimate for assault instead of murder for persons offenses). We next use these estimates to calculate the social cost of releasing each defendant, where we assign a social cost of $0 to cases without any misconduct and the lowest available cost amount to the 1% of cases with a misconduct type that is not categorized. We then perform our main analysis using the new social cost measure as the outcome variable.

misconduct.

A second concern is that the judges in our setting may care about both racial fairness and overall accuracy and that the judges we identify as low-skill may put more emphasis on racial fairness and less emphasis on accuracy. We explore this concern by measuring release disparities between white and non-white defendants with identical misconduct potential in Appendix Figure A.14, along with counterfactual racial disparities for the algorithm.[12] We find that, on average, approximately one-third of the judges in our setting generate lower release disparities than the algorithm when they make a discretionary override, leading to generally fairer release decisions than the algorithm alone. Most importantly, we find that there is a positive correlation between outperforming the algorithm in terms of accuracy and in terms of racial fairness in Table 2, columns 9 and 10. These findings indicate that the high-skill judges generally outperform the algorithm in terms of both accuracy and racial fairness, decreasing (but not eliminating) release disparities between white and non-white defendants with identical misconduct potential. Our main results are therefore unlikely to be explained by the low-skill judges putting more emphasis on racial fairness and less emphasis on accuracy.

A final concern is that the judges in our setting may consider a range of other objectives that are not permitted under the law, from the employment or family status of individuals to explicit discrimination against defendants based on their demographic characteristics or criminal charge. Such extra-legal objectives could complicate the interpretation of our estimates if the judges we identify as low-skill intentionally release individuals with a higher risk of pretrial misconduct due to these other objectives. We explore this concern by measuring judge performance separately for different defendant and case subgroups, finding very similar results by defendant race in Appendix Figure A.15, defendant age in Appendix Figure A.16, defendant education in Appendix Figure A.17, and case charge severity in Appendix Figure A.18. We find that at least 74% (SE: 6.4) of judges underperform the algorithm in any of these subgroups, indicating that our main results are unlikely to be explained by the low- and high-skill judges having different objective functions or preferences for release across different defendant subgroups.

Sampling Error. Our baseline estimates use the posterior average effect approach of Bonhomme and Weidner (2022) to account for sampling error in our judge-level estimates of $\Delta M_{j,s(j)}$. We show in Appendix Figure A.19 that we obtain very similar estimates of the distribution of $\Delta M_{j,s(j)}$ if we use the non-parametric empirical Bayes deconvolution approach of Efron (2016), where we first estimate a population prior of the judges' conditional misconduct rates relative to the algorithm and then estimate posterior means of each $\Delta M_{j,s(j)}$ estimate. We find very similar estimates of the share of judges with higher misconduct rates than the algorithm under the deconvolution approach at 92% compared to 90% (SE: 6.1) using the posterior average effects method. Both estimates are substantially larger than the unadjusted share of judges with higher misconduct than the algorithm, reflecting shrinkage due to the noise in the observed data.

Algorithmic Design. Our baseline analysis compares judges' outcomes to the recommendations of the proprietary algorithm that the judges see at the time of the bail decision. The algorithm itself is based on a

[12] We measure these release disparities using estimates of race-specific misconduct risk to rescale observational release rate comparisons in such a way that makes released white and non-white defendants comparable in terms of misconduct potential within each judge's defendant pool (Arnold, Dobbie, and Hull, 2022).

simple scoring system, which assigns points to various demographic, criminal history, and current charge characteristics as described in Section III. A natural concern is that the best judges in our setting may still underperform a more sophisticated machine learning algorithm. We explore this concern by comparing the judges to the gradient-boosted decision tree algorithm developed by Kleinberg et al. (2018) in Appendix Figure A.20.[13] We find that this more sophisticated machine learning algorithm is more accurate than the proprietary algorithm, with misconduct decreasing by an average of 0.4 percentage points across the observed release thresholds. The share of judges underperforming the more sophisticated algorithm is correspondingly higher than the share underperforming the proprietary algorithm, increasing to 95% (SE: 6.0).

V Potential Mechanisms

This section provides more suggestive evidence on the behavior underlying the differences in judge performance. We start by showing that the high- and low-skill judges use the observable information that is available to both the judges and the algorithm in a remarkably similar way, suggesting that their performance differences are likely driven by the use of private information that is not available to the algorithm. We then examine three particular ways that private information can result in inconsistency and noise. We conclude by using new survey data collected for this study to show the importance of the judges' stated preferences in explaining our findings.

V.A Observable Information

One explanation for our results is that the high- and low-skill judges differ in their use of the observable information $\mathbf{X}_i$ available to both judges and the algorithm. Table 2 shows that, on average, the high- and low-skill judges are equally likely to override the algorithm, meaning that any performance differences from using such observable information must be driven by the types of defendants for which the high- and low-skill judges choose to override the algorithm for.

First, we explore the possibility that the high- and low-skill judges override the algorithm across different parts of the risk score distribution in Figure 6. Overrides at different parts of the risk score distribution could explain our results if, for example, the high-skill judges only override the algorithm for particularly "close calls" where the recommendation changes from release to detain but the low-skill judges override the algorithm throughout the risk score distribution. The high- and low-skill judges could also react differently when the algorithmic recommendation discontinuously changes from release to detain, e.g., due to different cognitive biases related to anchoring. However, the data do not support these hypotheses. Figure 6 shows nearly identical release rates for high- and low-skill judges throughout the risk score distribution. Both types

[13] We follow Kleinberg et al. (2018) closely in constructing the gradient-boosted decision tree algorithm. The model is trained on the sample of released defendants for whom we observe pretrial misconduct, using the same algorithmic inputs as the proprietary algorithm. We rely on fivefold cross-validation to iterate over grids of values, and determine the optimal values of these parameters. Broadly speaking, a decision tree algorithm such as ours partitions the data space using binary splits. For example, an initial split might be based on the defendant's age; the second could further split the two resulting subsamples by the number of prior arrests. A gradient-boosted decision tree algorithm grows many such decision trees sequentially and then averages over the predictions of each iteration to form a final prediction for each observation.

of judges release most observably low-risk defendants, with release rates declining monotonically with the risk score. Both types of judges also exhibit nearly identical decreases in release rates at the risk score threshold where the algorithmic recommendation changes discontinuously.

We next explore the possibility that the high- and low-skill judges override the algorithm for observably different types of defendants within narrow risk score bins. Releasing observably different defendants within risk scores could explain our findings if, for example, the high-skill judges use additional information contained in the observable characteristics to release only the low-risk defendants. We explore this possibility in Panel A of Table 3 by regressing an indicator for pretrial release on the full set of algorithmic inputs $\mathbf{X}_i$ separately for high- and low-skill judges. We also control for detailed risk score fixed effects so that the coefficients reflect the additional weight (in percentage points) that judges place on each characteristic relative to the risk score. Column 1 reports these results for high-skill judges, column 2 reports these results for low-skill judges, and column 3 reports the p-value from a test of equality of the coefficients between high- and low-skill judges. We find that the high- and low-skill judges use the observable information that is available to both the judges and the algorithm in a remarkably similar way, even within these narrow risk score bins. While both types of judges do place additional weight on certain characteristics, they generally do so in the same direction. For example, high- and low-skill judges are both about 2 percentage points more likely to release defendants charged with a drug offense. Both sets of judges are also about 13-15 percentage points less likely to release defendants who are currently on probation or parole. However, there are no observable characteristics where there is a statistically significant difference between high- and low-skill judges at the 5% level. We find that the observable information also explains about the same amount of variation in the release decisions of the high- and low-skill judges, with an R^2 of 0.21 for the high-skill judges and 0.20 for the low-skill judges, and that the results do not change when we focus on just the observably high-risk defendants recommended for detention or the observably low-risk defendants recommended for release (Appendix Table A.6).

These findings suggest that the differences in judge performance are not driven by the use of such observable information but rather by the use of private information that is not available to the algorithm. We provide more direct support for this idea by building a new algorithm that predicts the high- and low-skill judges' release decisions using all of the information that is observable to both the judges and the original algorithm using a gradient-boosted decision tree, following Kleinberg et al. (2018). These predicted release decisions are then used to construct counterfactual decision rules for the high- and low-skill judges, where we order defendants in terms of their predicted probability of release and define judge-specific thresholds that yield each judge's original release rate. We then estimate the counterfactual misconduct rate under the predicted release decisions of each judge using the extrapolation approach described above, again separately for high- and low-skill judges. These extrapolations allow us to construct the misconduct rate under the high- and low-skill predicted judge release rule, holding fixed each judge's release rate. The results from this exercise tell us how the high- and low-skill judges would have performed if they had simply followed their own release tendencies using observable information (and/or private information correlated with this observable information set). We can then use the difference between the judges' actual performance and their predicted release decisions to shed light on the importance of private information, which presumably

drives most of the deviations from the predicted release tendencies.

Panel A of Figure 7 presents the conditional misconduct rate under the high- and low-skill predicted judge release rules. The solid red line shows the estimated conditional misconduct rate for high-skill predicted release decisions, while the solid blue line shows the estimated conditional misconduct rate for low-skill predicted release decisions. The dashed orange line shows the estimated conditional misconduct rate for the algorithm at different release rates for comparison. The misconduct rate under high- and low-skill predicted release rules is modestly higher than the misconduct rate under the original algorithm, suggesting that both types of judges make modest but predictable errors in their use of observable information that leads to some underperformance compared to the algorithm. Most importantly, the predicted decision rules yield nearly identical misconduct rates across the entire observed release distribution. These results again indicate that the high- and low-skill judges use the observable information available to both them and the algorithm in a very similar way, consistent with our findings above. But Figure 7 also shows that the high-skill judges consistently outperform their predicted judge release rule, suggesting that they can productively use private information that is not available to the algorithm to improve the accuracy of their decisions. By comparison, the low-skill judges underperform their predicted judge release rule, suggesting that they are instead adding noise and inconsistency to their decisions when they attempt to use such additional information.

V.B Private Information

Our results above show that high- and low-skill judges use the observable information $\mathbf{X}_i$ available to both them and the algorithm in a very similar way, suggesting that a key difference between them may be how they use private information $\mathbf{V}_{i,j}$ that is not available to the algorithm. We now provide three sets of results to more directly support this explanation and show how private information leads to noise and inconsistency for low-skill judges.

Pretrial Risk Assessment Reports. The first way private information may add inconsistency and noise for low-skill judges is the over- or underweighting of information contained in the detailed pretrial risk assessment reports, which largely capture the information set available to judges when they make their release decisions. We use the pretrial risk reports to encode several pieces of private information not used by the algorithm, such as demographic information on gender and race, whether the defendant is homeless, whether the defendant is missing a telephone number, and whether the detailed charge involves violence against adults or children. We also code whether the pretrial services officer recommended an override of the algorithmic recommendation and any aggravating factors supporting this override recommendation, such as the defendant posing a threat to others or suffering from mental illness. We explore whether the high- and low-skill judges use this private information differently in Panel B of Table 3 by regressing an indicator for pretrial release on this private information $\mathbf{V}_{i,j}$ separately for high- and low-skill judges, controlling for detailed risk score fixed effects and the full set of observable characteristics $\mathbf{X}_i$. We also construct a new algorithm that predicts the judges' release decisions using both the full set of observable characteristics $\mathbf{X}_i$ and the new private information $\mathbf{V}_{i,j}$ separately for high- and low-skill judges, with the conditional misconduct rates under the new release rules shown in Panel B of Figure 7.

We find several notable differences in how these judges use private information from the pretrial risk assessment reports. The estimates from Panel B of Table 3 show that the low-skill judges are much less likely to release defendants with no telephone number or defendants with an out-of-state address compared to high-skill judges, despite these factors not being particularly predictive of the risk of misconduct in unreported results. The low-skill judges are also much more likely to release individuals charged with a violent crime against an adult and defendants with aggravating factors compared to the high-skill judges, despite these factors being predictive of risk in unreported results. Panel B of Figure 7 also shows that there is a modest deterioration in performance for low-skill judges and a modest improvement in performance for the high-skill judges when we use both observable and private information to predict release decisions relative to use of only observable information (Panel A), although we caution that the estimates are statistically imprecise. Nevertheless, these results suggest that there are systematic differences in the use of private information that may play some role in explaining the performance differences across judges.[14]

Financial and Non-Financial Release Conditions. The second way that private information may add inconsistency and noise is through the judges' use of financial and non-financial release conditions that require private information that is not included in the algorithm. The judges in our setting can, for example, assign monetary bail to encourage defendants to appear in court and not engage in new criminal activity, as the assigned bail amount is forfeited and an arrest warrant is issued if a defendant engages in any form of pretrial misconduct. Monetary bail therefore potentially gives judges additional incentives to reduce misconduct beyond what is used by the algorithm, which never recommends assigning monetary bail (although recent work finds no evidence that monetary bail incentivizes individuals to refrain from pretrial misconduct (Ouss and Stevenson, Forthcoming)). In our setting, the algorithm and the corresponding risk assessment report do not provide information on a defendant's financial resources, meaning that the judges must use private information to accurately predict a defendant's ability to pay the assigned bail amount. The judges in our setting can also assign a range of non-financial release conditions such as requiring the defendant to undergo treatment for substance abuse disorders, counseling for mental health issues, no contact orders with victims, and supervision by pretrial services. These non-financial release conditions again require that the judges use private information to accurately assess and treat a defendant's needs.

Figure 8 explores whether high- and low-skill judges differ in their use of these financial and non-financial release conditions that require private information by regressing each release condition on an indicator for being a high-skill judge, controlling for detailed risk score fixed effects. We report the coefficient and standard errors from the high-skill indicator and plot 95% confidence intervals from standard errors clustered at the judge level. We find that low-skill judges are 9.9 percentage points (SE: 5.0) more likely to assign monetary bail than the high-skill judges conditional on the risk score. The low-skill judges also

[14] In Appendix Table A.7 we more formally decompose judge performance into predictable and non-predictable differences based on these predicted release decisions. We find that the low-skill judges increase conditional misconduct rates by 3.5 percentage points (SE: 0.5) relative to the algorithm on average, with 0.7 percentage points (SE: 0.6) explained by the predictable use of observable information, increasing to 1.0 percentage point (SE = 0.6) when we add private information, suggesting that the predictable use of additional private information is modestly worsening the performance of low-skill judges. The high-skill judges decrease conditional misconduct rates by 1.5 percentage points (SE: 0.7) relative to the algorithm on average, with 0.2 percentage points (SE: 0.6) explained by the predictable use of observable information and 0.1 percentage points (SE: 0.6) explained by the predictable use of observable and private information.

perform particularly poorly when they assign monetary bail (Appendix Figure A.21), suggesting that these judges may mistakenly release some high-risk defendants and mistakenly detain some low-risk ones when setting monetary bail. The high-skill judges are instead much more likely to use non-financial conditions of release that are meant to directly address the defendants' underlying needs. The high-skill judges are 9.3 percentage points (SE: 3.2) more likely to impose drug and alcohol conditions compared to the low-skill judges, 1.3 percentage points (SE: 1.1) more likely to impose mental health conditions, 7.2 percentage points (SE: 2.5) more likely to impose no contact conditions, and 12.8 percentage points (SE: 5.8) more likely to impose pretrial supervision. Taken together, these findings suggest that the low-skill judges may be unnecessarily adding inconsistency and noise to their release decisions by overusing financial release conditions and underusing non-financial release conditions that address the defendants' needs.

An Example of Unhelpful Private Information. The third way that private information may add inconsistency and noise to release decisions is through the judges' reactions to uninformative details or events. We focus on the judges' reactions to a highly salient but largely uninformative event to better understand one particular way that such unhelpful private information can lead to inconsistency and noise, specifically hearings held just after a case where a different and completely unrelated defendant is arrested for a homicide or violent first-degree felony while on pretrial release. These serious offenses consist primarily of homicides, rapes and sexual assaults, aggravated assaults, and kidnappings. These are highly salient adverse events for a bail judge but are arguably uninformative on misconduct risk given both the rarity of such events and the fact that the judge assigned to the case is generally not the judge who initially released the defendant.

We estimate a judge's reaction to this adverse event using the following event-study specification:

$$D_{i,j,t} = \sum_{k \neq -1} \beta_k \mathbf{1}\{K_{j,t} = k\} + \mathbf{U}_i'\boldsymbol{\omega} + \mathbf{W}_i'\boldsymbol{\gamma} + \alpha_j + \varepsilon_{i,j,t}, \tag{9}$$

where $D_{i,j,t}$ is an indicator variable for pretrial release in an unrelated case i assigned to judge j in shift t and $K_{j,t}$ is an indicator denoting the number of shifts since the adverse event, ranging from $k = -5$ to $k = 5$. We assign cases in the same shift as the adverse event to the omitted shift and focus on shifts $k = -4$ to $k = 4$, binning cases outside the focal shifts in $k = -5$ and $k = 5$. $\mathbf{U}_i$ is a vector of observable case and defendant characteristics (including $\mathbf{X}_i$ and a subset of $\mathbf{V}_{i,j}$), $\mathbf{W}_i$ is a vector of shift-by-time effects, and α_j are judge fixed effects. The coefficients of interest are β_k, which measure the probability of release for cases heard in the four shifts before and four shifts after the adverse event relative to the omitted shift at $k = -1$. We estimate the event-study specification for a balanced panel of 60 judges, including 9 judges who never experienced an adverse event (who are assigned to the omitted shift) and 51 judges who we observe for at least four shifts before and after the first time they experience an adverse event. We focus our main results on the first observed adverse event and include all bail hearings during our sample period (including bond modification hearings and hearings for defendants brought in for new bench warrants) so that we can observe the entire sequence of each judge's caseload. Standard errors are clustered at the judge level.[15]

[15] Our event-study specification relies on two identifying assumptions: (1) that there are no anticipation effects and (2) that the outcomes of all adoption groups would have evolved in parallel in the absence of the adverse event, including for the early-treated, late-treated, and never-treated groups. We show below that there are no anticipatory trends before the adverse event, consistent with

Figure 9 plots our event-study estimates and corresponding 95% confidence intervals. We also report average treatment effects, pooling the first four post-treatment shifts. We start with the full sample of cases in Panel A, where we observe a sharp decline in pretrial release rates immediately after the adverse event. The response is largest in the second shift after the event, with release rates returning to baseline levels by the fourth shift after the event. The magnitude of the response is substantial, with a 5.4 percentage point (SE: 1.7) decrease in pretrial release rates over the first four shifts following the adverse event. Panel B shows that these effects are driven by the low-skill judges, whose release rates decrease by 5.6 percentage points (SE: 2.2) following the adverse event compared to a statistically insignificant increase of 0.2 percentage points (SE: 2.2) for high-skill judges. These results are consistent with low-skill judges reacting more to the salient adverse event.

We also explore heterogeneity in the judges' reactions to better understand this behavioral response. Panel A of Appendix Figure A.24 presents results separately by defendant race, as non-white defendants are particularly representative of those arrested for serious violent crimes while on pretrial release (Kahneman and Tversky, 1972; Bordalo et al., 2016) and prior work documents substantial racial discrimination in bail decisions (Arnold, Dobbie, and Yang, 2018; Arnold, Dobbie, and Hull, 2021). We find that non-white defendants are 10.0 percentage points (SE: 2.3) less likely to be released after an adverse event compared to a statistically insignificant decrease of 0.7 percentage points (SE: 2.4) for white defendants. Panel B of Appendix Figure A.24 similarly shows that defendants currently on probation or parole are 9.4 percentage points (SE: 3.8) less likely to be released after an adverse event compared to only 3.6 percentage points (SE: 1.2) for defendants not on probation or parole. Recall that the judges substantially overweight this characteristic relative to the risk score (Table 3). Taken together, these results suggest that the reactions are concentrated among defendants who are either particularly representative of those arrested for serious violent crimes (Kahneman and Tversky, 1972; Bordalo et al., 2016) or who have observable characteristics that are particularly overweighted by judges (Bordalo, Gennaioli, and Shleifer, 2015; Sunstein, 2022).

We view these patterns as overreactions to a highly salient but largely uninformative event rather than to a rational updating of beliefs. Appendix Figure A.25 shows that the judges' reactions are concentrated among observably low-risk defendants who the algorithm recommends releasing such that judges increase their probability of harsh overrides by 5.2 percentage points (SE: 1.8) after the adverse event. In addition, Appendix Figure A.26 shows that conditional misconduct rates decrease by only a statistically insignificant 0.1 percentage points (SE: 2.5) following an adverse event despite the large decrease in release rates documented above. Both results, as well as the fact that judges' release rates return to baseline levels relatively soon after the adverse event, are inconsistent with most models of rational updating but consistent with a behavioral response. For example, the judges in our setting may be particularly prone to the "availability" heuristic described in Tversky and Kahneman (1974), where they overreact to recent and salient events that are "top of mind" (Bordalo, Gennaioli, and Shleifer, 2013; Bordalo et al., 2016; Sunstein, 2022).[16]

the first identifying assumption. We also show that our results are robust to restricting the control group to the never-treated judges (Appendix Figure A.22) and that our effects are not driven by changes in the types of cases assigned to judges before and after the adverse event (Appendix Figure A.23).

[16]Similar overreactions impact macroeconomic expectations (Bordalo et al., 2020), medical treatment decisions (Choudhry et al., 2006; Singh, 2021), coverage of crime and judicial errors (Philippe and Ouss, 2018), and sentencing reversals on appeal (Bhuller and Sigstad, 2021). As one such example, Singh (2021) finds that experiencing adverse obstetric events in one delivery mode (such

Our findings are also consistent with assimilation effects whereby the judges, in particular the low-skill judges, include information about the adverse salient event in their evaluation of subsequent cases (Bless and Schwarz, 1998, 2010). In this case, the salient adverse event may cause the judges to view future defendants as more risky than in the absence of the adverse news, resulting in harsher pretrial decisions.

V.C Survey Evidence on Judge Preferences

We conclude by using new survey data collected for this study to better understand the types of information the high- and low-skill judges consider when making release decisions, providing new insights into the kind of private information that adds valuable signal versus noise. The survey asked judges to rank the importance of different defendant and case characteristics when making the decision to impose monetary bail using the following prompt:

> *We are interested in learning what factors you use to make pretrial decisions. Below is a list of 19 different factors, including factors commonly considered by judges across the United States. For each factor, please tell us on a scale of 1–5, where 1 is not important and 5 is very important, how important each factor is for you in making a pretrial decision. Specifically, please tell us how important each factor is in the decision of whether to impose financial conditions of release (i.e., cash bail or bond).*

We focused on the observable factors that are included in the algorithm (such as charge type and prior criminal history), private demographic information that is not included in the algorithm (such as race and gender), and private non-demographic information that is not included in the algorithm (such as mental health condition and history, substance abuse diagnosis and history, and financial resources). We surveyed 28 of the judges in our sample, with similar response rates for the high- and low-skill judges. (The coefficient in a bivariate regression of indicator for high-skill on an indicator for survey response is 0.16 (SE: 0.12) with an $R^2 = 0.03$.) See Appendix B for additional details on the survey administration, response rates, and a full list of questions.

Figure 10 presents the bivariate relationship between being a high-skilled judge and three mutually exclusive indices for (1) observable information available to both the judges and the algorithm, (2) demographic private information available to only the judges, and (3) non-demographic private information available to only the judges, as well as each individual factor within the indices. The individual factors are indicator variables for placing greater than median weight on the factor, while the indices are simple group means of these indicator variables. Each point in Figure 10 represents estimates from OLS regressions of the relevant index variable or individual factor on an indicator for being a high-skill judge. We report the coefficient and standard errors on the high-skill judge indicator and plot 95% confidence intervals from robust standard errors throughout.

The survey data reveal several striking patterns. The high- and low-skill judges report using the observable information that is available to both the judges and the algorithm in a remarkably similar way, consistent with what we observe in the administrative court data in Table 3. High-skill judges report putting

as C-section versus vaginal delivery) makes the physician more likely to switch to the other delivery mode on the next patient, resulting in worse patient outcomes and inefficient resource use.

more weight on an index of observable information by a statistically insignificant 3.4 percentage points (SE: 11.8). We see a similar pattern for each component of the index. For example, high- and low-skill judges report placing statistically indistinguishable weight on observable factors, such as whether the current offense is a violent offense and whether the defendant has a past history of pretrial misconduct.

But there are important differences in how the high- and low-skill judges report using the private information that is not available to the algorithm, consistent with what we observe in the administrative court data in Table 3 and Figure 7. In the survey data, high-skill judges are 19.0 percentage points (SE = 9.8) less likely to report putting weight on private demographic information such as gender and race. These results are driven by differences in reported weight on defendant race, matching what we found in the administrative data in Table 3. In contrast, the high-skill judges are 16.4 percentage points (SE = 8.1) more likely to put weight on an index of private non-demographic information, driven by greater weight on factors such as mental health condition and history, substance abuse diagnosis and history, and financial resources. The greater weight that high-skill judges place on these factors is particularly illuminating in light of our findings in Appendix Figure A.21 that low-skill judges' underperformance could be driven by their ineffective use of monetary bail while high-skill judges' outperformance could be driven by their greater use of non-financial conditions such as treatment for substance abuse disorders.

Both these findings and the administrative court data suggest that the consideration of relevant private information can help explain why the high-skill judges can outperform the algorithm. The survey data suggest that we may be able to improve human performance by teaching the judges to focus only on the most relevant factors not included in the algorithm when deciding whether to override the algorithm. For example, the survey data suggest that instructing the judges to not consider defendant race and to not set monetary bail without considering the defendant's financial resources may help improve their performance when making an override decision. Our findings also suggest that judges may benefit from regular feedback on the private characteristics of formerly released defendants who engaged in misconduct to facilitate learning of relevant private information.

VI Conclusion

This paper shows that there is substantial variation in the effects of human discretion on the accuracy of decisions in the context of bail decisions. We estimate that 90% of the judges in our setting underperform the algorithm on average when they make discretionary overrides, with most making decisions that are no better than random. However, the remaining 10% outperform the algorithm and significantly decrease pretrial misconduct compared to an algorithmic counterfactual. We find that our results are unlikely to be explained by the use of observable information that is available to both the judges and the algorithm or by the judges having different objective functions. Rather, these performance differences are most likely driven by how the judges use the private information that is unavailable to the algorithm, with high-skill judges using such information to improve the accuracy of their decisions and low-skill judges only adding inconsistency and noise when they attempt to use such information.

Our findings suggest that there will not necessarily be a single correct policy on human oversight of algorithms if humans still make the final decision in a setting, as the impact of such policies depends on the

ability of the human decision-makers in a particular context. Our findings also suggest that we can increase the accuracy of decisions by allowing predictive algorithms to include the relevant private information used by high-skilled decision-makers. The quasi-experimental methods developed in this paper may also prove useful in measuring the impact of human discretion on the accuracy of decisions in other high-stakes settings. Our approach is appropriate whenever there is the quasi-random assignment of decision-makers and the objective of these decision-makers is both known and well measured. Our test can therefore be used to explore the impact of human discretion on the accuracy of decisions in other settings where algorithms are widely used, such as hiring, lending, and medical testing decisions.

References

Abaluck, Jason, Leila Agha, David C. Chan, Daniel Singer, and Diana Zhu. 2021. "Fixing Misallocation with Guidelines: Awareness vs. Adherence." *NBER Working Paper No. 27467*.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." *NBER Working Paper No. 31422*.

Albright, Alex. 2023. "The Hidden Effects of Algorithmic Recommendations." *Unpublished Working Paper*.

Andrews, Donald, and Marcia Schafgans. 1998. "Semiparametric Estimation of the Intercept of a Sample Selection Model." *Review of Economic Studies*, 65(3): 497–517.

Anwar, Shamena, Shawn D. Bushway, and John Engberg. 2023. "The Impact of Defense Counsel at Bail Hearings." *Science Advances*, 9(18).

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics*, 133(4): 1885–1932.

Arnold, David, Will Dobbie, and Peter Hull. 2021. "Measuring Racial Discrimination in Algorithms." *AEA Papers and Proceedings*, 111: 49–54.

Arnold, David, Will Dobbie, and Peter Hull. 2022. "Measuring Racial Discrimination in Bail Decisions." *American Economic Review*, 112(9): 2992–3038.

Berk, Richard. 2017. "An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism." *Journal of Experimental Criminology*, 13(2): 193–216.

Bhuller, Manudeep, and Henrik Sigstad. 2021. "Feedback and Learning: The Causal Effects of Reversals on Judicial Decision-Making." *Unpublished Working Paper*.

Bless, Herbert, and Norbert Schwarz. 1998. "Context Effects in Political Judgement: Assimilation and Contrast as a Function of Categorization Processes." *European Journal of Social Psychology*, 28(2): 159–172.

Bless, Herbert, and Norbert Schwarz. 2010. "Mental Construal and the Emergence of Assimilation and Contrast Effects: The Inclusion/Exclusion Model." In *Advances in Experimental Social Psychology*. Vol. 42, 319–373. Elsevier.

Bonhomme, Stéphane, and Martin Weidner. 2022. "Posterior Average Effects." *Journal of Business & Economic Statistics*, 40(4): 1849–1862.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics*, 131(4): 1753–1794.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice Under Risk." *Quarterly Journal of Economics*, 127(3): 1243–1285.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013. "Salience and Consumer Choice." *Journal of Political Economy*, 121(5): 803–843.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2015. "Salience Theory of Judicial Decisions." *Journal of Legal Studies*, 44(S1): S7–S33.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. 2020. "Overreaction in Macroeconomic Expectations." *American Economic Review*, 110(9): 2748–2782.

Brinch, Christian, Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy*, 125(4): 985–1039.

Chamberlain, Gary. 1986. "Asymptotic Efficiency in Semi-Parametric Models with Censoring." *Journal of Econometrics*, 32(2): 189–218.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *Quarterly Journal of Economics*, 137(2): 729–783.

Choudhry, Niteesh K., Geoffrey M. Anderson, Andreas Laupacis, Dennis Ross-Degnan, Sharon-Lise T. Normand, and Stephen B. Soumerai. 2006. "Impact of Adverse Events on Prescribing Warfarin in Patients with Atrial Fibrillation: Matched Pair Analysis." *BMJ: British Medical Journal*, 332(7534): 141–145.

Currie, Janet M., and W. Bentley MacLeod. 2020. "Understanding Doctor Decision Making: The Case of Depression Treatment." *Econometrica*, 88(3): 847–878.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review*, 108(2): 201–240.

Dube, Oeindrila, Sandy Jo MacArthur, and Anuj Shah. 2022. "A Cognitive View of Policing." *Unpublished Working Paper*.

Efron, Bradley. 2016. "Empirical Bayes Deconvolution Estimates." *Biometrika*, 103(1): 1–20.

Eren, Ozkan, and Naci Mocan. 2018. "Emotional Judges and Unlucky Juveniles." *American Economic Journal: Applied Economics*, 10(3): 171–205.

Finkelstein, Amy, Petra Persson, Maria Polyakova, and Jesse Shapiro. 2022. "A Taste of Their Own Medicine: Guideline Adherence and Access to Expertise." *American Economic Review: Insights*, 4(4): 507–526.

Frandsen, Brigham R., Lars J. Lefgren, and Emily C. Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review*, 113(1): 253–277.

Heckman, James J. 1990. "Varieties of Selection Bias." *American Economic Review*, 80(2): 313–318.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. 2018. "Discretion in Hiring." *Quarterly Journal of Economics*, 133(2): 765–800.

Hull, Peter. 2020. "Estimating Hospital Quality with Quasi-Experimental Data." *Unpublished Working Paper*.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2020. "Simple Rules to Guide Expert Classifications." *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(3): 771–800.

Kahneman, Daniel, and Amos Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology*, 3(3): 430–454.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment.* Little, Brown Spark and Company.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Ludwig, Jens, and Sendhil Mullainathan. 2023. "Machine Learning as a Tool for Hypothesis Generation." NBER Working Paper No. 31017.

Mellers, Barbara, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S. Emlen Metz, Lyle Ungar, Michael M. Bishop, Michael Horowitz, Ed Merkle, and Philip Tetlock. 2015. "The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics." *Journal of Experimental Psychology: Applied*, 21(1): 1–14.

Miller, Ted R., Mark A. Cohen, David I. Swedler, Bina Ali, and Delia V. Hendrie. 2021. "Incidence and Costs of Personal and Property Crimes in the USA, 2017." *Journal of Benefit-Cost Analysis*, 12(1): 24–54.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters." *Econometrica*, 86(5): 1589–1619.

Mullainathan, Sendhil. 2002. "A Memory-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 117(3): 735–774.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *Quarterly Journal of Economics*, 137(2): 679–727.

Narasimhan, Balasubramanian, and Bradley Efron. 2020. "deconvolveR: A G-Modeling Program for Deconvolution and Empirical Bays Estimation." *Journal of Statistical Software*, 94(11): 1–20.

Ouss, Aurélie, and Megan Stevenson. Forthcoming. "Does Cash Bail Deter Misconduct?" *American Economic Journal: Applied Economics*.

Philippe, Arnaud, and Aurélie Ouss. 2018. "No Hatred or Malice, Fear or Affection: Media and Sentencing." *Journal of Political Economy*, 126(5): 2134–2178.

Rambachan, Ashesh. 2022. "Identifying Prediction Mistakes in Observational Data." *Unpublished Working Paper*.

Satopää, Ville A., Marat Salikhov, Philip E. Tetlock, and Barbara Mellers. 2021. "Bias, Information, Noise: The BIN Model of Forecasting." *Management Science*, 67(12): 7599–7618.

Singh, Manasvini. 2021. "Heuristics in the Delivery Room." *Science*, 374(6565): 324–329.

Stevenson, Megan. 2018. "Assessing Risk Assessment in Action." *Minnesota Law Review*, 103: 303–384.

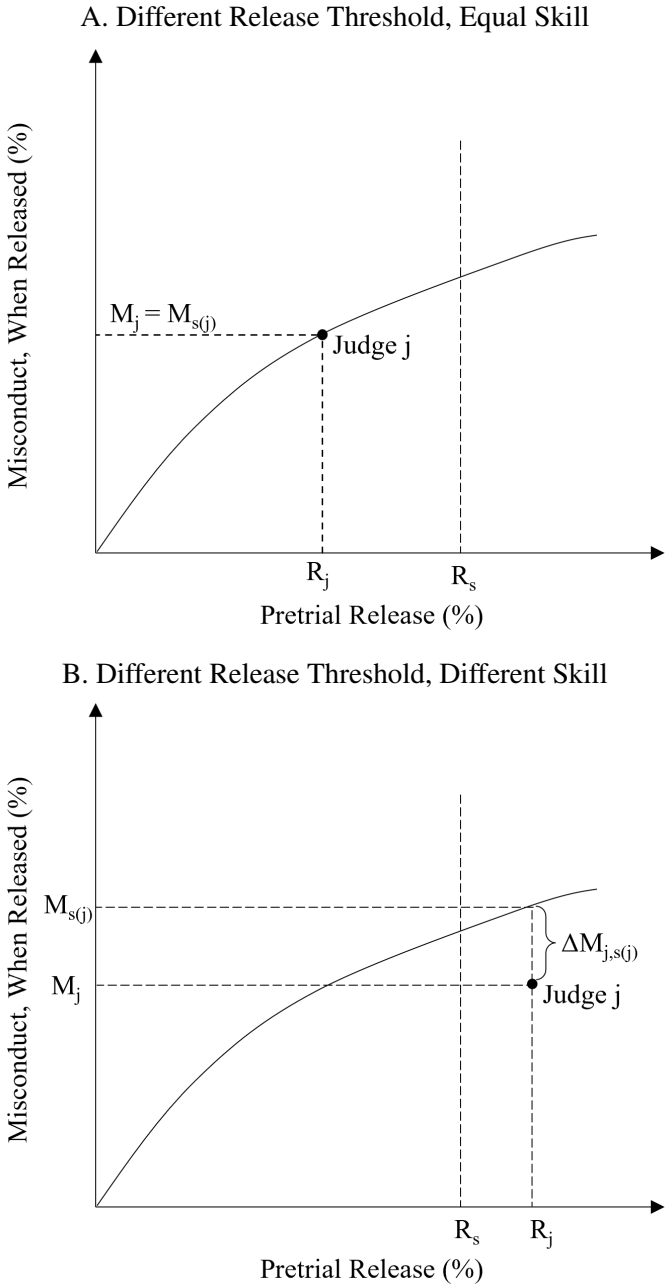
Stevenson, Megan T., and Jennifer L. Doleac. 2022. "Algorithmic Risk Assessment in the Hands of Humans." *Unpublished Working Paper*.

Sunstein, Cass R. 2022. "Governing by Algorithm? No Noise and (Potentially) Less Bias." *Duke Law Journal*, 71(6): 1175–1205.

Tetlock, Philip E., and Dan Gardner. 2015. *Superforecasting: The Art and Science of Prediction.* Random House.

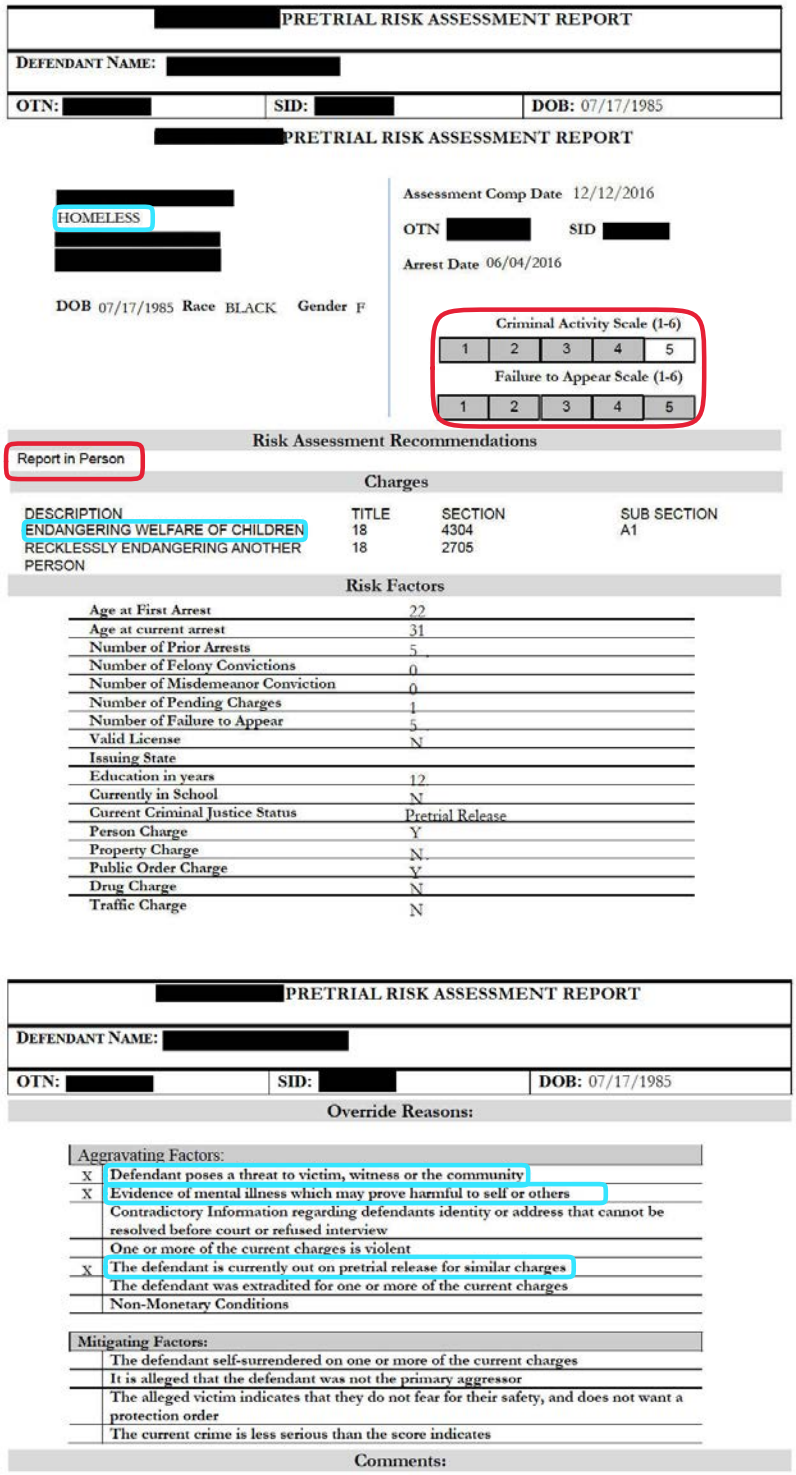
Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131.

Figure 1: Hypothetical Variation in Release Thresholds and Predictive Skill



Notes. This figure plots release rates against misconduct rates among released defendants for a hypothetical judge, along with counterfactual misconduct rates among released defendants for a hypothetical algorithm. Panel A varies the release threshold and fixes predictive skill compared to the algorithmic counterfactual. Panel B varies both the release threshold and predictive skill compared to the algorithmic counterfactual.

Figure 2: Example Pretrial Risk Assessment Report



[redacted] PRETRIAL RISK ASSESSMENT REPORT

DEFENDANT NAME: [redacted]

OTN: [redacted] SID: [redacted] DOB: 07/17/1985

[redacted] PRETRIAL RISK ASSESSMENT REPORT

[redacted]
HOMELESS
[redacted]
[redacted]

DOB 07/17/1985 Race BLACK Gender F

Assessment Comp Date 12/12/2016

OTN [redacted] SID [redacted]

Arrest Date 06/04/2016

Criminal Activity Scale (1-6)

1	2	3	4	5

Failure to Appear Scale (1-6)

1	2	3	4	5

Risk Assessment Recommendations

Report in Person

Charges

DESCRIPTION	TITLE	SECTION	SUB SECTION
ENDANGERING WELFARE OF CHILDREN	18	4304	A1
RECKLESSLY ENDANGERING ANOTHER PERSON	18	2705	

Risk Factors

Age at First Arrest	22
Age at current arrest	31
Number of Prior Arrests	5
Number of Felony Convictions	0
Number of Misdemeanor Conviction	0
Number of Pending Charges	1
Number of Failure to Appear	5
Valid License	N
Issuing State	
Education in years	12
Currently in School	N
Current Criminal Justice Status	Pretrial Release
Person Charge	Y
Property Charge	N
Public Order Charge	Y
Drug Charge	N
Traffic Charge	N

[redacted] PRETRIAL RISK ASSESSMENT REPORT

DEFENDANT NAME: [redacted]

OTN: [redacted] SID: [redacted] DOB: 07/17/1985

Override Reasons:

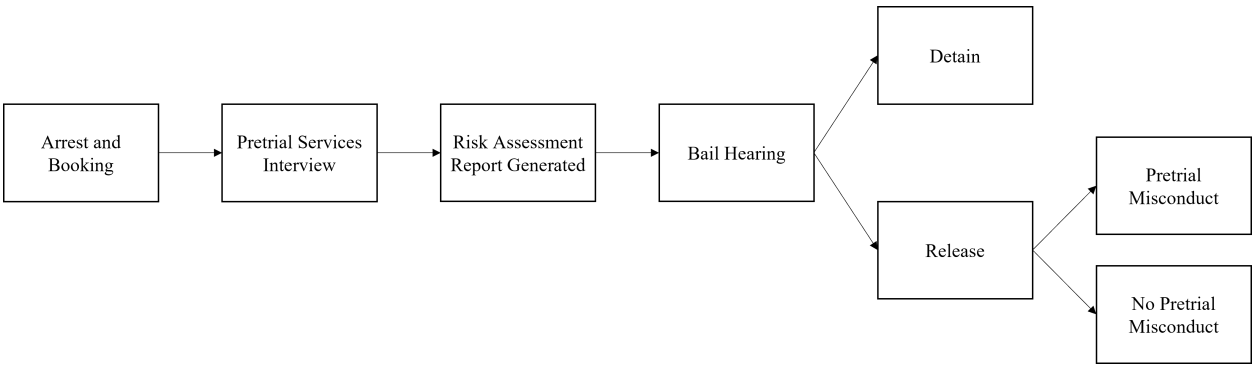
Aggravating Factors:	
X	Defendant poses a threat to victim, witness or the community
X	Evidence of mental illness which may prove harmful to self or others
	Contradictory Information regarding defendants identity or address that cannot be resolved before court or refused interview
	One or more of the current charges is violent
X	The defendant is currently out on pretrial release for similar charges
	The defendant was extradited for one or more of the current charges
	Non-Monetary Conditions

Mitigating Factors:	
	The defendant self-surrendered on one or more of the current charges
	It is alleged that the defendant was not the primary aggressor
	The alleged victim indicates that they do not fear for their safety, and does not want a protection order
	The current crime is less serious than the score indicates

Comments:

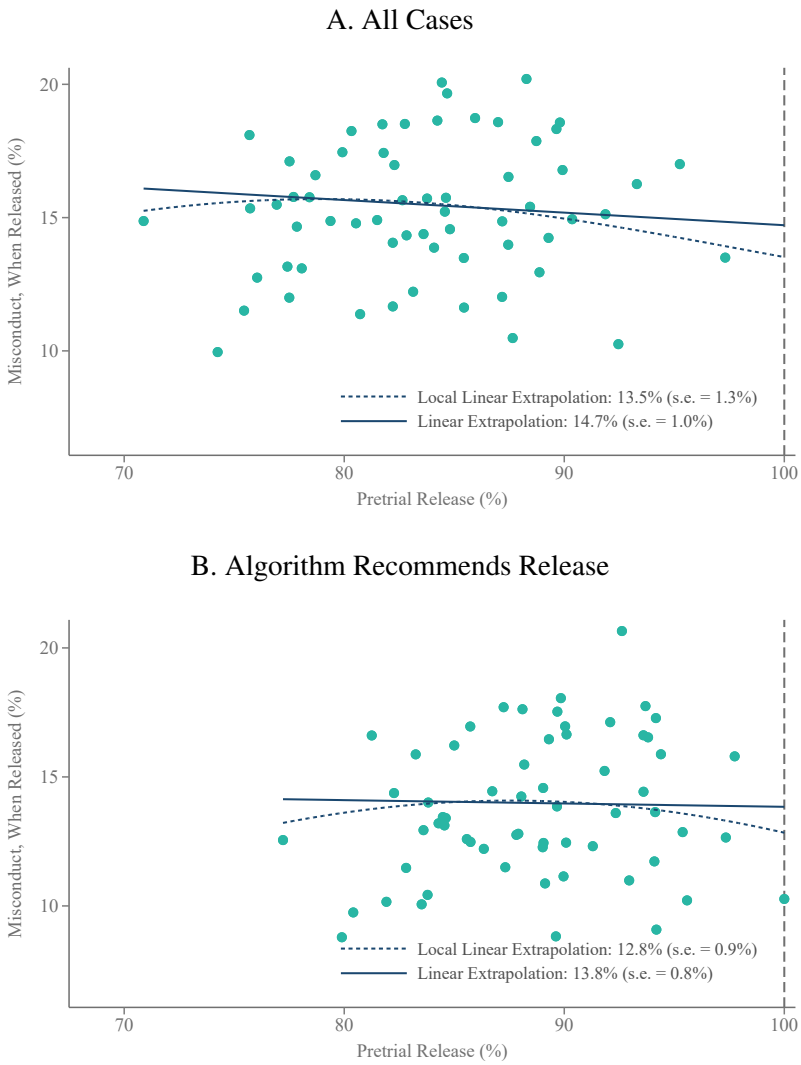
Notes. This figure shows an example risk assessment report in our setting. Red ovals indicate the risk assessment scores and algorithmic recommendation. Blue ovals indicate examples of private information not included in the algorithm. See the text for additional details.

Figure 3: Pretrial Process



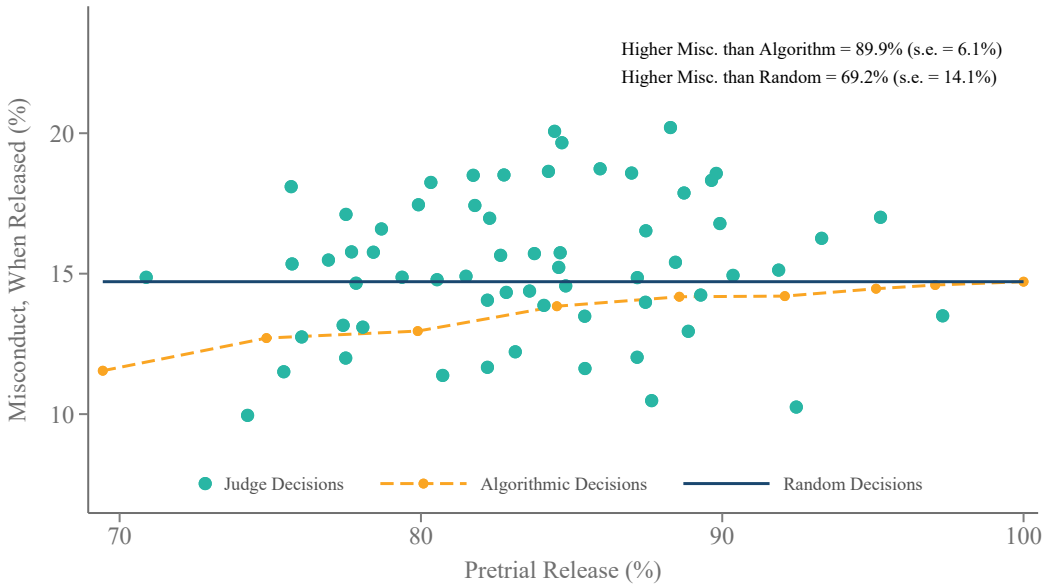
Notes. This figure shows the pretrial process starting from the arrest and booking of the defendant to his potential post-release outcomes. See the text for additional details.

Figure 4: Extrapolations of Release and Conditional Misconduct Rates



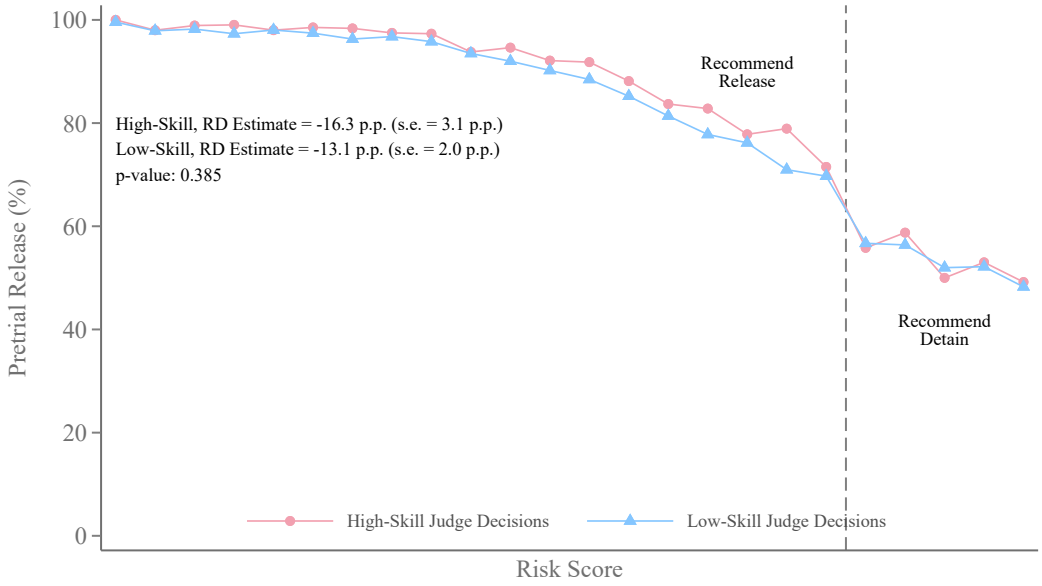
Notes. This figure plots judge release rates against misconduct rates among released defendants at two algorithmic risk score cutoffs. Each point represents the mean release and conditional misconduct rate for each judge, adjusted for shift-by-time fixed effects. Panel A reports results for the full sample of cases, corresponding to an algorithmic release rate of 100%. Panel B restricts the sample to cases where the algorithm recommends release, corresponding to an algorithmic release rate of 84.5%. Each panel also plots local linear and linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated predicted misconduct rate among released defendants. The local linear regression uses a Gaussian kernel with a fixed bandwidth. We report the estimated intercept and standard error at a cutoff-specific release rate of 100% under each extrapolation, which equals the estimated average misconduct risk for the relevant sample of defendants. Standard errors are estimated by taking random draws from the distributions of the estimated judge-specific release rates and conditional misconduct rates and then recalculating the threshold-specific extrapolations.

Figure 5: Conditional Misconduct Rates Relative to the Algorithm



Notes. This figure plots release rates against misconduct rates among released defendants for the 62 judges in our sample, along with counterfactual misconduct rates among released defendants for algorithmic and random decisions. Conditional misconduct rates under the algorithm are estimated using linear extrapolations of mean risk at different risk score cutoffs as illustrated in Figure 4. Conditional misconduct rates under the random release rule are estimated using linear extrapolations of mean risk for the full sample as described in detail in the main text. All estimates adjust for shift-by-time fixed effects. The figure also reports the fraction of judges with higher misconduct rates compared to the algorithmic and the random release rules, computed from these estimates as posterior average effects. Standard errors are estimated by taking random draws from the distributions of the estimated judge-specific release rates and conditional misconduct rates and then recalculating the statistics of interest.

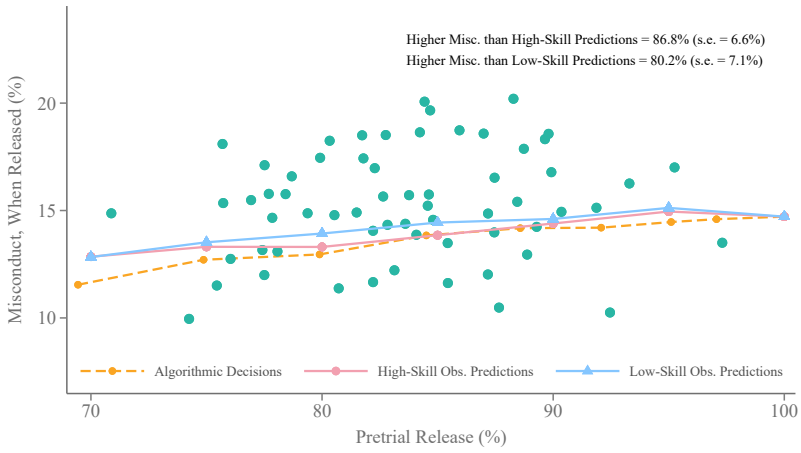
Figure 6: Pretrial Release Rates by Judge Skill



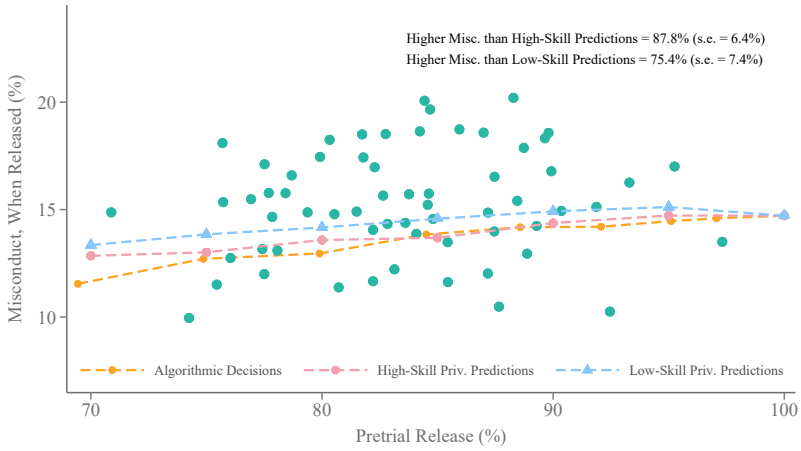
Notes. This figure plots pretrial release rates against algorithmic NCA risk scores by judge skill. The dashed vertical line indicates the score where the algorithmic recommendation changes from release to detain. We also report regression discontinuity estimates and standard errors using a local randomization approach, and the p-value from a test of equality.

Figure 7: Predicted Pretrial Release Decisions

A. Predictions Based on Observable Information

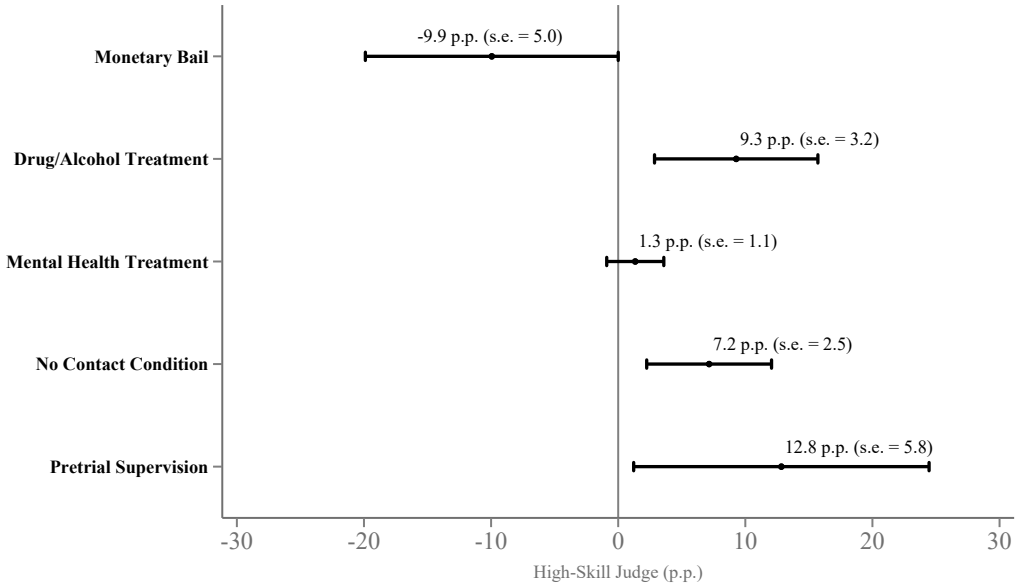


B. Predictions Based on Observable and Private Information



Notes. This figure plots release rates against misconduct rates among released defendants, along with counterfactual misconduct rates from the high- and low-skill judges' predicted release decisions. Panel A constructs the predicted release decisions using the observable characteristics in the original algorithm. Panel B constructs the predicted release decisions using the observable characteristics in the original algorithm and the private information listed in Table 1. Conditional misconduct rates under the predicted release decisions are estimated using linear extrapolations at different release cutoffs as described in the main text. All estimates adjust for shift-by-time fixed effects. The figure also reports the fraction of judges with higher misconduct rates compared to the predicted release decisions, computed from these estimates as posterior average effects. Standard errors are estimated by taking random draws from the distributions of the estimated judge-specific release rates and conditional misconduct rates and then recalculating the statistics of interest.

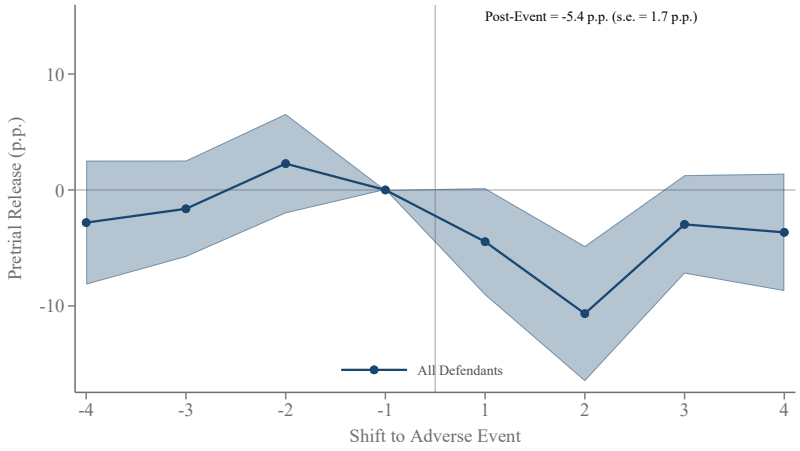
Figure 8: Financial and Non-Financial Release Conditions of High-Skill Judges



Notes. This figure reports OLS estimates of regressions of indicators for financial and non-financial release conditions on an indicator for being a high-skill judge. Each point represents the coefficient from a bivariate regression of the indicated variable on an indicator for being a high-skill judge. We also plot 95% confidence intervals from robust standard errors clustered by judge.

Figure 9: Effect of Adverse Events on Pretrial Release

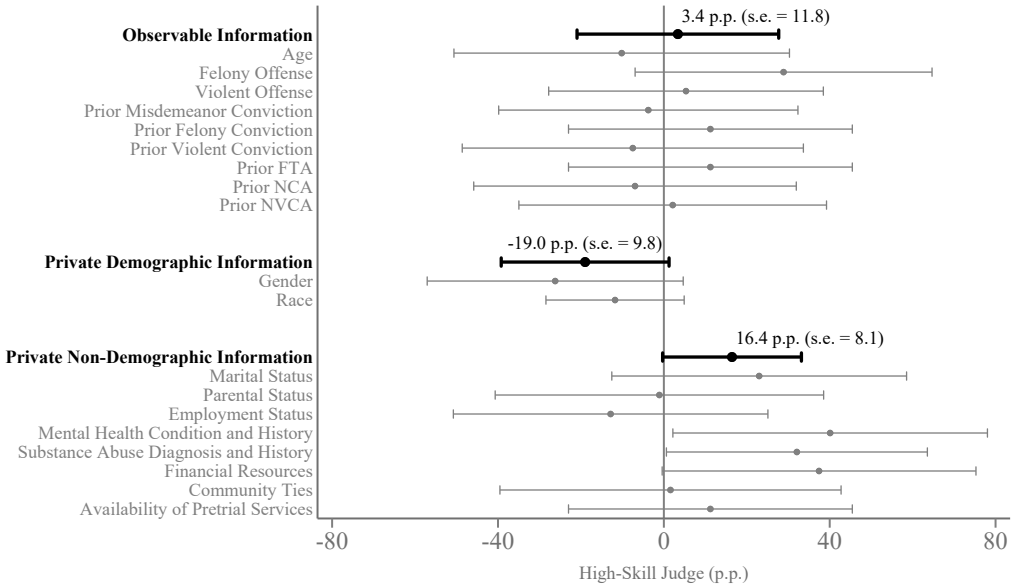
A. Full Sample



B. By Judge Skill

Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release. Panel A reports results for the full sample, and Panel B reports results separately for high-skilled and low-skilled judges. The horizontal axis denotes time, in shifts, relative to the adverse event. The estimated effect is normalized to zero in the shift before the adverse event. The shaded regions are 95% confidence intervals from robust standard errors clustered by judge. We also report the average effect and standard error across the four post-event shifts and, when relevant, the p-value from a test of equality. See the text for additional details on the sample and regression specification.

Figure 10: Preferences and Beliefs of High-Skill Judges



Notes. This figure reports OLS estimates of regressions of judge preferences and beliefs on an indicator for being a high-skilled judge. Information on the judges' preferences and beliefs comes from a survey that asked judges to rank the importance of different defendant and case characteristics when making the decision to impose monetary bail. Each point represents the coefficient from a bivariate regression of the indicated variable on an indicator for being a high-skilled judge. The bolded rows are index variables constructed using the mean values of each of the individual variables, where each individual variable is an indicator for reporting an above median weight on the relevant case or defendant characteristic. We also plot 95% confidence intervals from robust standard errors. Each regression only includes the 28 judges who responded to the survey.

Table 1: Descriptive Statistics

	All Cases	Recommend Detain: Lenient Override	Recommend Detain: Follow Algorithm	Recommend Release: Harsh Override	Recommend Release: Follow Algorithm
A. Pretrial Release	(1)	(2)	(3)	(4)	(5)
Released Before Trial	0.83	1.00	0.00	0.00	1.00
Share ROR	0.10	0.02	—	—	0.11
Share Non-Monetary	0.52	0.15	—	—	0.56
Share Monetary Bail	0.37	0.82	—	—	0.32
Share Remanded	0.01	0.01	—	—	0.00
B. Observable Information					
Age at Current Arrest	34.59	33.02	33.04	35.50	34.80
Age at First Arrest	21.63	16.91	16.59	20.06	22.85
Prior Arrests	9.39	17.86	19.89	12.86	6.96
Prior Felonies	1.39	3.04	3.52	2.10	0.91
Prior Misdemeanors	2.38	4.70	5.26	3.26	1.73
Pending Charges	0.57	1.94	2.07	0.52	0.27
Property Charge	0.20	0.26	0.29	0.26	0.18
Drug Charge	0.28	0.47	0.40	0.27	0.24
Public Order Charge	0.45	0.50	0.57	0.53	0.43
Traffic Charge	0.14	0.28	0.17	0.06	0.14
Parole/Probation	0.27	0.46	0.65	0.52	0.18
Pretrial Release	0.32	0.91	0.89	0.33	0.20
Person Charge	0.41	0.17	0.20	0.38	0.46
C. Private Information					
Male	0.74	0.84	0.87	0.84	0.71
White	0.44	0.38	0.37	0.39	0.46
Homeless	0.05	0.07	0.13	0.13	0.03
No Telephone	0.08	0.09	0.14	0.17	0.07
Out-of-State Address	0.03	0.01	0.01	0.05	0.03
Violent Charge Against an Adult	0.48	0.26	0.32	0.49	0.53
Violent Charge Against a Child	0.04	0.02	0.02	0.04	0.05
Any Aggravating Condition	0.11	0.00	0.00	0.14	0.12
Override Recommendation	0.08	0.05	0.01	0.31	0.06
D. Pretrial Misconduct, When Released					
Any Misconduct	0.16	0.29	—	—	0.14
Share NCA Only	0.74	0.62	—	—	0.76
Share FTA Only	0.17	0.21	—	—	0.16
Share NCA and FTA	0.09	0.17	—	—	0.08
Cases	37,855	3,142	2,721	3,784	28,208

Notes. This table reports descriptive statistics for our analysis sample. The sample consists of bail hearings assigned to judges between October 16, 2016 and March 16, 2020, as described in the text. Information on case and defendant characteristics and pretrial outcomes is derived from court records as described in the text. Pretrial release is defined as ever being released before case disposition. ROR (release on recognizance) is defined as being released without any conditions. FTA is defined as failing to appear at a required court appearance. NCA is defined as a rearrest before case disposition. An indicator for a person charge is not included in the NCA predictive algorithm but is included under Panel B for completeness. Column 1 reports statistics for the full sample of cases. Columns 2 and 3 restrict the sample to cases where the algorithm recommends detention. Columns 4 and 5 restrict the sample to cases where the algorithm recommends release.

Table 2: Characteristics of High-Skill Judges

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Override Rate (0-100)	-1.44									-1.25
	(1.67)									(1.73)
Above Median Experience		0.73								0.43
		(11.93)								(12.99)
Male			-9.89							-1.29
			(15.20)							(16.53)
White				11.58						9.79
				(19.23)						(21.99)
Registered Republican					21.13					24.91
					(14.60)					(15.31)
Law Degree						17.98				10.77
						(12.40)				(15.03)
Former Prosecutor							2.98			-7.05
							(20.54)			(24.73)
Former Police Officer								-24.62		-24.62
								(11.75)		(14.86)
White vs. Non-White Disparity (0-100)									-16.40	-16.20
									(6.76)	(6.93)
R^2	0.01	0.00	0.01	0.00	0.04	0.04	0.00	0.04	0.06	0.18
Judges	62	62	62	62	62	62	62	62	62	62

Notes. This table reports OLS estimates of regressions of an indicator for being a high-skill judge on judge characteristics. Information on the judge demographics is derived from publicly available voter data and official publications. Judge skill, override rates, and white vs. non-white release disparities are estimated using the administrative court data as described in the text. The white vs. non-white release disparities are empirical Bayes posteriors computed using a standard shrinkage procedure. Robust standard errors are reported in parentheses. See the text for additional details.

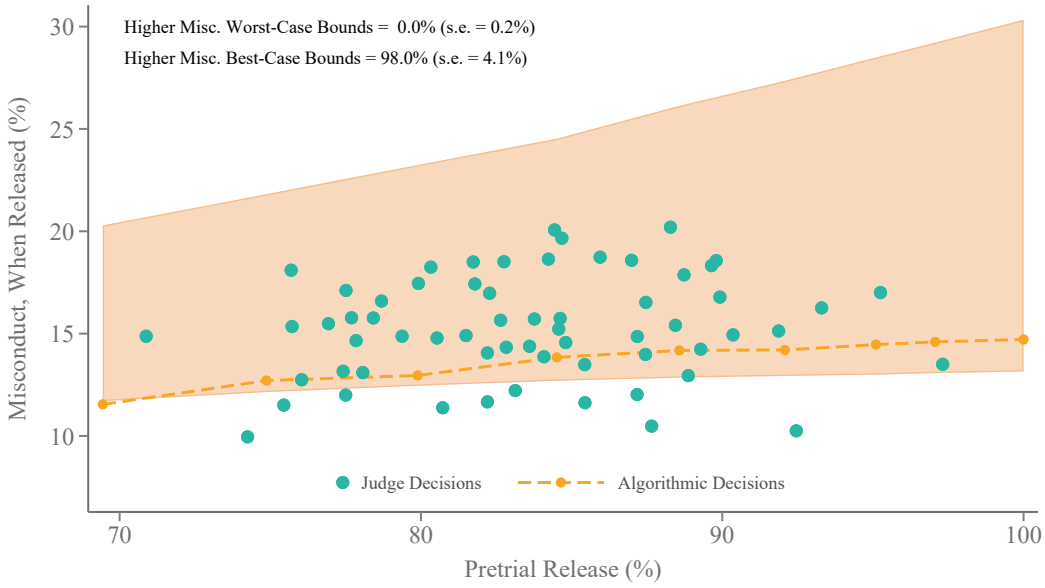
Table 3: Characteristics of Released Defendants

	High-Skill Judges	Low-Skill Judges	p-value
A. Observable Information	(1)	(2)	(3)
Age at Current Arrest	0.76	0.08	0.22
	(0.47)	(0.31)	
Age at First Arrest	-0.11	-0.09	0.81
	(0.06)	(0.05)	
Prior Arrests	-0.30	-0.16	0.30
	(0.11)	(0.08)	
Prior Felonies	0.34	-0.14	0.11
	(0.24)	(0.20)	
Prior Misdemeanors	0.25	0.09	0.60
	(0.27)	(0.16)	
Pending Charges	-0.31	-1.52	0.21
	(0.89)	(0.47)	
Property Charge	-1.92	-1.96	0.98
	(1.65)	(0.68)	
Drug Charge	2.58	2.48	0.94
	(1.16)	(0.65)	
Public Order Charge	-2.12	-3.85	0.08
	(0.87)	(0.42)	
Traffic Charge	7.21	8.60	0.23
	(1.00)	(0.76)	
Parole/Probation	-12.53	-14.53	0.34
	(1.84)	(0.94)	
Pretrial Release	-0.18	2.37	0.15
	(1.54)	(1.01)	
B. Private Information			
Male	-3.29	-2.46	0.29
	(0.69)	(0.37)	
White	0.01	1.35	0.14
	(0.69)	(0.59)	
Homeless	-11.79	-13.56	0.46
	(1.96)	(1.42)	
No Telephone	-6.40	-9.63	0.07
	(1.62)	(0.71)	
Out-of-State Address	-3.26	-9.10	0.02
	(1.97)	(1.61)	
Violent Charge Against an Adult	-4.43	-2.04	0.03
	(0.77)	(0.80)	
Violent Charge Against a Child	0.76	1.31	0.75
	(1.28)	(1.14)	
Any Aggravating Condition	-0.48	2.33	0.03
	(1.12)	(0.68)	
Override Recommendation	-21.07	-23.75	0.17
	(1.69)	(0.90)	
Risk Score FE	Yes	Yes	
Cases	7,909	29,946	

Notes. This table reports OLS estimates of regressions of an indicator for release on case and defendant characteristics with NCA risk score fixed effects. Column 1 reports results for high-skill judges with lower conditional misconduct rates than the algorithm holding fixed release rates, column 2 reports results for low-skill judges with higher conditional misconduct rates than the algorithm holding fixed release rates, and column 3 reports the p-value on the difference. Standard errors clustered by judge are reported in parentheses.

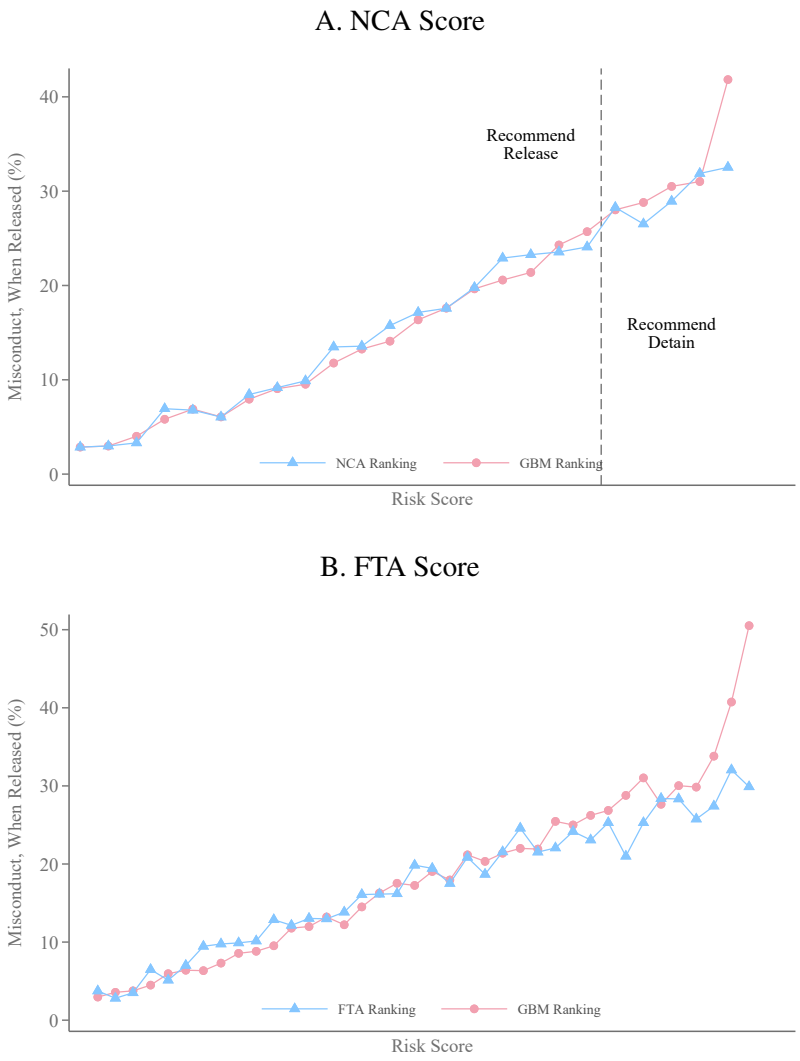
Appendix A: Additional Figures and Tables

Figure A.1: Uninformative Worst- and Best-Case Bounds



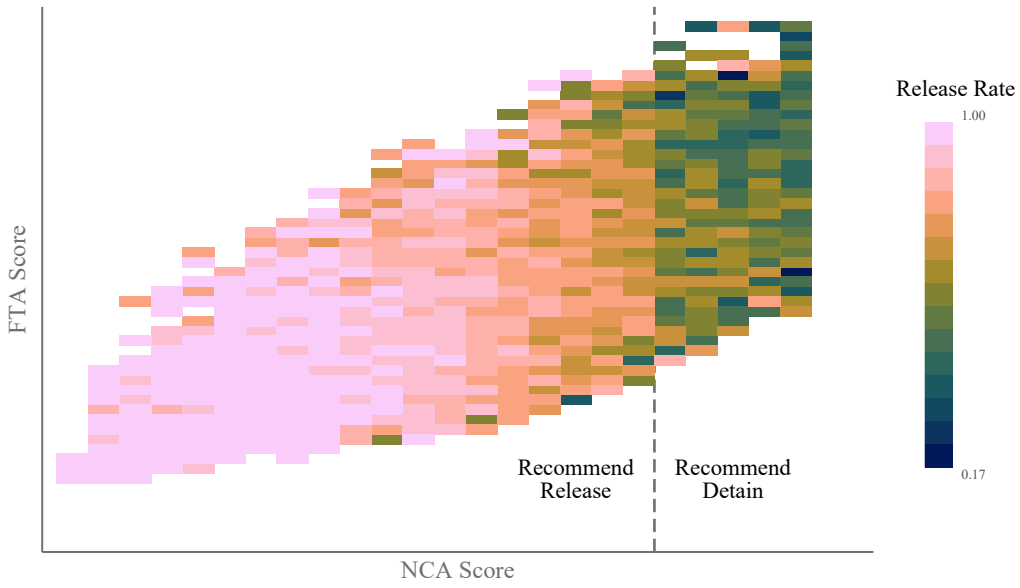
Notes. This figure plots the worst- and best-case bounds of the algorithmic counterfactual. The region between the constructed worst- and best-case bounds of the algorithmic line is shaded in orange. We also include quasi-experimental estimates of the algorithmic counterfactual for comparison. All estimates adjust for shift-by-time fixed effects. See the notes for Figure 5 for additional details.

Figure A.2: Conditional Misconduct Rates by FTA and NCA Risk Scores



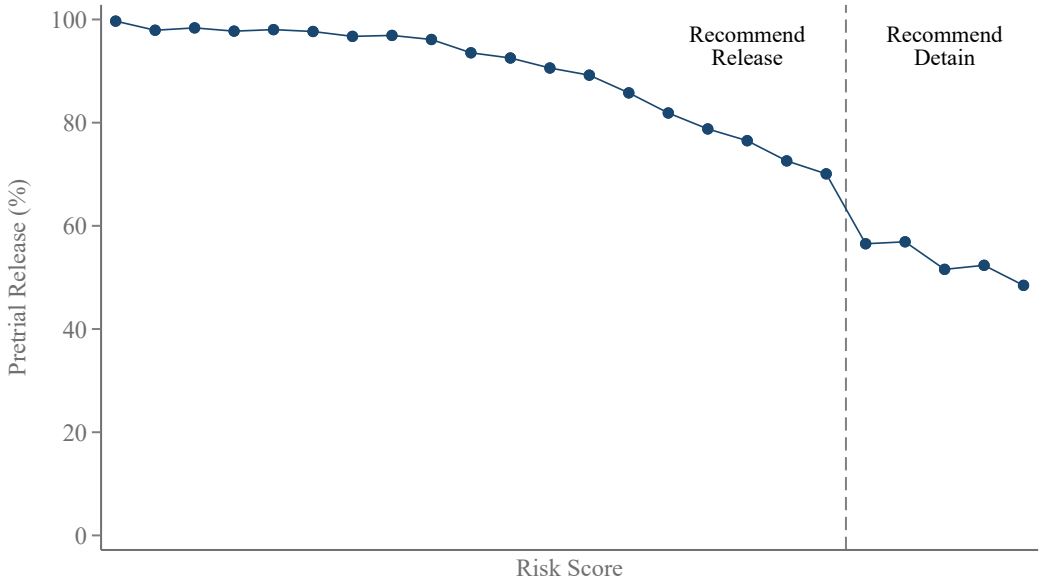
Notes. This figure plots average conditional misconduct rates by FTA and NCA algorithmic risk scores. Panel A plots the conditional misconduct rate for each NCA risk score. Panel B plots the conditional misconduct rate for each FTA risk score, where we aggregate small cells so that each contains at least 150 cases. Both panels also plot the conditional misconduct rate for a gradient-boosted decision trees algorithm constructed using the same observable characteristics as the original algorithm and any pretrial misconduct as an outcome. See the text for additional details.

Figure A.3: Pretrial Release Rates by FTA and NCA Risk Scores



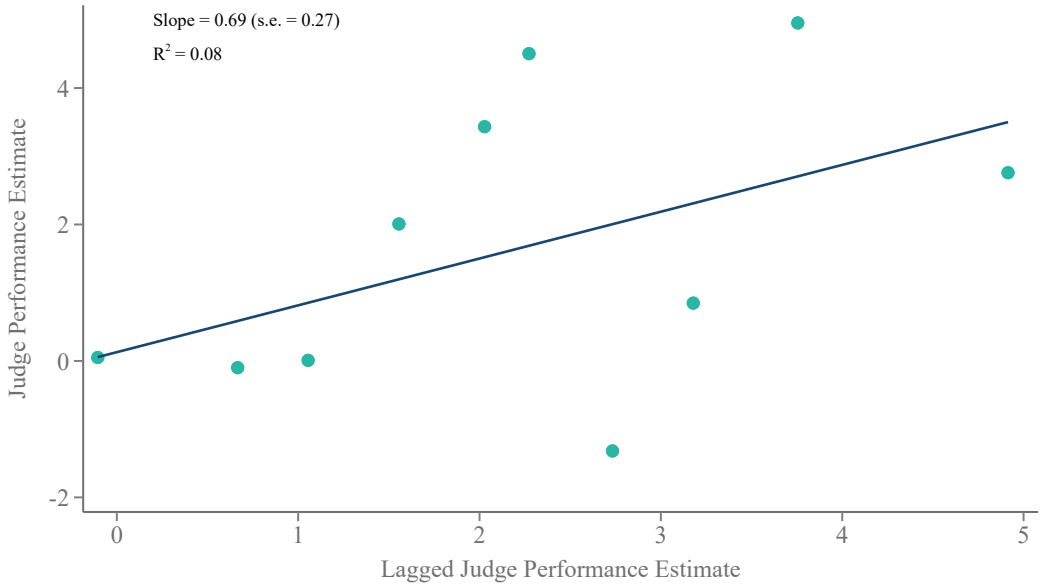
Notes. This figure plots average release rates by FTA and NCA algorithmic risk scores. We limit the sample to cells with at least 5 cases. The dashed vertical line indicates the NCA score where the algorithmic recommendation changes from release to detain.

Figure A.4: Pretrial Release Rates by NCA Risk Score



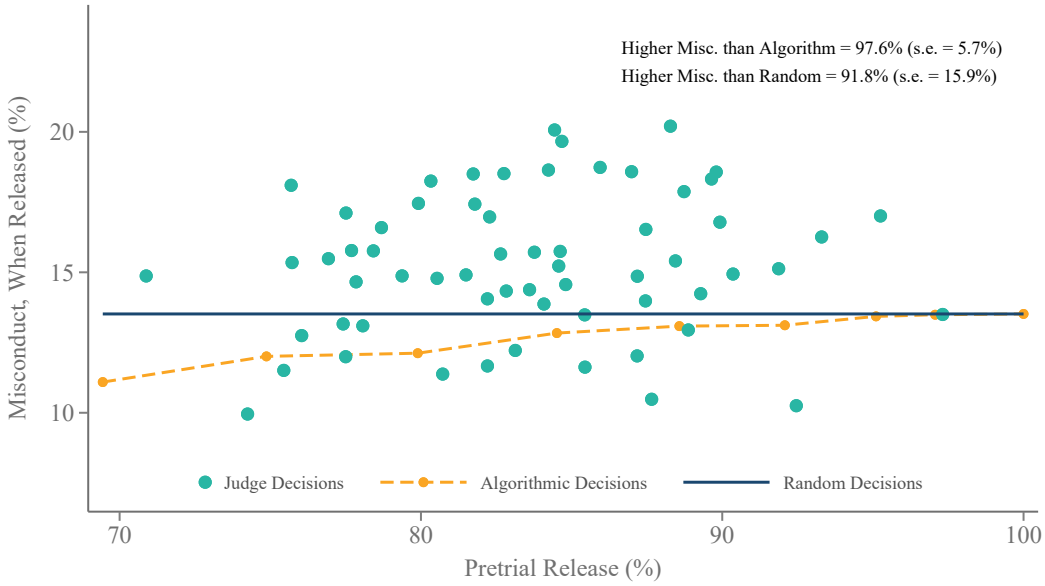
Notes. This figure plots pretrial release rates by NCA algorithmic risk scores. The dashed vertical line indicates the risk score where the algorithmic recommendation changes from release to detain.

Figure A.5: Out-of-Sample Predictive Power of Judge Performance



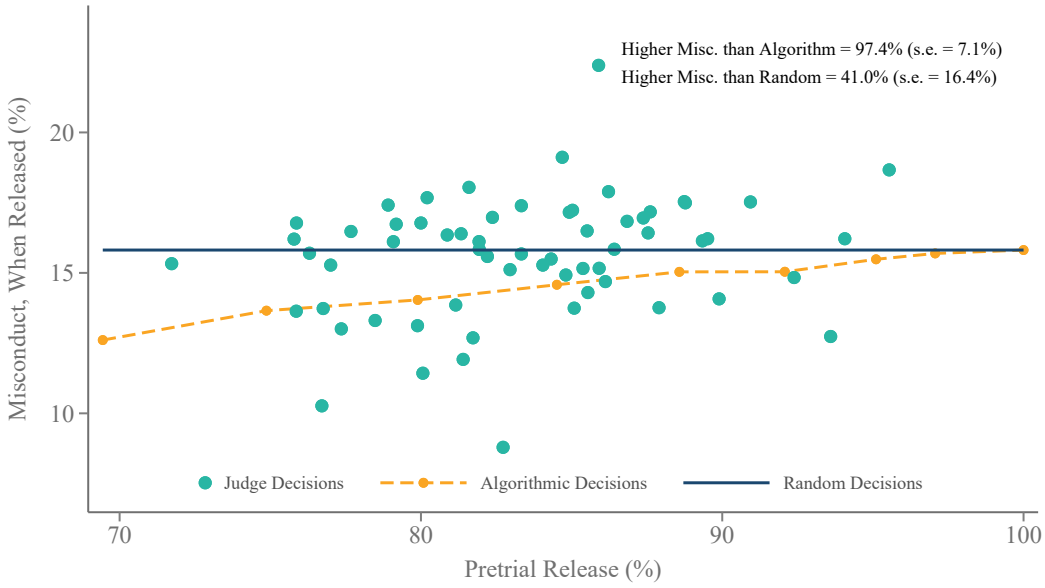
Notes. This figure shows the relationship between current judge performance estimates and lagged judge performance posteriors for the 54 judges with at least 100 cases in each period. Lagged judge performance is measured using the first half of cases that each judge sees in our sample period and current judge performance is measured using the remaining second half of cases. We compute posteriors of lagged judge performance using a conventional empirical Bayes "shrinkage" procedure and weight by estimates of the inverse variance of the current judge performance estimates. Adjusting for the noise in the judge performance estimates yields a bias-corrected R^2 of 0.41, indicating that the correlation between predictions in the first period and the latent true performance in the second period is $\sqrt{0.41} = 0.64$. The judge-level data are presented in 10 bins. The slope, standard error, and R^2 reported are for a linear regression fit on the underlying judge-level data. See the text for additional details.

Figure A.6: Results with Local Linear Extrapolations of Conditional Misconduct Rates



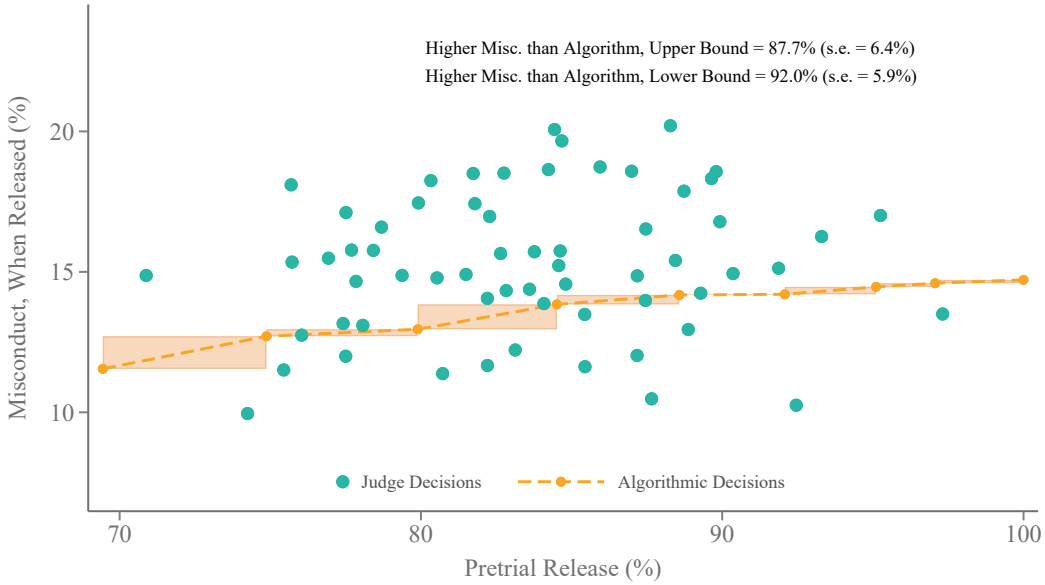
Notes. The figure shows results when we construct counterfactual misconduct rates under the algorithm using local linear extrapolations of mean risk. See the notes for Figure 5 for additional details.

Figure A.7: Results Without Shift-by-Time Effects



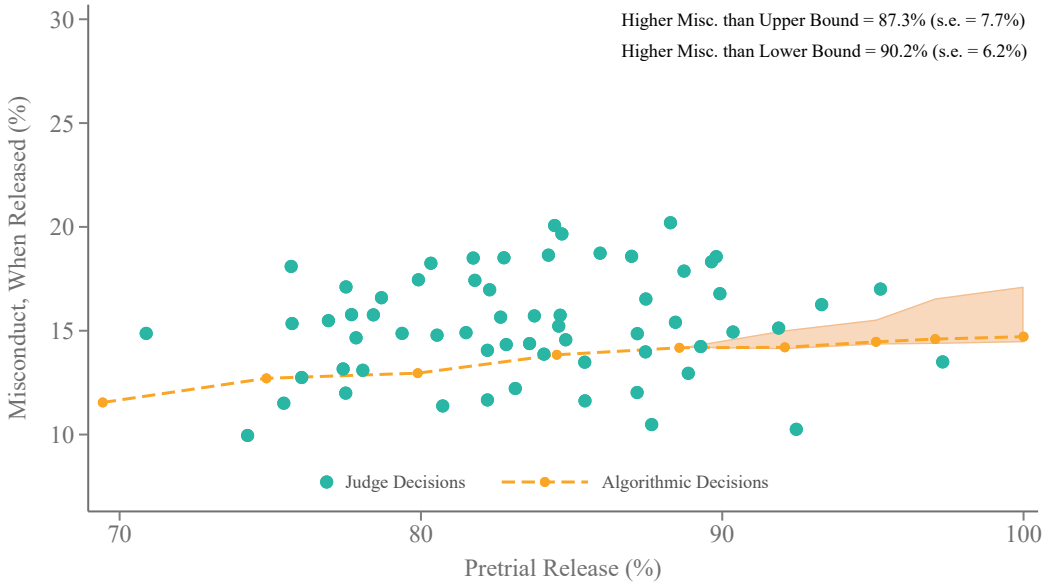
Notes. This figure shows results when we omit shift-by-time effects. See the notes for Figure 5 for additional details.

Figure A.8: Stepwise Connections of Discrete Risk Scores



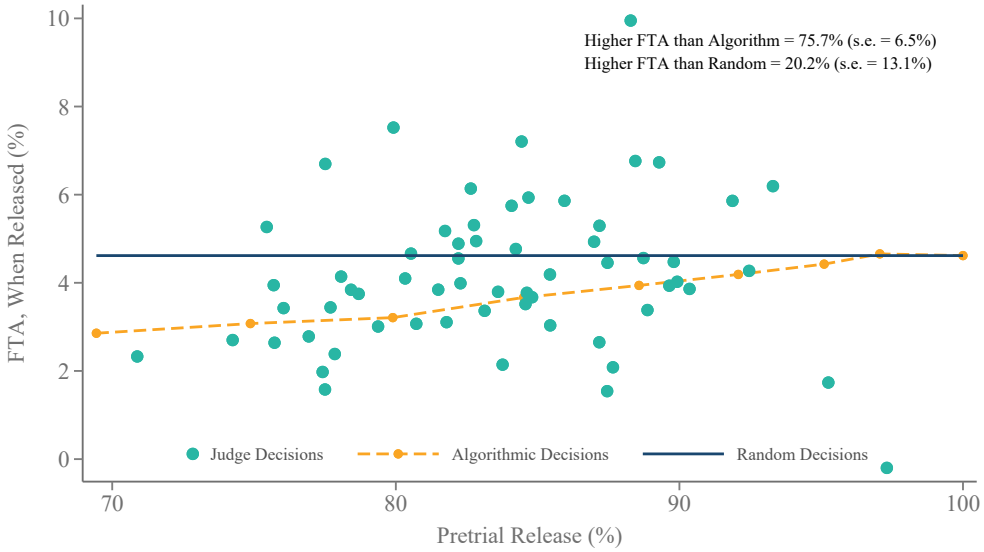
Notes. This figure shows the stepwise connections of the discrete risk scores. See the notes for Figure 5 for additional details.

Figure A.9: Results with Extrapolations to the Most Lenient Judge



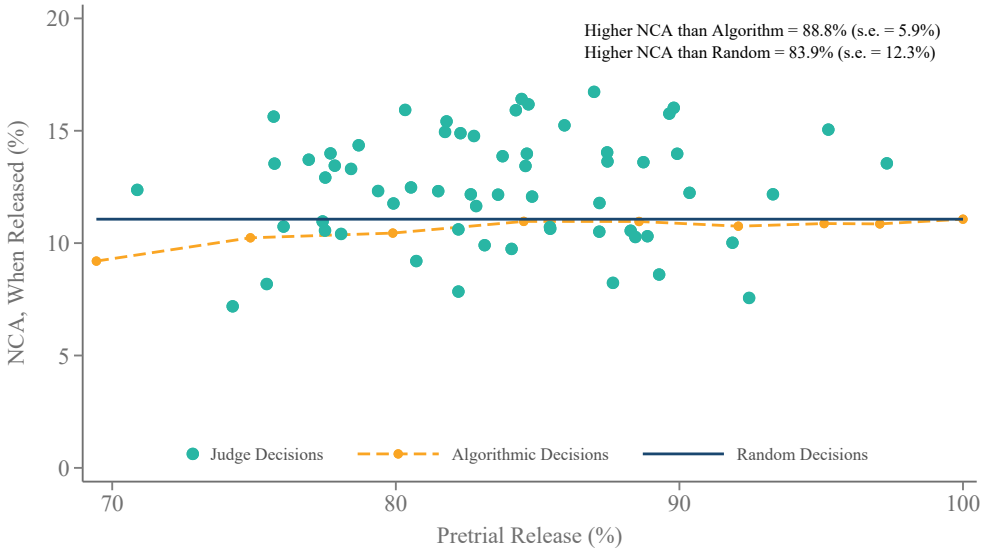
Notes. This figure shows results when we construct counterfactual misconduct rates under the algorithm using extrapolations to only the most lenient judge at each risk score cutoff and then calculate worst- and best-case bounds for the remaining fraction of defendants. The region between the constructed worst- and best-case bounds of the algorithmic line is shaded in orange. All estimates adjust for shift-by-time fixed effects. See the notes for Figure 5 for additional details.

Figure A.10: Results for FTA Only



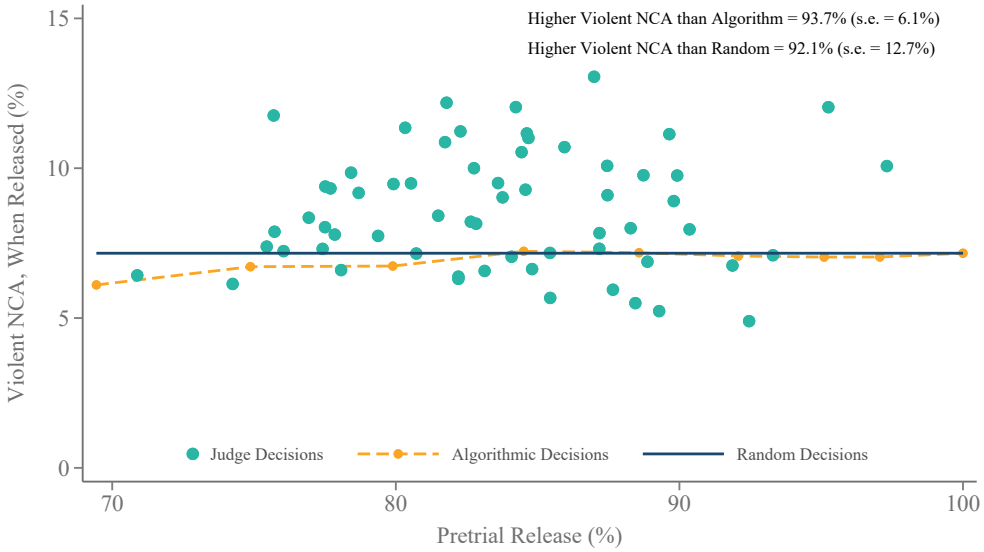
Notes. This figure shows results where pretrial misconduct is defined as FTA only. See the notes for Figure 5 for additional details.

Figure A.11: Results for NCA Only



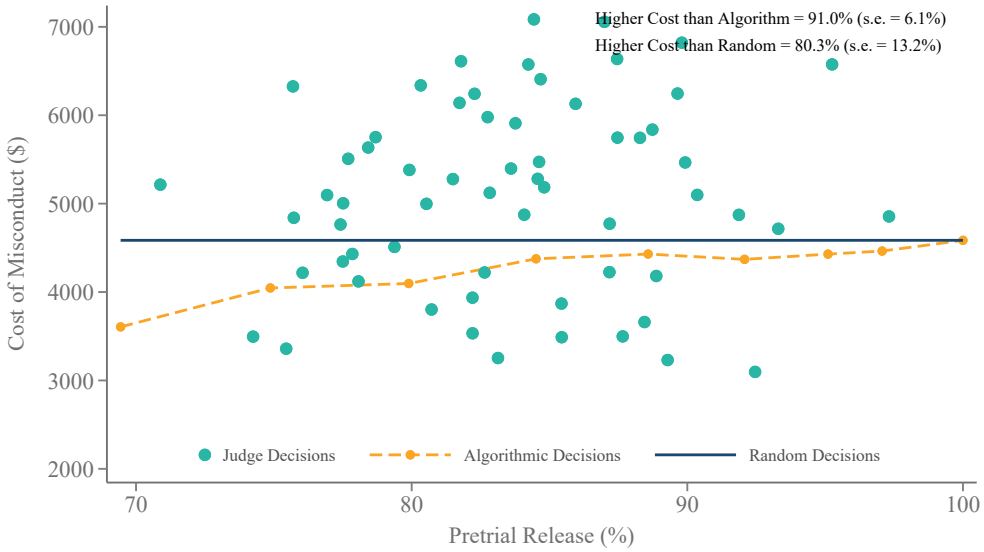
Notes. This figure shows results where pretrial misconduct is defined as NCA only. See the notes for Figure 5 for additional details.

Figure A.12: Results for Violent NCA Only



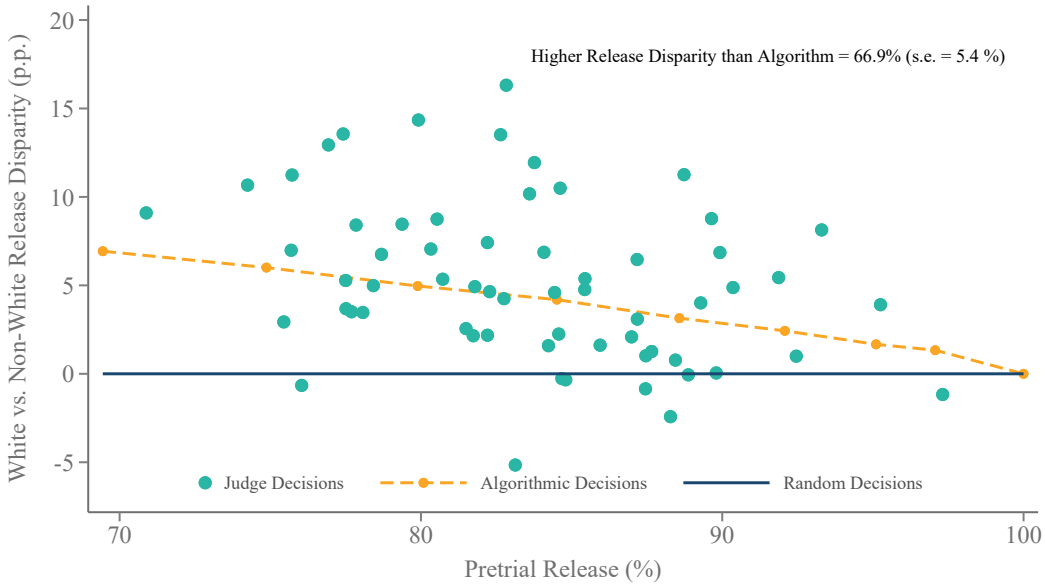
Notes. This figure shows results where pretrial misconduct is defined as violent NCA only. See the notes for Figure 5 for additional details.

Figure A.13: Results for the Social Cost of Misconduct



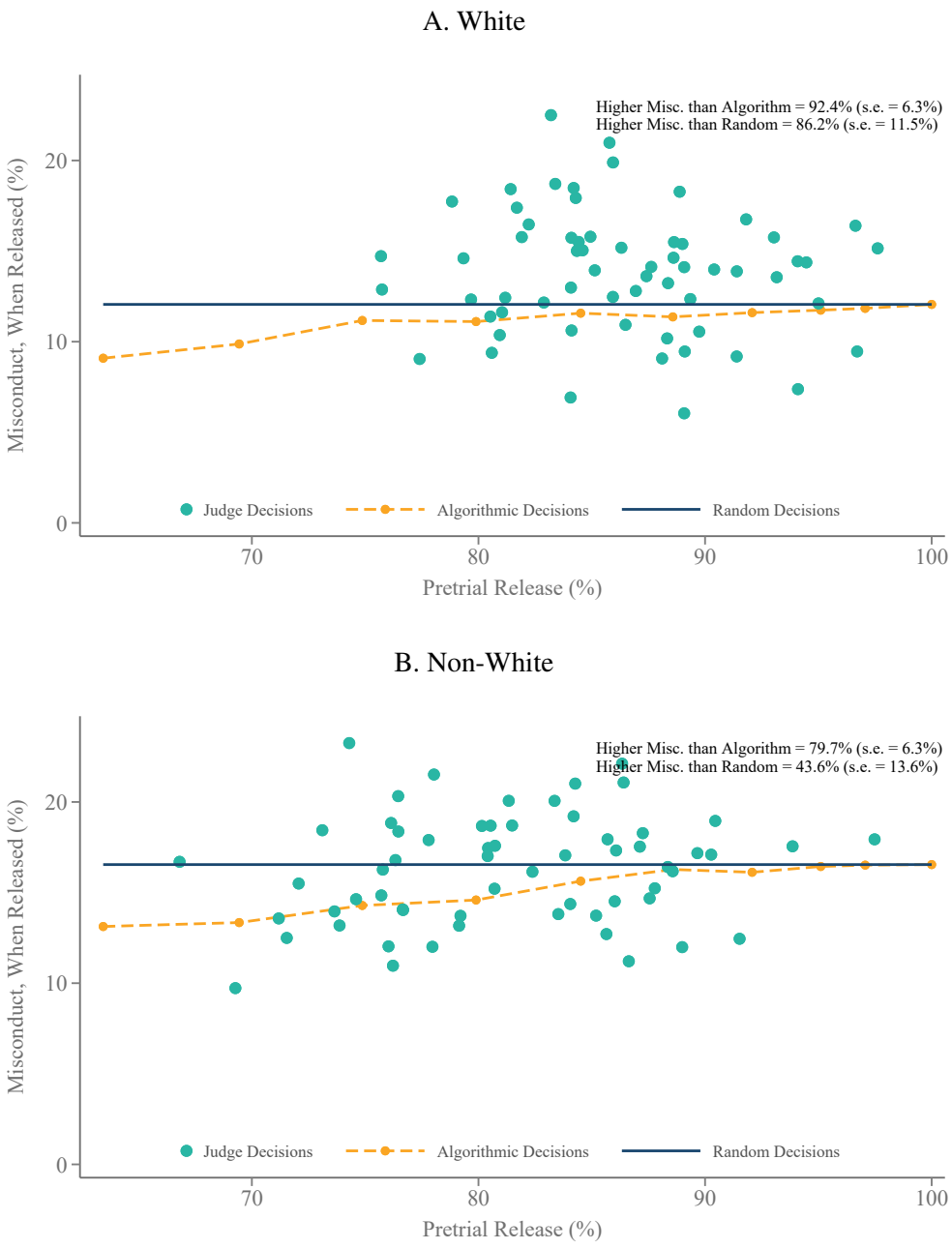
Notes. This figure shows results for the social cost of pretrial misconduct. We first calculate the social cost for FTAs, DUIs, drug offenses, motor vehicle offenses, person offenses, property offenses, public order offenses, and weapons offenses using the estimates from Dobbie, Goldin, and Yang (2018) and Miller et al. (2021). Within each of these misconduct types, we use the lowest social cost estimate available. We next use these estimates to calculate the social cost of releasing each defendant, where we assign a social cost of $0 to cases without any misconduct and the lowest available cost amount to the 1.0% of cases with a misconduct type that is not categorized. We then perform our main analysis using the new social cost measure as the outcome variable. See the notes for Figure 5 for additional details.

Figure A.14: White vs. Non-White Release Disparities Relative to the Algorithm



Notes. This figure plots release rates against release disparities between white and non-white defendants with identical misconduct potential, along with counterfactual disparities for algorithmic and random decisions. All estimates are based on a linear extrapolation of the race-specific misconduct rates and adjust for shift-by-time fixed effects. The figure also reports the fraction of judges with higher release disparities compared to the algorithmic release rule, computed from these estimates as posterior average effect. Standard errors are estimated by taking random draws from the distributions of the estimated judge-specific release rates and conditional misconduct rates and then recalculating the statistics of interest.

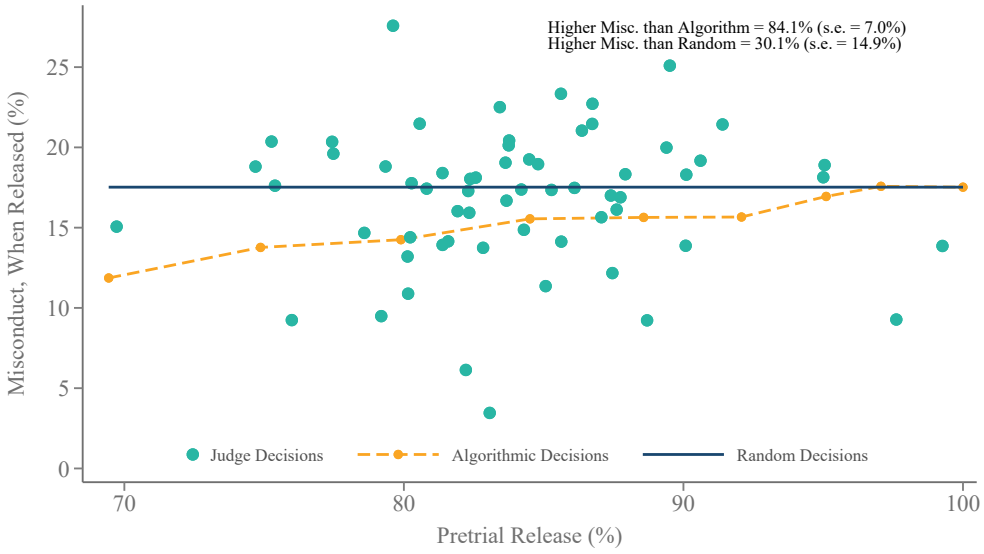
Figure A.15: Results by Defendant Race



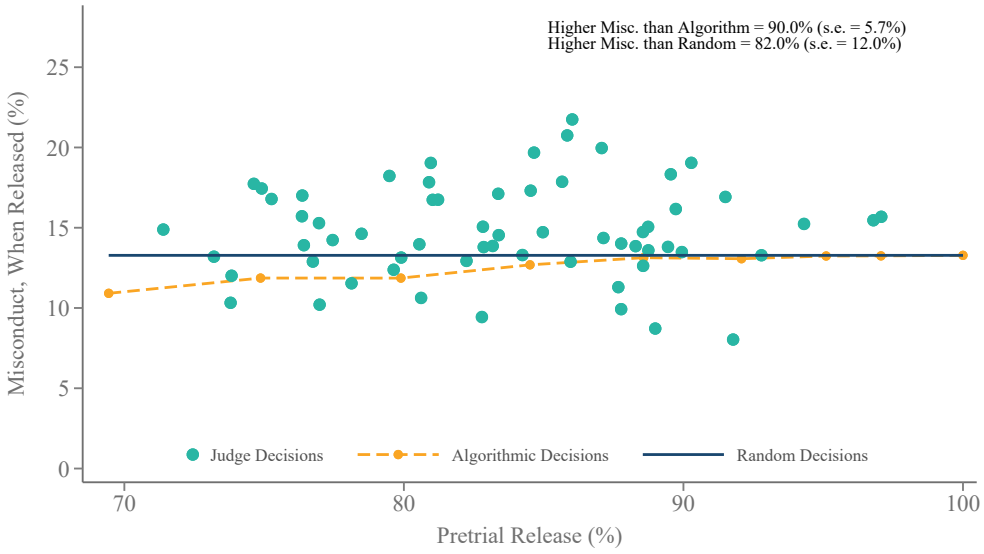
Notes. This figure shows results by defendant race. Each panel plots judge-specific release rates and conditional misconduct rates, the algorithmic counterfactual and the random release rule for each subgroup. The figures also report the fraction of judges with higher misconduct rates compared to the algorithmic and the random release rules for each subgroup. See the notes for Figure 5 for additional details. Panel B omits one outlier point.

Figure A.16: Results by Defendant Age

A. Age 25 or Younger



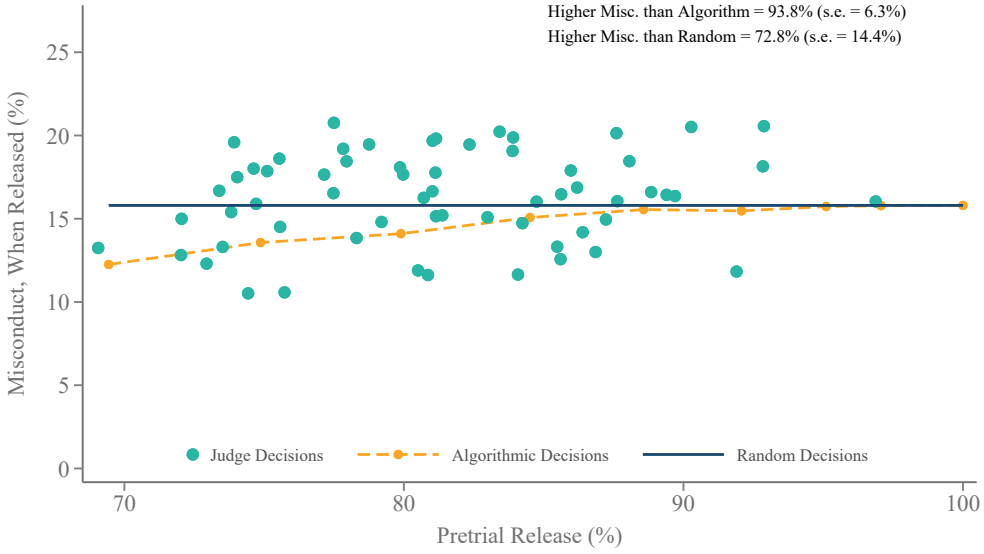
B. Age 26 or Older



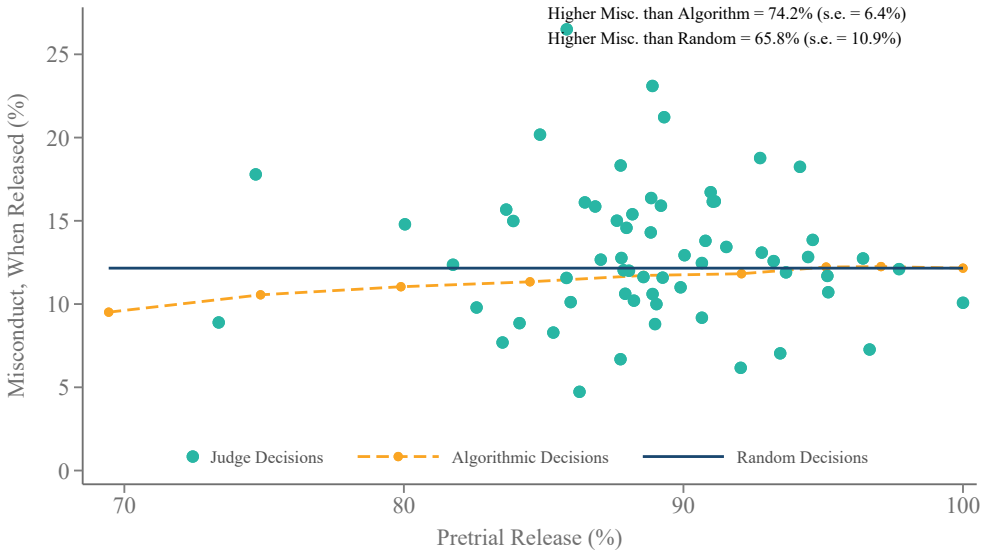
Notes. This figure shows results by defendant age. Each panel plots judge-specific release rates and conditional misconduct rates, the algorithmic counterfactual and the random release rule for each subgroup. The figures also report the fraction of judges with higher misconduct rates compared to the algorithmic and the random release rules for each subgroup. See the notes for Figure 5 for additional details.

Figure A.17: Results by Defendant Education

A. 12 or Fewer Years of Schooling



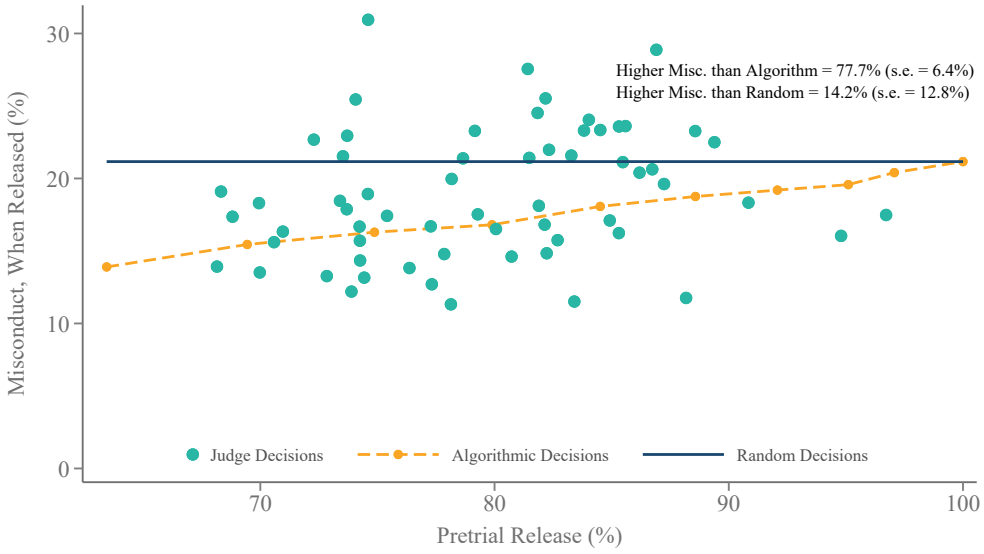
B. More than 12 Years of Schooling



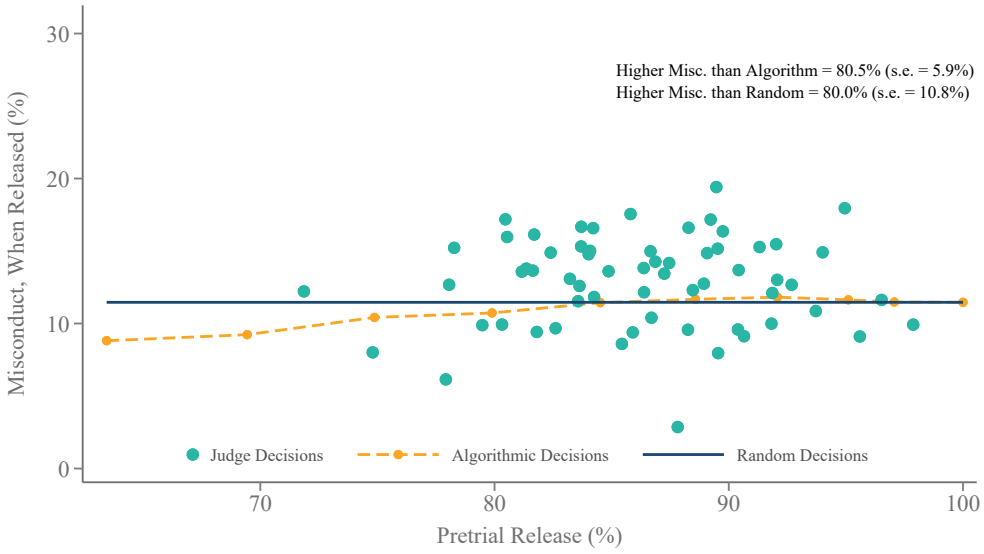
Notes. This figure shows results by defendant education. Each panel plots judge-specific release rates and conditional misconduct rates, the algorithmic counterfactual and the random release rule for each subgroup. The figures also report the fraction of judges with higher misconduct rates compared to the algorithmic and the random release rules for each subgroup. See the notes for Figure 5 for additional details.

Figure A.18: Results by Charge Type

A. Felony Charge

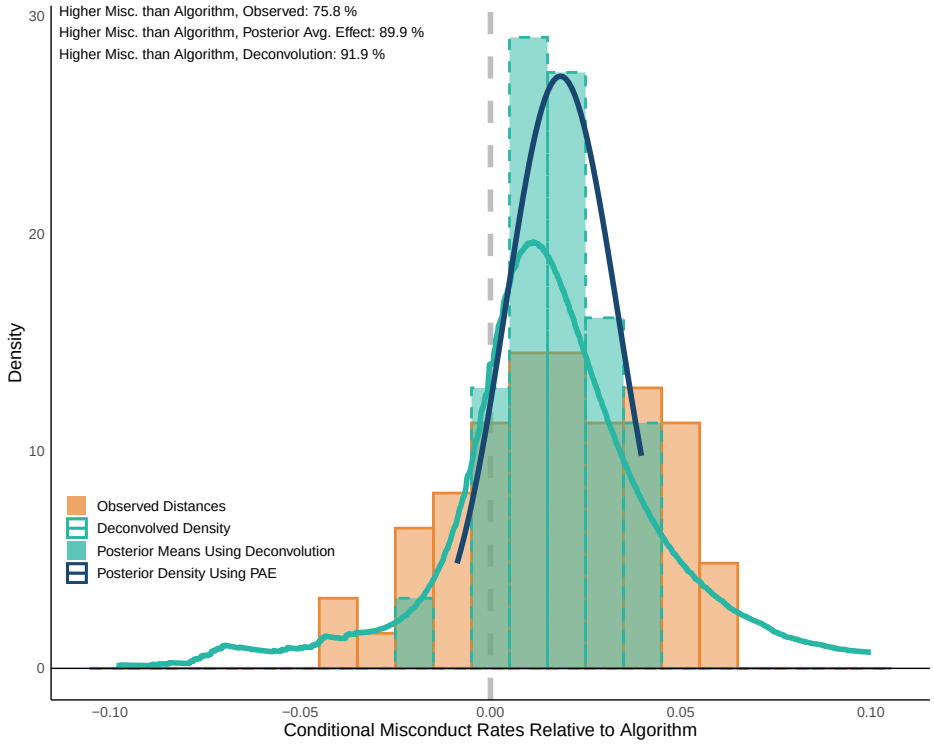


B. Non-Felony Charge



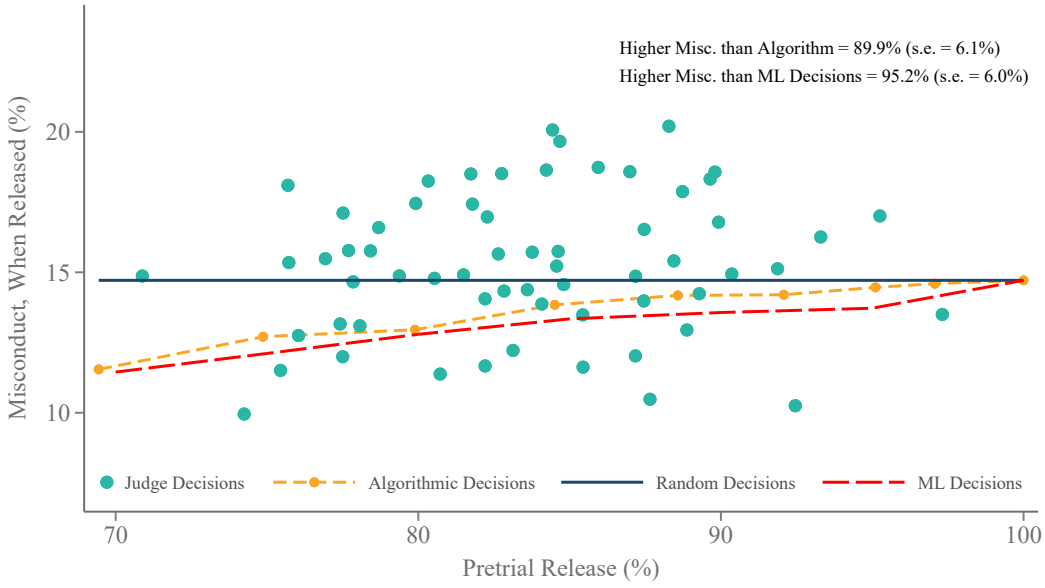
Notes. This figure shows results by whether or not the defendant had a felony charge. Each panel plots judge-specific release rates and conditional misconduct rates, the algorithmic counterfactual and the random release rule for each subgroup. The figures also report the fraction of judges with higher misconduct rates compared to the algorithmic and the random release rules for each subgroup. See the notes for Figure 5 for additional details.

Figure A.19: Distribution of Judge Performance Relative to the Algorithm



Notes. This figure plots estimates of the distribution of judge performance relative to the algorithm, as measured by the distance between each judge's conditional misconduct rate and algorithmic counterfactual at the same release rate. The navy line plots the estimate of the posterior distribution of conditional misconduct rates relative to the algorithm using the posterior average effects approach of Bonhomme and Weidner (2022). The orange histogram shows the observed distribution of judge performance relative to the algorithm. The light green line shows the deconvolved prior density of the population distribution of conditional misconduct rates relative to the algorithm estimated using the deconvolveR package from Narasimhan and Efron (2020). The light green histogram plots the posterior means of the conditional misconduct rates relative to the algorithm relying on the deconvolved prior density. The figure also reports the fraction of judges with higher misconduct rates compared to the algorithmic counterfactual in the observed data, as computed using posterior average effects, and as computed using the posterior means from the deconvolution approach.

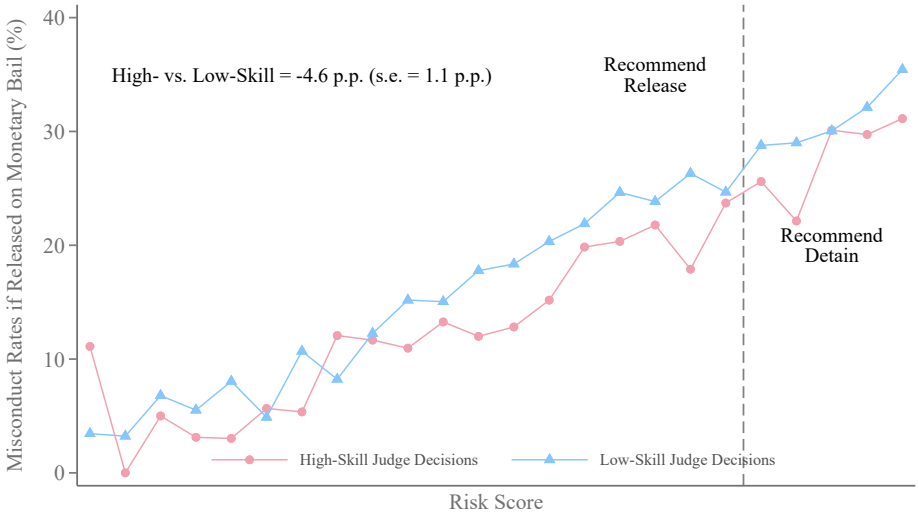
Figure A.20: Comparison to Machine Learning Algorithm



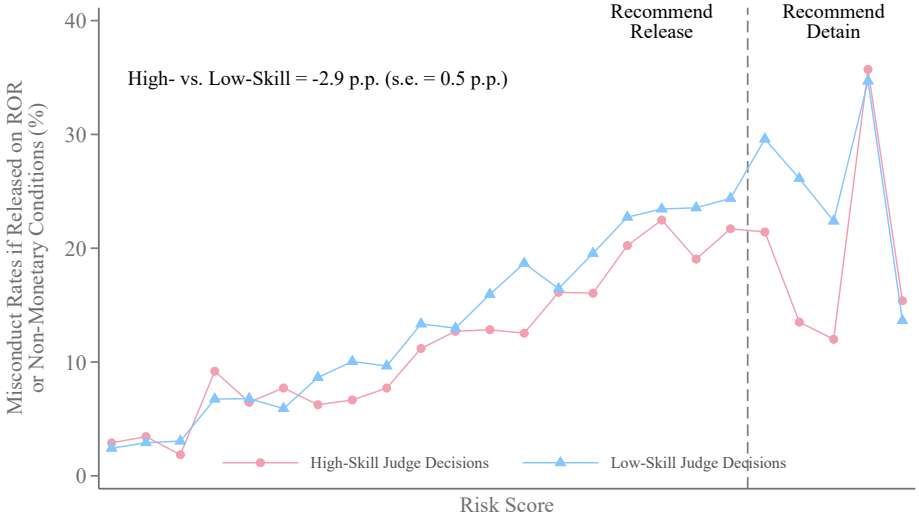
Notes. The figure shows results for a gradient-boosted decision trees algorithm constructed using the same observable characteristics as the original algorithm. See the notes for Figure 5 for additional details.

Figure A.21: Conditional Misconduct Rates by Release Type and Judge Skill

A. Misconduct Rates if Released on Monetary Bail

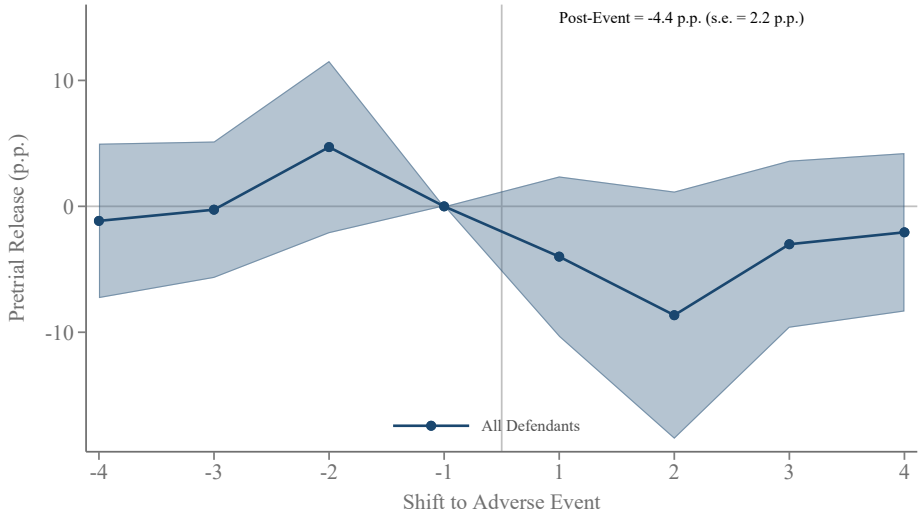


B. Misconduct Rates if Released on ROR or Non-Monetary Conditions



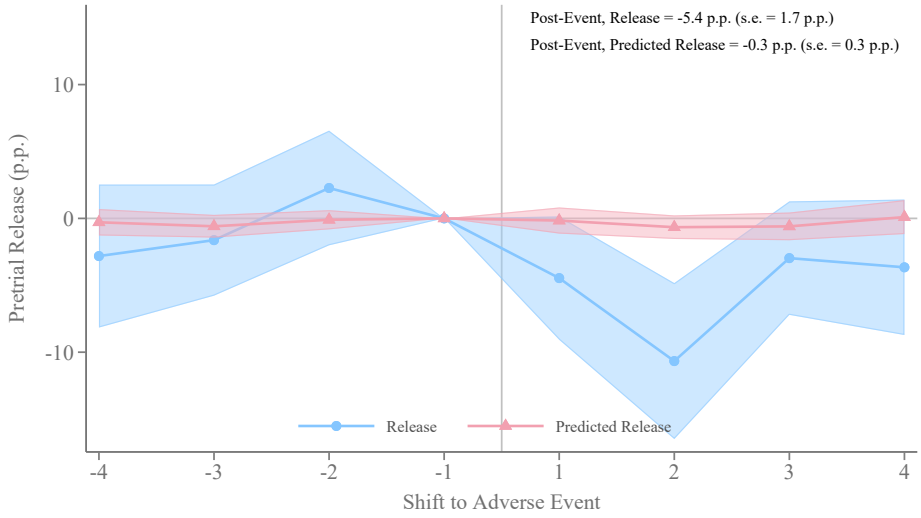
Notes. This figure plots conditional misconduct rates if released on monetary bail and conditional misconduct rates if released under ROR or non-monetary conditions against algorithmic NCA risk scores. The dashed vertical line indicates the score where the algorithmic recommendation changes from release to detain. All panels report the regression coefficient from a regression of the indicated outcome on a high-skill judge indicator, the algorithmic score, and shift-by-time fixed effects. Robust standard errors clustered at the judge level are included in parentheses. The p-value on a test for equality of the coefficients from Panels A and B is 0.238.

Figure A.22: Effect of Adverse Events with Never-Treated Judges as the Control Group



Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release, restricting the control group to never-treated judges. We first group the treated judges according to the date when they first heard a case involving a defendant arrested for a serious violent felony while on pretrial. We then create a separate dataset for each treated judge with only the never-treated judges as a control group. Finally, we stack all the datasets to compare the outcomes of each cohort of treated judges to the outcomes of the never-treated judges. See footnote 15 for additional details.

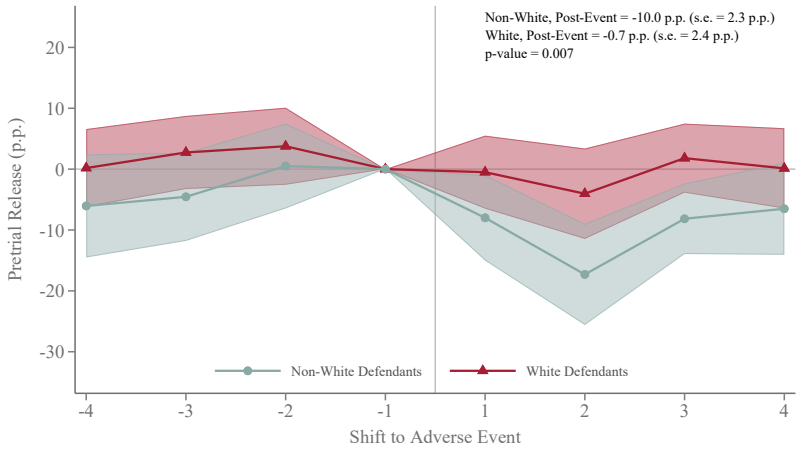
Figure A.23: Effect of Adverse Events on Predicted Release



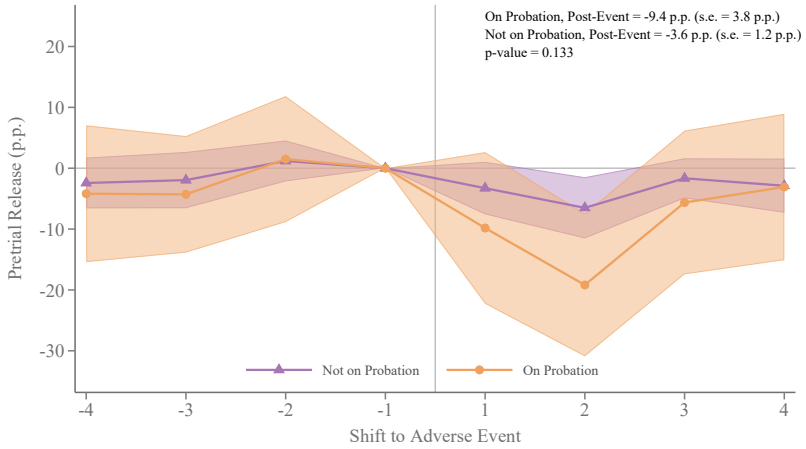
Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release on predicted pretrial release. For each case, we predict release using all the private and observable information available (see Table 3). See the notes of Figure 9 for additional details.

Figure A.24: Effect of Adverse Events on Pretrial Release by Defendant Race and Probation Status

A. By Defendant Race

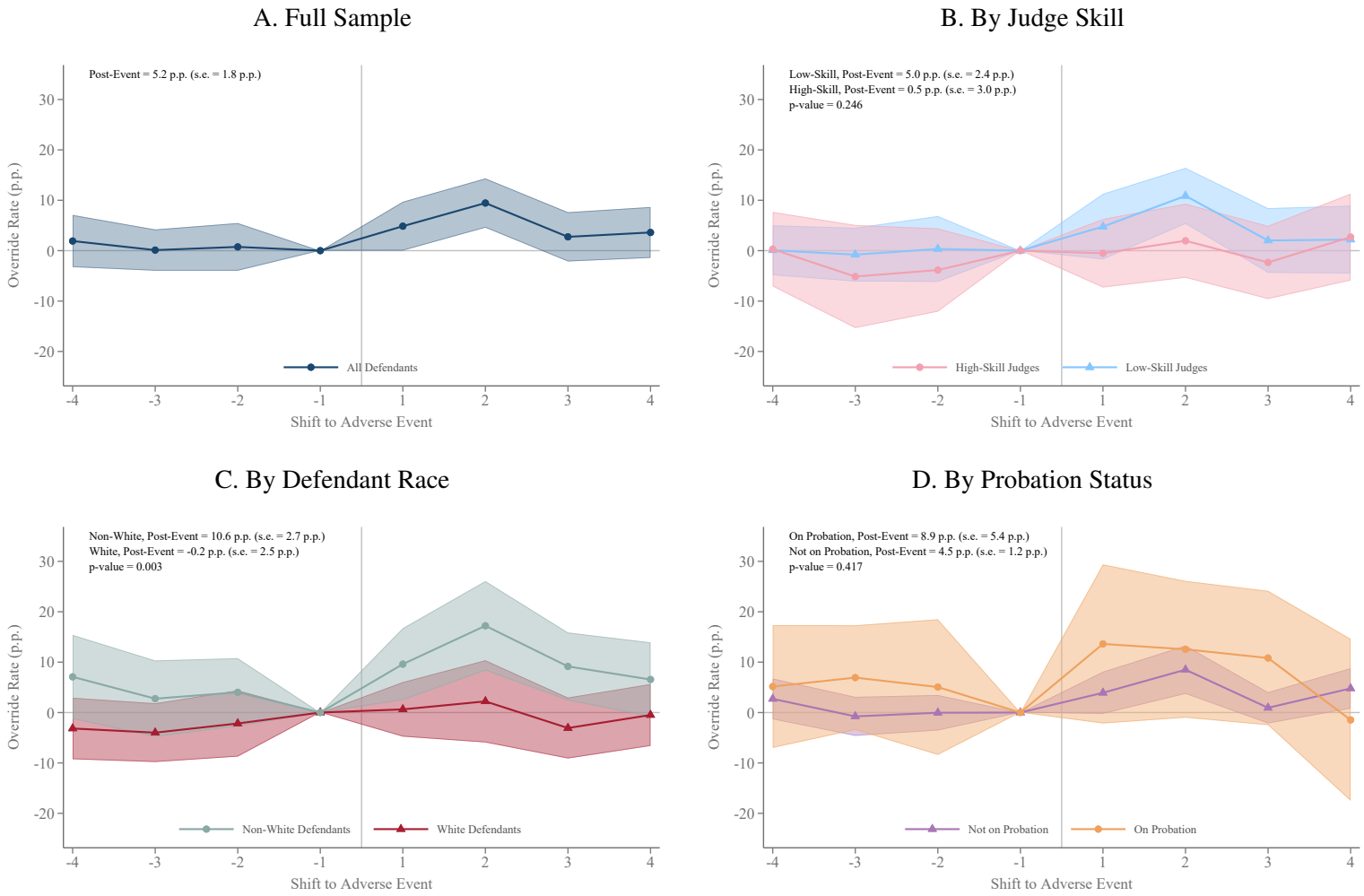


B. By Probation Status



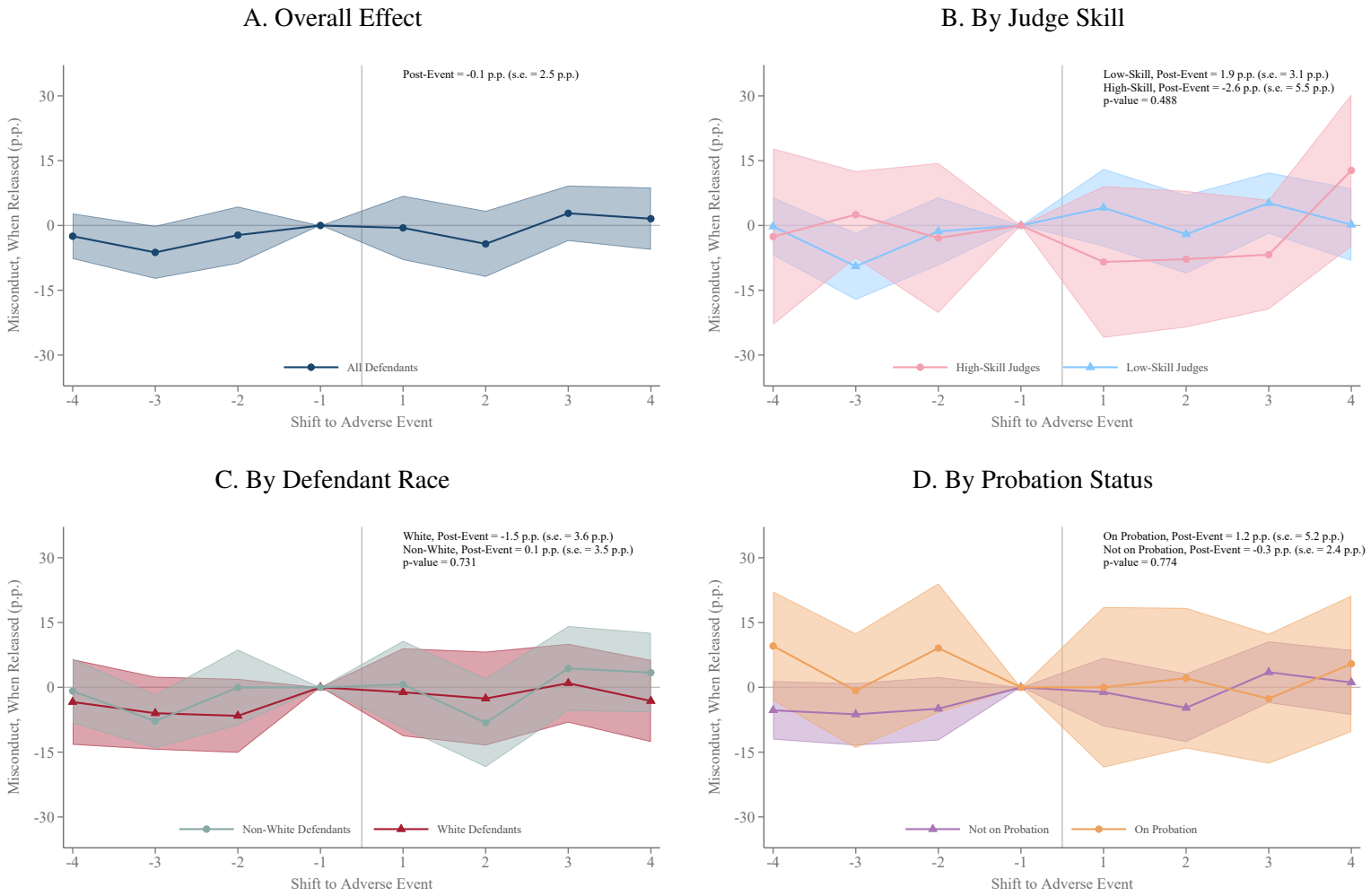
Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release by defendant race and probation status. See the notes of Figure 9 for additional details.

Figure A.25: Effect of Adverse Events on Harsh Overrides



Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release on harsh overrides for observably low-risk defendants. See the notes of Figure 9 for additional details.

Figure A.26: Effect of Adverse Events on Conditional Misconduct Rates



Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release on conditional misconduct rates. See the notes of Figure 9 for additional details.

Table A.1: Algorithmic Release Recommendations

	NCA Score = 1	**NCA Score = 2**	**NCA Score = 3**	**NCA Score = 4**	**NCA Score = 5**	**NCA Score = 6**
FTA Score = 1	Release	Release	Release	Release + Phone	Release + In Person	Detention
FTA Score = 2	Release	Release	Release	Release + Phone	Release + In Person	Detention
FTA Score = 3	Release	Release	Release + Phone	Release + In Person	Release + In Person	Detention
FTA Score = 4	Release + Phone	Release + Phone	Release + In Person	Release + In Person	Release + In Person	Detention
FTA Score = 5	Release + Phone	Release + Phone	Release + In Person	Release + In Person	Release + In Person	Detention
FTA Score = 6	Release + Phone	Release + Phone	Release + In Person	Release + In Person	Release + In Person	Detention

Notes. This table shows the automatic release recommendations generated by the FTA and NCA binned scores described in the text. Release recommendations indicate release with no conditions, or ROR. Release and phone recommendations indicate release with the condition of making regular phone check-ins. Release and in person recommendations indicate release with the condition of making regular in-person check-ins. Detention recommendations indicate detention with no money bail option.

Table A.2: Tests of Quasi-Random Judge Assignment

	All Defendants	Recommend Release	Recommend Detain
	(1)	(2)	(3)
Age at Current Arrest	-0.00000	0.00000	-0.00006
	(0.00002)	(0.00002)	(0.00005)
Age at First Arrest	-0.00001	-0.00002	-0.00003
	(0.00002)	(0.00002)	(0.00013)
Prior Arrests	0.00002	0.00002	-0.00001
	(0.00003)	(0.00004)	(0.00006)
Prior Felonies	-0.00003	-0.00005	0.00015
	(0.00008)	(0.00009)	(0.00021)
Prior Misdemeanors	-0.00000	0.00000	0.00012
	(0.00007)	(0.00009)	(0.00015)
Pending Charges	-0.00008	0.00016	0.00003
	(0.00021)	(0.00036)	(0.00030)
Property Charge	0.00052	0.00104	-0.00177
	(0.00059)	(0.00063)	(0.00140)
Drug Charge	0.00018	-0.00002	0.00115
	(0.00034)	(0.00032)	(0.00104)
Public Order Charge	-0.00049	-0.00032	-0.00064
	(0.00038)	(0.00040)	(0.00087)
Traffic Charge	0.00028	0.00054	-0.00091
	(0.00048)	(0.00053)	(0.00090)
Parole/Probation	0.00015	0.00030	-0.00019
	(0.00036)	(0.00032)	(0.00108)
Pretrial Release	0.00063	0.00073	0.00040
	(0.00092)	(0.00106)	(0.00136)
Male	-0.00014	0.00010	-0.00121
	(0.00045)	(0.00044)	(0.00121)
White	-0.00008	-0.00015	0.00058
	(0.00036)	(0.00037)	(0.00076)
Homeless	-0.00115	-0.00103	-0.00301
	(0.00073)	(0.00096)	(0.00138)
No Telephone	-0.00139	-0.00118	-0.00186
	(0.00076)	(0.00071)	(0.00158)
Out-of-State Address	-0.00001	-0.00006	0.00014
	(0.00076)	(0.00079)	(0.00393)
Violent Charge Against an Adult	-0.00082	-0.00042	-0.00251
	(0.00053)	(0.00057)	(0.00135)
Violent Charge Against a Child	-0.00009	-0.00019	0.00409
	(0.00069)	(0.00069)	(0.00273)
Any Aggravating Condition	-0.00028	-0.00048	-0.00240
	(0.00067)	(0.00068)	(0.01224)
Override Recommendation	-0.00108	-0.00111	-0.00543
	(0.00061)	(0.00070)	(0.00217)
Joint p-value	[0.278]	[0.247]	[0.062]
Shift x Time FE	Yes	Yes	Yes
Cases	37,855	31,992	5,795

Notes. This table reports OLS estimates of regressions of judge leniency on case and defendant characteristics. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a given bail judge. All regressions control for shift-by-time fixed effects. The p-values reported at the bottom of each column are from F-tests of the joint significance of the variables listed in the rows. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Table A.3: First-Stage Effects of Judge Leniency

	All Defendants	Recommend Release	Recommend Detain
	(1)	(2)	(3)
Leave-Out Judge Leniency	95.4	94.0	123.1
	(6.3)	(6.5)	(25.2)
Shift x Time FE	Yes	Yes	Yes
Mean Release Rate	82.8	88.2	53.6
Cases	37,855	31,992	5,795

Notes. This table reports OLS estimates of regressions of an indicator for pretrial release on judge leniency. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a given bail judge. All regressions control for shift-by-time fixed effects. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses. After accounting for the shift-by-time fixed effects, there are 68 singleton observations in the recommend detain sample, which are automatically dropped in the regression in column 3.

Table A.4: Extrapolations of Conditional Misconduct Rates

	Release Rate	Linear Extrapolation	Local Linear Extrapolation
	(1)	(2)	(3)
NCA Score ≤ 15	69.4	11.5	11.1
		(0.7)	(0.7)
NCA Score ≤ 16	74.9	12.7	12.0
		(0.8)	(0.8)
NCA Score ≤ 17	79.9	13.0	12.1
		(0.8)	(0.8)
NCA Score ≤ 18	84.5	13.8	12.8
		(0.8)	(0.9)
NCA Score ≤ 19	88.6	14.2	13.1
		(0.8)	(1.0)
NCA Score ≤ 20	92.1	14.2	13.1
		(0.8)	(1.0)
NCA Score ≤ 21	95.1	14.5	13.4
		(0.9)	(1.1)
NCA Score ≤ 22	97.1	14.6	13.5
		(1.0)	(1.2)
NCA Score ≤ 23	100.0	14.7	13.5
		(1.0)	(1.3)
Shift x Time FE		Yes	Yes
Judges		62	62

Notes. This table reports extrapolation-based estimates of the conditional misconduct rate at each risk score cutoff. Column 1 reports the fraction of defendants released at each risk score cutoff. Column 2 reports results based on a linear extrapolation. Column 3 reports results based on a local linear extrapolation with a Gaussian kernel and a fixed bandwidth. All columns control for shift-by-time fixed effects in the estimation of the judge-specific release rates and misconduct rates. Standard errors are obtained through a bootstrapping procedure described in the text and appear in parentheses. See the text for additional details.

Table A.5: Conditional Misconduct Rates Relative to the Algorithm and Judge Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Override Rate (0-100)	0.20									0.20
	(0.12)									(0.11)
Above Median Experience		1.08								0.15
		(0.81)								(0.77)
Male			-1.05							-1.62
			(0.96)							(1.14)
White				-0.32						-1.05
				(1.21)						(1.11)
Registered Republican					1.64					1.07
					(0.92)					(0.89)
Law Degree						-0.99				-0.35
						(0.86)				(0.84)
Former Prosecutor							-0.69			0.14
							(0.92)			(1.01)
Former Police Officer								2.54		2.37
								(0.71)		(0.93)
White vs. Non-White Disparity (0-100)									-0.08	-0.20
									(0.52)	(0.43)
R^2	0.06	0.04	0.02	0.00	0.07	0.03	0.00	0.15	0.00	0.30
Judges	62	62	62	62	62	62	62	62	62	62

Notes. This table reports OLS estimates of regressions of each judge's conditional misconduct rate relative to the algorithm on judge characteristics. Information on the judge demographics is derived from publicly available voter data and official publications. Judge performance, override rates, and white vs. non-white release disparities are estimated using the administrative court data as described in the text. The white vs. non-white release disparities are empirical Bayes posteriors computed using a standard shrinkage procedure. Robust standard errors are reported in parentheses. See the text for additional details.

Table A.6: Characteristics of Released Defendants

	Recommend Release			Recommend Detain		
	High-Skill Judges	Low-Skill Judges	p-value	High-Skill Judges	Low-Skill Judges	p-value
A. Observable Information	(1)	(2)	(3)	(4)	(5)	(6)
Age at Current Arrest	0.29	0.09	0.73	2.89	-1.17	0.07
	(0.50)	(0.29)		(1.74)	(1.41)	
Age at First Arrest	-0.10	-0.11	0.85	-0.35	0.56	0.07
	(0.06)	(0.05)		(0.44)	(0.26)	
Prior Arrests	-0.39	-0.16	0.09	-0.16	-0.14	0.92
	(0.11)	(0.08)		(0.18)	(0.11)	
Prior Felonies	0.31	-0.19	0.12	0.49	-0.02	0.39
	(0.27)	(0.19)		(0.46)	(0.37)	
Prior Misdemeanors	0.70	0.10	0.07	-0.47	-0.10	0.52
	(0.29)	(0.17)		(0.52)	(0.26)	
Pending Charges	0.31	-3.44	0.00	0.19	-0.88	0.38
	(0.92)	(0.78)		(1.09)	(0.60)	
Property Charge	-1.91	-2.18	0.86	-0.76	-0.83	0.99
	(1.28)	(0.66)		(4.73)	(1.82)	
Drug Charge	1.90	1.95	0.97	5.50	4.77	0.84
	(0.81)	(0.77)		(3.23)	(1.50)	
Public Order Charge	-2.00	-3.47	0.09	-1.39	-5.37	0.41
	(0.77)	(0.38)		(4.63)	(1.52)	
Traffic Charge	5.36	7.17	0.11	14.98	13.82	0.70
	(1.04)	(0.72)		(2.19)	(2.16)	
Parole/Probation	-10.27	-13.54	0.13	-18.39	-18.06	0.94
	(1.85)	(0.97)		(4.00)	(1.89)	
Pretrial Release	1.15	4.25	0.04	-1.67	3.85	0.26
	(1.23)	(1.03)		(4.02)	(2.65)	
B. Private Information						
Male	-2.60	-1.98	0.50	-7.00	-3.61	0.43
	(0.84)	(0.37)		(3.70)	(2.42)	
White	0.42	1.03	0.49	-2.94	1.83	0.13
	(0.71)	(0.53)		(2.71)	(1.64)	
Homeless	-11.96	-13.78	0.54	-7.98	-11.50	0.50
	(2.30)	(1.86)		(4.44)	(2.71)	
No Telephone	-8.63	-9.18	0.77	3.45	-10.67	0.00
	(1.74)	(0.88)		(4.05)	(2.56)	
Out-of-State Address	-1.96	-8.04	0.00	-10.81	-16.25	0.77
	(1.33)	(1.53)		(17.10)	(6.60)	
Violent Charge Against an Adult	-1.60	-1.54	0.95	-15.19	-2.63	0.00
	(0.79)	(0.65)		(3.50)	(2.40)	
Violent Charge Against a Child	0.50	1.52	0.54	19.07	3.65	0.05
	(1.33)	(1.04)		(6.30)	(4.79)	
Any Aggravating Condition	1.54	3.49	0.07	13.14	-26.81	0.14
	(0.89)	(0.60)		(23.00)	(13.81)	
Override Recommendation	-28.75	-27.73	0.64	36.83	24.56	0.03
	(1.83)	(1.15)		(4.64)	(2.99)	
Risk Score FE	Yes	Yes		Yes	Yes	
Cases	6,665	25,327		1,244	4,619	

Notes. This table reports OLS estimates of regressions of an indicator for release on case and defendant characteristics with NCA risk score fixed effects. Columns 1 and 2 report results for cases where the algorithm recommends release. Columns 4 and 5 report results for cases where the algorithm recommends detain. Columns 1 and 4 report results for high-skill judges, columns 2 and 5 report results for low-skill judges, and columns 3 and 6 report the p-value on the difference. Standard errors clustered by judge are reported in parentheses.

Table A.7: Decomposing the Judges' Performance

	All Judges		High-Skill Judges		Low-Skill Judges	
	(1)	(2)	(3)	(4)	(5)	(6)
Judge Misconduct vs. Algorithm	2.42	2.42	-1.54	-1.54	3.47	3.47
	(0.48)	(0.48)	(0.68)	(0.68)	(0.51)	(0.51)
Predictable Performance Differences	0.54	0.74	0.20	0.13	0.72	0.95
	(0.56)	(0.58)	(0.56)	(0.57)	(0.62)	(0.59)
Non-Predictable Performance Differences	1.88	1.68	-1.74	-1.67	2.75	2.52
	(0.48)	(0.48)	(0.66)	(0.67)	(0.51)	(0.50)
Observable Information	Yes	Yes	Yes	Yes	Yes	Yes
Private Information	No	Yes	No	Yes	No	Yes
Judges	62	62	19	19	43	43

Notes. This table reports estimates decomposing the impact of human discretion on conditional misconduct rates into the share that is predictable based on observable inputs versus non-predictable based on observable inputs. Columns 1-2 report results for all judges, columns 3-4 for high-skill judges with lower conditional misconduct rates than the algorithm holding fixed release rates, and columns 5-6 for low-skill judges with higher conditional misconduct rates than the algorithm holding fixed release rates. The first row reports the average difference between the conditional misconduct rate for each judge and the counterfactual misconduct rates under the algorithm at the same release rate, the second row reports the share of this performance difference that is predictable from the included factors, and the third row reports the share of this performance difference that is not predictable from the included factors. Conditional misconduct rates under the algorithm are estimated using linear extrapolations of mean risk at different risk score cutoffs. All estimates adjust for shift-by-time fixed effects. Standard errors are estimated by taking random draws from the distributions of the estimated judge-specific release rates and conditional misconduct rates and then recalculating the threshold-specific ATE extrapolations and statistics of interest. See the text for additional details.

Appendix B: Judge Survey

This appendix summarizes the most important details of the judge survey. We begin by describing how we reached out to the judges and conducted the survey. We then provide summary statistics for our survey sample and a full list of the questions.

Outreach. Between May 2021 and January 2022, we worked with the jurisdiction to email the active full-time bail judges to participate in our survey. We successfully contacted 40 of the 62 judges in our sample, with 33 consenting to the survey and 28 completing the relevant questions described below. The judges who completed the relevant survey questions are not disproportionately high- or low-skilled based on our classification described in the main text, with the coefficient in a bivariate regression of indicator for high-skill on an indicator for survey response at 0.16 (SE: 0.12) with an $R^2 = 0.03$.

Survey Questions. Following a series of short introduction materials, the judges were provided with the following instructions:

> *We are interested in learning what factors you use to make pretrial decisions. Below is a list of 19 different factors, including factors commonly considered by judges across the United States. For each factor, please tell us on a scale of 1–5, where 1 is not important and 5 is very important, how important each factor is for you in making a pretrial decision. Specifically, please tell us how important each factor is in the decision of whether to impose financial conditions of release (i.e., cash bail or bond).*

We considered a list of 19 different factors, including factors commonly considered by judges across the United States. These 19 factors were chosen on the basis of extensive focus groups with bail judges that we conducted in 2020 and 2021. Subsection B.1 provides the complete list of questions.

Summary Statistics. We create indicator variables for giving above median importance to each factor given the variation in reported importance (e.g., most judges ranked the importance of race as either 1 or 2). We also construct mutually exclusive indices for observable information, private demographic information, and private non-demographic information by taking a simple mean of each of the indicators in the relevant index. Summary statistics are presented in Table B.1.

Table B.1: Summary Statistics of Judge Characteristics and Survey Responses

	All Judges	Low-Skill Judges	High-Skill Judges
	(1)	(2)	(3)
Observable Information Index	0.69	0.67	0.71
Age	0.61	0.65	0.55
Felony Offense	0.64	0.53	0.82
Violent Offense	0.79	0.76	0.82
Prior Misdemeanor Conviction	0.75	0.76	0.73
Prior Felony Conviction	0.75	0.71	0.82
Prior Violent Conviction	0.50	0.53	0.45
Prior FTA	0.75	0.71	0.82
Prior NCA	0.68	0.71	0.64
Prior NVCA	0.71	0.71	0.73
Demographic Information Index	0.16	0.24	0.05
Gender	0.25	0.35	0.09
Race	0.07	0.12	0.00
Private Information Index	0.62	0.55	0.72
Marital Status	0.68	0.59	0.82
Parental Status	0.64	0.65	0.64
Employment Status	0.71	0.76	0.64
Mental Health Condition and History	0.39	0.24	0.64
Substance Abuse Diagnosis and History	0.71	0.59	0.91
Financial Resources	0.50	0.35	0.73
Community Ties	0.54	0.53	0.55
Availability of Pretrial Services	0.75	0.71	0.82
Judges	28	17	11

Notes. The table reports descriptive statistics for the judges in our survey sample. Column 1 reports statistics for all judges in our survey sample, column 2 for low-skill judges in our survey sample, and column 3 for high-skill judges in our survey sample. The individual factors are indicator variables for placing greater than median weight on the factor, while the indices are simple means of these indicator variables. See the text for more details on the construction of these variables.

B.1 Survey Questions

Q1. How important is the defendant's age at current arrest for you in making a pretrial decision?

Q2. How important is whether or not the defendant is charged with a felony offense for you in making a pretrial decision?

Q3. How important is whether or not the defendant is charged with a violent offense for you in making a pretrial decision?

Q4. How important is a prior misdemeanor conviction for you in making a pretrial decision?

Q5. How important is a prior felony conviction for you in making a pretrial decision?

Q6. How important is a prior violent conviction for you in making a pretrial decision?

Q7. How important is prior failure to appear for you in making a pretrial decision?

Q8. How important is prior re-arrest for any new crime while out on release for you in making a pretrial decision?

Q9. How important is prior re-arrest for any new violent crime while out on release for you in making a pretrial decision?

Q10. How important is the defendant's gender for you in making a pretrial decision?

Q11. How important is the defendant's race for you in making a pretrial decision?

Q12. How important is the defendant's marital status for you in making a pretrial decision?

Q13. How important is the defendant's parental status for you in making a pretrial decision?

Q14. How important is the defendant's employment status for you in making a pretrial decision?

Q15. How important is the defendant's mental health condition and history for you in making a pretrial decision?

Q16. How important is the defendant's substance abuse (drug and alcohol) diagnosis and history for you in making a pretrial decision?

Q17. How important are the defendant's financial resources for you in making a pretrial decision?

Q18. How important are the defendant's community ties for you in making a pretrial decision?

Q19. How important is the availability of pretrial services for you in making a pretrial decision?