Algorithmic Recommendations and Human Discretion*

Victoria Angelova, Harvard University
Will Dobbie, Harvard University
Crystal S. Yang, Harvard University

March 2025

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.

*We thank Alex Albright, David Arnold, Ian Ayres, David Chan, Mitch Hoffman, Peter Hull, Michael Luca, Jack Mountjoy, Sendhil Mullainathan, Manish Raghavan, Ashesh Rambachan, 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.

I Introduction

Human decisions are often mistaken, noisy, and biased. Hiring managers often fail to identify the most productive candidates; emergency room physicians often fail to identify the patients with the highest expected risk of having a heart attack; and bail judges often fail to identify the defendants with the highest expected risk of engaging in pretrial misconduct. These costly mistakes have contributed to the rapid adoption of predictive algorithms in a range of high-stakes settings where clearly defined and socially important decisions hinge on individualized predictions.

Yet these same settings still require that a human decision-maker 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. Hiring managers might interview seemingly less productive candidates with excellent soft skills; emergency room physicians might order additional testing for seemingly low-risk patients with nausea; and bail judges might release seemingly high-risk defendants with mental health issues that can be addressed with mandated treatment. The key open question is whether allowing for such human discretion over an algorithm can yield more accurate decisions than an algorithm working alone.

Estimating the impact of human discretion is complicated by an important selection challenge. The challenge is that we generally only observe outcomes for individuals endogenously selected for treatment by the human decision-maker. Productivity and turnover are only observed among the selected subset of candidates hired by firms; heart blockages are only observed among the selected subset of patients who received additional testing by hospitals; and pretrial misconduct outcomes are only observed among the selected subset of defendants released by bail judges. As a result, we cannot directly measure counterfactual outcomes under the algorithm unless the human decision-maker and algorithm happen to select the same individuals for treatment.

This paper develops new quasi-experimental tools to solve this selection challenge and 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. To help guide their decisions, bail judges are often given an algorithmic risk assessment that predicts the likelihood of misconduct and recommends whether to release or detain the defendant. Like many other high-stakes settings, the bail judges frequently override these algorithmic recommendations, despite influential work showing that such algorithms can outperform bail judges working alone (Kleinberg et al., 2018).

In the first part of the paper, we measure the impact of human discretion over an algorithm on the accuracy of decisions by leveraging 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 estimate counterfactual outcomes under the algorithm 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, thereby solving the selection problem introduced by the bail judges detaining some defendants that the algorithm would have released. 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.

The second part of the paper uses our quasi-experimental approach to measure the impact of human discretion over an algorithm in a large, mid-Atlantic city. We find that the judges in our setting substantially underperform the algorithm when they make a discretionary override, increasing pretrial misconduct by an average of 15% at the judges' existing release rates. The negative average impact of human discretion 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. But we also find that 10% of the judges outperform the algorithm when they make a discretionary override, suggesting that a human and algorithm working together can potentially outperform automated release decisions. These high-performing judges have similar demographics, political affiliations, and years of experience as the low-performing judges. The only notable difference is that the high-performing judges are less likely to have previously worked in law enforcement compared to the low-performing judges.

The final part of the paper explores what interventions could improve the judges' performance by studying the mechanisms underlying these performance differences. We show that the high- and low-performing judges primarily differ in how they use private information that is not available to the algorithm. The low-performing judges meaningfully underperform an alternative algorithm that we construct, which predicts release decisions (not pretrial misconduct) using the observable information that is available to both the judges and the original algorithm. This finding indicates that the low-performing judges add noise and inconsistency to their decisions when they attempt to use private information that is not available to the algorithm. The added noise and inconsistency come from a variety of channels, including the consistent overweighting of factors that are not predictive of misconduct risk; the mistaken release of some high-risk defendants and mistaken detention of some low-risk defendants when setting monetary bail; and the overreaction to highly salient but largely uninformative events, such as hearing a case where a different defendant is arrested for a serious violent offense. As a result, the low-performing judges are far worse at filtering out the noise in private information relative to the high-performing judges, leading to

a deterioration in the low-performing judges' performance relative to both the high-performing judges and the algorithm. By comparison, we find no support for the idea that the high- and low-performing judges consider different objectives or differ in their use of observable information.

We supplement these findings with an original survey that asked the judges in our sample to rank the importance of different factors when making bail decisions, yielding new insights into the types of private information that add valuable signal versus noise. The survey reveals striking differences in the importance attached to the private information that is not available to the algorithm. The low-performing judges place far greater importance on demographic factors such as race, while the high-performing judges place far greater importance on non-demographic factors such as mental health, substance abuse, and financial resources. These findings suggest that we may be able to improve the accuracy of release decisions by teaching the judges to focus only on the most relevant pieces of private information not included in the algorithm.

Our results are an important proof of concept that the most skilled 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 the impact of human oversight policies will depend on the predictive abilities of the human decision-makers. As a result, there will not necessarily be a single correct human oversight policy. 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. An important question for future work is how to improve the predictive abilities of human decision-makers by learning from the most skilled human decision-makers and, when that is not possible, how to constrain or guide the least skilled decision-makers.[1]

Our paper complements an important literature showing that predictive algorithms theoretically outperform human decision-makers working alone (e.g., Berk, 2017; Jung et al., 2017; Mullainathan and Obermeyer, 2022). Kleinberg et al. (2018) is a seminal paper in this area, using the quasi-random assignment of judges and bounding techniques to show that predictive algorithms can substantially outperform bail judges working alone. Yet recent work shows that judges frequently override the recommendations generated by such predictive algorithms in real-world settings (e.g., Stevenson, 2018; Albright, 2023; Stevenson and Doleac, Forthcoming). We connect these two streams of work by developing new tools to measure the impact of these overrides on social welfare. We show that it is possible to identify counterfactual outcomes under the algorithm using the quasi-random assignment of decision-makers to individuals. Leveraging these new

[1]For example, recent work by Satopää et al. (2021) shows that we can improve forecasting performance in government-funded tournaments by reducing noise versus increasing information or reducing bias and Sun et al. (2022) show that incorporating predictions of human decisions into an algorithm can lead to lower override rates and increased accuracy among warehouse workers.

quasi-experimental tools, we find that the theoretical advantages of predictive algorithms documented in prior work are largely undermined by allowing human decision-makers to oversee these algorithms and make the final decision. These findings highlight the importance of understanding how humans use algorithms in practice, an essential component to designing optimal algorithmic systems where humans are the final decision-makers (Ludwig and Mullainathan, 2021).

Our paper also adds to a burgeoning literature studying the impact of human discretion over algorithmic or evidence-based guidelines in various contexts. This literature generally finds that human discretion reduces the accuracy of decisions. For example, Hoffman, Kahn, and Li (2018) find that hiring managers with high override rates end up with worse overall hires. 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, while Kesavan and Kushwaha (2020) find that automating decisions at an automobile replacement parts retailer increased firm profitability. Most recently, Agarwal et al. (2023) find that providing algorithmic predictions to professional radiologists only increases the probability of making a correct decision when contextual information is also included. We contribute to this literature by developing new tools to measure the impact of human discretion and by providing additional evidence that human discretion over an algorithm reduces the accuracy of decisions on average. We add important nuance to this finding, however, by showing that the most skilled human decision-makers can still outperform an algorithm by distinguishing between valuable private information and noise.

The 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 outcome of interest is both known and well-measured among the subset of treated individuals. Such assignment variation is already widely used in economics, suggesting our new tools may be applied in many other settings where predictive algorithms are currently deployed, such as in hiring decisions (Hoffman, Kahn, and Li, 2018), lending decisions (Dobbie et al., 2021), medical diagnoses (Chan, Gentzkow, and Yu, 2022), and foster care placement decisions (Baron et al., 2024).[2]

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.

[2]In contrast, it is more difficult to measure the impact of human discretion in settings when the outcome of interest is not well-measured or there are multiple objectives. For example, college admissions officers often balance goals of diversity, academic achievement, and tuition revenue that make comparisons to algorithmic predictions difficult.

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.

By law, each decision-maker is meant to align $D_{i,j}$ with Y_i^*. 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 $Y_i^* = 1$). In practice, however, judges may differ in their predictions of which individuals fall into which category or may consider a different set of objectives altogether. 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 predicts 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 her private information $\mathbf{V}_{i,j}$. The judge uses this information set to form her subjective assessment of each individual's appropriateness for release, $h_{i,j}$, which can be a prediction of Y_i^* if the judge follows the legal objective, but can also be an assessment of appropriateness for release based on other factors if the judge has other objectives. The judge's

subjective assessment for each individual i comes from the mapping $h_j(\cdot) : \bar{a} \times \bar{D} \times \mathscr{X} \times \mathscr{V} \to [0,1]$, where $\bar{a} = [0,1]$ and $\bar{D} \in \{0,1\}$. We allow the function $h_j(\cdot)$ to vary by j to allow for the possibility that each judge places different weights on each piece of information. We simply assume that each judge releases individuals in order of her subjective assessment, implying that the judge's decision rule can be represented by a threshold τ_j with $D_{i,j} = \mathbf{1}[h_{i,j} \leq \tau_j]$, where τ_j can be interpreted as judge j's preference for release.[3] 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 rankings for some or all individuals. This issue has complicated efforts to measure the relative performance 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 challenge and to jointly identify predictive performance 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 the judge's relative level of performance at the 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]The release threshold τ_j can be microfounded with a simple model where the judge weighs a perceived cost of release relative to a perceived cost of detention. For example, suppose that the judge's utility function can be represented as $\pi_j(Y^*, D) = -a_j Y^* D - c_j(1 - Y^*)(1 - D)$, where a_j represents the judge's perceived cost of releasing an individual who commits misconduct (type II error), c_j represents the judge's perceived cost of detaining an individual who does not commit misconduct (type I error), and a_j can differ from c_j. This utility function results in a decision rule of $D_{i,j} = \mathbf{1}[h_{i,j} \leq \tau_j]$ where $\tau_j = \frac{c_j}{a_j + c_j}$. Intuitively, as the relative perceived cost of a type II versus type I error increases, judges lower their release threshold τ_j. In such a model, comparing each judge's observed misconduct rate to the counterfactual misconduct rate under the algorithm at the same release rate is equivalent to holding fixed the ratio of perceived type II and type I errors.

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 rankings for some or all individuals.

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.

There are two main reasons why the judge's subjective ranking of individuals may differ from the algorithm's ranking such that $\Delta M_{j,s(j)} \neq 0$. The first possibility is that the judge may consider other objectives that are not permitted under the law, from exhibiting mercy to younger defendants to explicitly discriminating against defendants based on their demographic characteristics or criminal charge. The judges could also care about certain types of pretrial misconduct, such as being rearrested for a violent felony, more than other types of misconduct, such as failing to appear at court or being arrested for a misdemeanor. For example, some judges report to our research team that, "Any crime is a wide variety of crimes. I am willing to risk some misdemeanors, but not a serious felony." In contrast, the algorithm is designed to minimize overall pretrial misconduct across all defendant subgroups.[4]

The second possibility is that the judge shares the same objective as the algorithm, but disagrees with the algorithm's misconduct predictions and rankings for some or all individuals. The judge may, for example, systematically over- or underweight the observable information $\mathbf{X}_i$ that is available to both the judge and the algorithm. The effect of such over- or underweighting on accuracy is theoretically ambiguous as many existing pretrial algorithms (including the one we

[4]In theory, predictive algorithms can be designed to prioritize different forms of misconduct or different subgroups of defendants. For example, predictive algorithms can be designed to minimize the social cost of pretrial misconduct or to minimize unwarranted racial disparities in release rates among defendants with similar misconduct potential. We take the current algorithmic design as given and measure how human discretion over the algorithm impacts the accuracy of decisions. We return to this issue in Section V when considering the mechanisms underlying our results.

study) are deliberately simple and may only roughly approximate $E[Y_i^* \mid \mathbf{X}_i]$. The judge may also use private information $\mathbf{V}_{i,j}$ that is not available to the algorithm, such as aggravating risk factors or the defendant's appearance. 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.

II.B Empirical Challenges

Estimating the impact of human discretion over an algorithm on the accuracy of decisions is complicated by an important selection challenge, as we only observe misconduct outcomes among the selected subset of defendants that a judge chooses to release before trial.

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, individuals who are detained by the judge ($D_{i,j} = 0$) cannot engage in misconduct, and so $Y_i = 0$ for these individuals regardless of true 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 $\mathbf{V}_{i,j}$. 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 the true misconduct potential Y_i^* among the defendants whom the judge chooses to release before trial. However, we are unable to directly measure the misconduct rate under the algorithmic counterfactual, $M_{s(j)}$, because there are generally some defendants whom the algorithm would release ($D_{i,s(j)} = 1$) but the judge does not ($D_{i,j} = 0$).

Consider a simple comparison of judge j's misconduct rate to the misconduct rate of the algorithmic counterfactual at risk score threshold $s(j)$ using the observed misconduct outcomes among released defendants:

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

Equation (4) shows that a simple comparison based on the observed misconduct outcomes among released defendants will yield biased estimates of $M_{s(j)}$ and, as a result, biased estimates of $\Delta M_{j,s(j)}$ unless judge release decisions are uncorrelated with true misconduct potential among the relevant set of cases, such that $E[Y_i^* | a(\mathbf{X}_i) \leq s(j)] = E[Y_i^* | D_i = 1, a(\mathbf{X}_i) \leq s(j)]$. Such a scenario is unlikely

in practice, motivating our empirical exercise.

II.C 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 a hypothetical algorithm at each possible release rate, capturing the observed misconduct rate among released defendants at each release rate if individuals were released in order of their algorithmic risk score $a(\mathbf{X}_i)$. The dashed vertical line at R_s denotes the release preference of the algorithmic designer at risk score threshold s, which corresponds to the threshold where the algorithm changes its recommendation from release to detain. We also plot a hypothetical release rate and conditional misconduct rate for a hypothetical judge j under different scenarios, highlighting the logic underlying our approach. Panel A depicts a scenario where judge j has a lower release threshold than the algorithmic designers ($R_j < R_s$). However, the judge still follows the algorithm's ranking of individuals and overrides the algorithm solely because of differences in release preferences. As a result, her conditional misconduct rate is equal to the algorithm's conditional misconduct rate at her release rate such that $\Delta M_{j,s(j)} = 0$.

Panel B illustrates an alternative scenario where judge j has a higher release threshold than the algorithmic designers ($R_j > R_s$). In this scenario, judge j overrides the algorithm because she has a different ranking of individuals compared to the algorithm, which may be a result of a different set of objectives, or the different use of observable information or additional private information even with the same objective as the algorithm. In this particular case, she makes more accurate predictions than the algorithm and achieves a lower conditional misconduct rate such that $\Delta M_{j,s(j)} < 0$. The same underlying logic applies to a scenario where the judge makes less accurate predictions than the algorithm and achieves a higher conditional misconduct rate. This logic applies even if the judge has the exact same release threshold as the algorithmic designer because the judge may have a different ranking of individuals compared to the algorithm.

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. The empirical challenge is that we do not observe the counterfactual misconduct rate under the algorithm, $M_{s(j)}$, and therefore cannot directly measure $\Delta M_{j,s(j)}$.

III Setting and Data

III.A Our Setting

We study the impact of human discretion on the accuracy of pretrial release decisions in 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 the risk of 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 only case and defendant characteristics they deem relevant to minimizing the risk of 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). Bail judges 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.

To help guide their decisions, bail judges in our setting are provided with an algorithmic risk assessment that predicts the likelihood of misconduct and recommends whether to release or detain the defendant. The judges are provided with separate FTA and NCA risk scores, which are both based on the defendant's 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 case and demographic characteristics included as algorithmic inputs in the FTA and NCA risk scores depend on whether there was a statistically significant association with the relevant misconduct outcome in the data used to build each 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 summed to yield a detailed risk score for each misconduct outcome. These detailed risk scores are then binned into aggregate risk scores that range from 1 to 6, with lower risk scores indicating a lower probability of misconduct and higher risk scores indicating a higher probability of misconduct. While judges are only shown these aggregate risk scores, they are all trained on the risk scoring system and are provided with the complete set of factors used to calculate the more detailed scores.

The FTA and NCA risk scores are highly predictive of pretrial misconduct among released defendants and closely track the performance of the more sophisticated gradient-boosted decision tree algorithm developed by Kleinberg et al. (2018), at least for most defendants (Appendix Figure A.1). The fact that the FTA and NCA risk scores closely track the gradient-boosted decision tree algorithm suggests that our results are likely externally valid in pretrial settings deploying more

sophisticated algorithms, a point we return to in Section IV.

The algorithmic recommendation is automatically generated using a combination of the FTA and NCA risk scores (Appendix Table A.1). The lowest combinations of risk scores yield a recommendation of release with no conditions, with these cases making up approximately 25% of the cases in our data. More moderate combinations of risk scores yield a recommendation of release with regular phone or in-person check-ins, with these still 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 recommendation to detain. Release decisions are also much more responsive to the NCA score in practice, with sharp changes in release rates when the NCA score changes from 5 to 6 (Appendix Figure A.2). We also show results comparing the judges to the FTA risk score and a linear combination of the FTA and NCA scores.

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 recommendation ($D_{i,s}$ in our model) and aggregate FTA and NCA scores ($a(\mathbf{X}_i)$) discussed above. The report also includes information on the defendant's arrest date and date of birth and all of the observable risk factors used by the algorithm ($\mathbf{X}_i$), such as the defendant's age at first arrest and number of prior arrests. The report also includes private information that is not used by the algorithm ($\mathbf{V}_{i,j}$), such as the defendant's race and gender, a detailed description of the charges, the defendant's residence, and a telephone number if available. Additional private information comes from the pretrial services officer, who can recommend an override of the algorithmic recommendation in specific situations with supervisor approval when there are aggravating factors. These override recommendations occur in about 8% of the cases in our sample, with nearly all of the recommendations being to detain defendants that the algorithm recommends releasing.

Local rules do not bind judges to the algorithmic recommendation and state that the algorithm "must not be the only means of reaching the bail determination." Given the scope for human discretion, some judges in our setting publicly state that when setting conditions for release, they would "take into account the county risk assessment tool," while others state that they do not believe in "binding judges to...risk assessment algorithms." In practice, 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%. Yet we also see that the judges 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 (Appendix Figure A.3).

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 (e.g., Albright, 2023; Stevenson and Doleac, Forthcoming).

We exploit four additional features of the pretrial system in our setting for our analysis. First, the bail judges generally make their decisions based on the pretrial risk assessment report (Figure 2) before speaking with parties at the bail hearing. The bail judges typically announce these decisions at the start of the bail hearing, where the defendant appears via videoconference from the local jail. The bail judge generally does not ask any questions of the defendant and the hearings last less than five minutes on average. The prosecutor is almost never present at the bail hearing, and public defenders provide only 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.

Second, the legal objective of bail judges in our setting is both narrow and measurable. Judges are directed to release the vast majority of individuals while minimizing the risk of pretrial misconduct. This 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. This legal objective is embodied in various principles of pretrial release, such as the American Bar Association's Pretrial Release Standard, which states that "In deciding pretrial release, the judicial officer should assign the least restrictive condition(s) of release that will reasonably ensure a defendant's attendance at court proceedings and protect the community, victims, witnesses or any other person" (American Bar Association, 1981). Bail judges in our context also reported to our research team in focus groups that they focus on how their decisions impact appearance and re-arrest rates among released defendants. They also view the trade-off between pretrial release and pretrial misconduct as "very relevant," and that it "gets at the real issue, public safety."

Third, we follow the prior literature in treating bail judges as making binary decisions: releasing low-risk defendants (generally by release on recognizance (ROR), non-monetary conditions, or setting a low monetary bail amount) and detaining high-risk defendants (generally by setting a high monetary bail amount or outright detaining them).[5] This binary classification is particularly

[5]In our setting, judges decide whether to release on recognizance, release with non-monetary conditions such as a requirement to report to pretrial services, impose monetary bail, or detain defendants. 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.

well suited to our setting, as the judges in this jurisdiction 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, 2023), 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 how the use of monetary bail may help to explain 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, 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 full-time judges, 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 following booking in the jurisdiction's main jail between October 16, 2016, and March 16, 2020, corresponding to when the jurisdiction adopted the recent iteration of the algorithm and when it 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. The data also contain information on all of the factors used to calculate the FTA and NCA risk scores, the 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 monetary bail. We categorize defendants as either released (including ROR, release with non-monetary conditions, and release conditional on paying monetary bail) or detained (including non-payment of monetary 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.[6] 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 (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 (dropping 2,144 cases). Finally, we omit observations where the case is assigned to a judge with fewer than 100 cases 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. Judges override the algorithm's recommendation in 54% of observably high-risk cases and 12% of observably low-risk cases. The lenient overrides among observably high-risk cases and harsh overrides among observably low-risk cases are also not random. For example, defendants receiving a lenient override have fewer prior arrests and convictions (both felony and misdemeanor), are less likely to be on parole or probation, and are more likely to be charged with drug or traffic offenses. Defendants receiving a lenient override are also less likely to be male, homeless, and missing a telephone number, 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 number, more likely to reside out of state, and more likely to have an aggravating condition and a pretrial services override recommendation.

[6]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. Following Arnold, Dobbie, and Yang (2018), Dobbie, Goldin, and Yang (2018), and Arnold, Dobbie, and Hull (2022), we include these miscoded cases in our estimation sample because there is no way to identify miscodings when defendants do not commit pretrial misconduct. Our results are also unchanged if we drop these miscoded cases instead.

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 misconduct rate at the same release rate, $M_{s(j)}$. Our approach, as described below, only requires that the threshold-specific average misconduct risk parameters used to construct the algorithmic counterfactual (i.e., $M_{s(j)}$ for all j) can be accurately extrapolated from the data. Our approach for estimating the impact of human discretion proceeds in three main 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 algorithmic release threshold. The key insight underlying our empirical strategy is that measuring the algorithmic counterfactual at a given judge's release rate is equivalent to estimating the average misconduct risk of defendants with risk scores at or below the relevant algorithmic release threshold $s(j)$: $M_{s(j)} = E[Y_i^* | a(\mathbf{X}_i) \leq s(j)]$. When judges are as-good-as-randomly assigned, the average misconduct risk at a given algorithmic release threshold is common to all judges. As a result, the algorithmic counterfactual at that release threshold is captured by the threshold-specific average misconduct risk.

The required threshold-specific misconduct risk parameters can be estimated from quasi-experimental variation in pretrial release and misconduct rates, building on the "identification at infinity" approach in Arnold, Dobbie, and Hull (2022). To build intuition for our empirical strategy, consider a setting with as-good-as-random judge assignment and a hypothetical bail judge j^* who releases all or nearly all of her defendants with risk scores at or below a given algorithmic release threshold $s(j)$ for $j \neq j^*$, regardless of their true potential for pretrial misconduct. This hypothetical judge may release all or nearly all defendants for two different reasons. The first reason is that she is supremely lenient and releases all or nearly all defendants in the full population, not just at or below the relevant algorithmic release threshold $s(j)$. The second reason is that she is supremely compliant with the algorithm at or below this particular release threshold so she releases all or nearly all defendants with risk scores at or below $s(j)$ but detains some defendants with higher risk scores. In either case, this hypothetical judge's release rate for defendants with risk scores at or below the release threshold $s(j)$ is close to 100%:

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

This means that the threshold-specific average misconduct rate among defendants released by the hypothetical judge approximates the average misconduct risk of defendants with risk scores at or

below the algorithmic release threshold $s(j)$, at least in large samples:

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

The first equality in both Equations (5) and (6) follows from the as-good-as-random judge assignment. In large enough samples, the decisions of such a hypothetical judge who is as-good-as-randomly assigned can be used to estimate the threshold-specific average misconduct risk parameters needed for our analysis without additional assumptions.

In the absence of such a hypothetical judge, the required threshold-specific average misconduct risk parameters can be reliably estimated using linear or local linear extrapolations of release and misconduct rate variation across as-good-as-randomly assigned judges. The validity of our extrapolation-based approach requires that we reliably estimate the relationship between the misconduct rates among released individuals and release rates among judges with release rates close to 100%.[7] Linear extrapolations can be justified by functional form assumptions about the average relationship between true misconduct potential Y_i^* and judges' subjective assessments of each individual's appropriateness for release, $h_{i,j}$, analogous to shape restrictions on MTE functions (Brinch, Mogstad, and Wiswall, 2017; Mogstad, Santos, and Torgovitsky, 2018). For example, a linear MTE model yields a linear relationship between conditional misconduct rates and release rates, such that linear extrapolation can reliably estimate the required misconduct risk parameters.[8] Local linear extrapolations can accommodate substantially more complex and non-linear decision-making models, with consistency following from observing a large number of instrument values

[7]Consider the population of defendants with risk scores at or below an arbitrary threshold s. Suppose without loss of generality that $h_{i,j}|\tau_j \sim U(0,1)$, where τ_j is judge j's release threshold for defendants with risk scores at or below s. Let the average misconduct rate (averaged over i) for defendants with ranking $h_{i,j} = h$ by a judge j with release threshold τ_j be $E[Y_i^*|\tau_j, h_{i,j} = h, a(\mathbf{X}_i) \leq s] = g(\tau_j, h|a(\mathbf{X}_i) \leq s)$. Given quasi-random case assignment, the average misconduct potential for all defendants assigned to each judge must equal the average misconduct rate, M_s, such that $E[Y_i^*|\tau_j, a(\mathbf{X}_i) \leq s] = \int_0^1 g(\tau_j, h|a(\mathbf{X}_i) \leq s)dh = M_s$ for all judges. Now, define the conditional misconduct rate for a judge j with release threshold τ_j as:

$$\begin{aligned} E[Y_i^*|D_{i,j} = 1, \tau_j, a(\mathbf{X}_i) \leq s] &= E[Y_i^*|h_{i,j} \leq \tau_j, a(\mathbf{X}_i) \leq s] \\ &= \frac{1}{\tau_j}\int_0^{\tau_j} g(\tau_j, h|a(\mathbf{X}_i) \leq s)dh = K(\tau_j|a(\mathbf{X}_i) \leq s), \end{aligned}$$

where by definition $K(1|a(\mathbf{X}_i) \leq s) = M_s$. Our extrapolation methodology requires us to reliably estimate this $K(\tau_j)$ function as the release threshold τ_j approaches 1.

[8]Formally, $g(\cdot)$: $E[Y_i^*|\tau_j, h_{i,j}, a(\mathbf{X}_i) \leq s] = M_s + (h_{i,j} - \frac{1}{2})$ is the most general functional form that yields a linear $K(\tau_j|a(\mathbf{X}_i) \leq s)$ function, corresponding to the linear MTE model in Brinch, Mogstad, and Wiswall (2017).

as the sample grows and with maximal choice probabilities approaching one (Hull, 2020). In our setting, the relatively large number of judges with large caseloads and high release rates at the relevant release thresholds helps make these local linear extrapolations reliable. For example, we observe 23 judges with a release rate of 85% or higher in the full sample and 46 judges with a release rate of 85% or higher in the sample where the algorithm recommends release.

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. This allows us to estimate $M_{s(j)}$ for each judge 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 algorithmic counterfactual at the same release rate $M_{s(j)}$.

Finally, we summarize 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. We note, however, that the judge-level estimates of $\Delta M_{j,s(j)}$ are likely to involve substantial sampling error, particularly for the judges with relatively few cases. As a result, the raw share of judges with observed misconduct rates that are higher and lower than the algorithmic counterfactual is unlikely to be a valid estimate of the judge's true 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.[9] We focus on the share of judges with misconduct rates that are higher and lower than the algorithmic counterfactual throughout the paper, as the judge-specific estimates are generally too imprecise to reliably 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 rates $\hat{R}_j$ and conditional misconduct rates $\hat{M}_j$ and then recalculate the threshold-specific extrapolations and statistics of interest. The Online Appendix provides additional details on how we account for the conditional random assignment of

[9]Estimation of the posterior average effects depends on two distributional assumptions. First, we assume that 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, $\widehat{\Delta M}_{j,s(j)} = \Delta M_{j,s(j)} + \varepsilon_j$, where $\varepsilon_j \sim N(0, \Sigma_j)$. Second, we assume that the unknown true judge-specific effects are drawn from a normal distribution with unknown hyperparameters, $\Delta M_{j,s(j)} \sim N(\overline{\Delta M}, \Lambda)$. 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 show below 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.

bail judges and the discrete algorithmic scores.

Our quasi-experimental approach to estimating the impact of human discretion on the accuracy of pretrial decisions 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). Our approach is most closely related to Arnold, Dobbie, and Hull (2022), who study racial disparities in pretrial release rates conditional on a defendant's potential for pretrial misconduct. Arnold, Dobbie, and Hull (2022) show that omitted variables bias in pretrial release rate comparisons can be purged by using the quasi-random assignment of judges to estimate average pretrial misconduct risk by race. Baron et al. (2024) show that a similar quasi-experimental approach can be used to study racial disparities in multiple stage systems such as the child welfare system. We build on this work by showing how the quasi-random assignment of judges can be used to estimate average misconduct risk at different algorithmic risk thresholds.

An important advantage of our extrapolation-based approach is that it can be justified without a conventional first-stage monotonicity assumption, which recent work has questioned in the context of quasi-random decision-makers (Chan, Gentzkow, and Yu, 2022; Frandsen, Lefgren, and Leslie, 2023). Like Arnold, Dobbie, and Hull (2022) and Baron et al. (2024), our quasi-experimental approach only requires that the average misconduct risk parameters can be accurately extrapolated from the data. The extrapolated average misconduct risk parameters are valid so long as the *average* relationship between conditional misconduct rates and release rates across judges can be reliably estimated (at least at high release rates).[10]

[10]We consider a simple extension of our conceptual framework described in Section II to show how the relationship between conditional misconduct rates and release rates across judges can be reliably estimated even when the conventional first-stage monotonicity assumption is violated. For a population of individuals i with risk scores at or below an algorithmic threshold s, let each judge's release decision be given by $D_{i,j} = \mathbf{1}[h_{i,j} \leq \tau_j | a(\mathbf{X}_i) \leq s]$, and denote by λ_j the judge's skill at predicting true misconduct. Suppose without loss of generality that $h_{i,j}|\tau_j, \lambda_j \sim U[0,1]$ and that $E[Y_i^*|\lambda_j, \tau_j, h_{i,j}, a(\mathbf{X}_i) \leq s] = M_s + \lambda_j(h_{i,j} - \frac{1}{2})$, where M_s represents the average misconduct risk in the population of defendants with risk scores at or below s. Importantly, this model allows for violations of conventional monotonicity since judges can differ both in their rankings of individuals by appropriateness for release, $h_{i,j}$, and in their relative skill at predicting misconduct, λ_j. Nevertheless, the relationship between conditional misconduct rates and release rates across judges can still be reliably estimated. For example, if $E[\lambda_j|\tau_j]$ is constant in τ_j (i.e., $E[\lambda_j|\tau_j] = a$), the average conditional misconduct rates, $E[Y_i^*|D_{i,j} = 1, \tau_j, a(\mathbf{X}_i) \leq s]$ are linear in judge release rates, $E[D_{i,j}|a(\mathbf{X}_i) \leq s] = \tau_j$. Specifically, $E[Y_i^*|D_{i,j} = 1, \tau_j, a(\mathbf{X}_i) \leq s] = M_s + \frac{1}{2}a(\tau_j - 1)$, such that linear extrapolation can estimate M_s. Non-parametric local linear extrapolations can accommodate a broader range of shape restrictions on $E[\lambda_j|\tau_j]$ under the assumptions discussed in the main text.

IV.B Counterfactual Misconduct Under the Algorithm

Figure 3 illustrates our extrapolation-based estimation of average misconduct risk at 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 nearly 85%. The horizontal axis in each panel plots estimated release rates for each of the 62 judges in our data ($\hat{R}_j$) among defendants with algorithmic risk scores below the relevant threshold. The estimated release rates are regression adjusted 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 ($\hat{M}_j$) among defendants with algorithmic risk scores below the relevant threshold, again adjusted for shift-by-time fixed effects but not for sampling error.

The vertical intercepts of the lines of best fit, at a release rate of 100%, provide estimates of the threshold-specific average misconduct rates. The lines of best fit are obtained from 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 simple linear extrapolation yields a mean misconduct estimate of 13.8% (SE: 0.8) for cases where the algorithm recommends release, which corresponds to a release rate of nearly 85%. 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 the release rates observed in our sample. The results from these extrapolations are plotted in Figure 4 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 in the Online Appendix. We take the linear

extrapolation as our baseline specification for estimating the impact of discretion on the accuracy of decisions and explore the robustness of our results to alternative mean risk estimates below.

IV.C Effects of Human Discretion on Conditional Misconduct Rates

Figure 4 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 estimates are adjusted for shift-by-time fixed effects but not for sampling error. The dashed orange line shows the 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), which 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 estimated judge conditional misconduct rates and release rates, even before we compare judge performance 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 fixed the release rate. Second, there is 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.[11] 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).

The comparison of judge performance 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 higher than the algorithmic counterfactual at the same re-

[11] These results provide evidence against the standard monotonicity assumption, which implies that judges whose objective is to minimize misconduct will have the same conditional misconduct rate at the same release rate (Chan, Gentzkow, and Yu, 2022; Frandsen, Lefgren, and Leslie, 2023).

lease rate. Using the posterior average effect approach to account for sampling error, we estimate that 90% (SE: 6.1) of judges 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 algorithmic counterfactual. These results indicate that we could substantially decrease misconduct rates by automating release decisions, as the typical judge in our setting is either less skilled at predicting misconduct or considers a different set of objectives than the algorithm.

Our estimates show that there are both statistically and economically significant costs of human discretion over an algorithm in our setting. They 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).

However, the negative system-wide impact of human 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 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 performance of the human decision-makers in a particular context.

We next ask whether we can identify high- versus low-performing 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-performing 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 higher. The second group consists of high-performing judges who are 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.[12]

[12]There is a consistent set of high- and low-performing judges that drive the variation in performance, suggesting that our measure captures true performance differences and not idiosyncratic noise. For example, we find that the correlation between the judges' performance estimates in

Table 2 shows results from OLS regressions of an indicator for being a high-performing 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, Table 2), ruling out the explanation that high-performing judges are simply more likely to follow the algorithmic recommendations compared to low-performing ones. In addition, there is no statistically significant relationship between judge performance and years of experience, gender, race, political affiliation, or whether the judge has a law degree or is a former prosecutor (columns 2–7, Table 2). However, one 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 a high-performing judge (column 8, Table 2). There is also evidence that high-performing judges generally outperform the algorithm in terms of both accuracy and racial fairness in release (column 9, Table 2).[13] However, none of the estimates from Table 2 suggests a clear explanation for the variation in judge performance.

IV.D Robustness and Extensions

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 consider a range of alternative specifications when estimating the distribution of judge performance relative to the algorithmic counterfactual, finding qualitatively similar results when using local linear extrapolations (Appendix Figure A.4a), our baseline linear extrapolations that omit shift-by-time fixed effects (Appendix Figure A.4b), and best- and worst-case step functions that connect the discrete risk scores (Appendix Figure A.4c). 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 lower and upper bounds for the remaining fraction of defendants (Appendix Figure A.4d). In this final specification where we only extrapolate to the most lenient judge, 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.

Sampling Error. Our baseline estimates use the posterior average effect approach of Bonhomme

the first and second half of our sample is 0.64 after accounting for statistical noise, indicating significant out-of-sample forecasting power even when using only half of the available data.

[13] We measure 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, following the procedure developed by Arnold, Dobbie, and Hull (2022).

and Weidner (2022) to account for sampling error in our judge-level estimates of $\Delta M_{j,s(j)}$. We obtain very similar estimates if we use the non-parametric empirical Bayes deconvolution approach of Efron (2016) to account for sampling error in our judge-level estimates. Here, we first estimate a population prior of the judges' conditional misconduct rates relative to the algorithmic counterfactual and then estimate posterior means of each $\Delta M_{j,s(j)}$ estimate (Appendix Figure A.5). Both our baseline estimates and the new non-parametric estimates are substantially larger than the unadjusted share of judges with higher misconduct than the algorithm, reflecting shrinkage due to noise in the observed data.

Algorithmic Comparison. Our baseline estimates compare judge performance to an algorithmic counterfactual based on the NCA risk scores that the judges see at the time of the bail decision. These NCA risk scores are generated using the simple scoring system as described in Section III. We obtain very similar estimates if we compare the judges to an algorithmic counterfactual based on the FTA risk scores that the judges also see at the time of the bail decision (Appendix Figure A.6a) or a linear combination of the FTA and NCA risk scores that, in theory, the judges could be using based on what they see at the time of the bail decision (Appendix Figure A.6b). We also find qualitatively similar results when we compare judge performance to a more sophisticated gradient-boosted decision tree algorithm (Appendix Figure A.6c), which speaks to the potential external validity of our results as risk scoring algorithms improve over time. The more sophisticated gradient-boosted decision tree algorithm is more accurate than the NCA and FTA risk scores, with misconduct rates decreasing by an average of 0.4 percentage points across the observed release thresholds. The share of judges underperforming the more sophisticated gradient-boosted decision tree algorithm is correspondingly higher than the share underperforming the proprietary algorithm, increasing to 93% (SE: 6.2). These results suggest there still remains a small share of judges who can potentially outperform even a sophisticated machine learning algorithm, indicating that our results are likely externally valid to pretrial settings deploying substantially more sophisticated algorithms either now or in the near future.

V Potential Mechanisms

This section explores the types of interventions that could improve judge performance by studying the mechanisms underlying the performance differences we observe. We start by showing that our results are unlikely to be explained by the judges having different objective functions across different defendant subgroups or different preferences over types of misconduct outcomes. We then show that the judges primarily differ in how they use private information that is not available to the algorithm, as opposed to how they use the observable information that is available to both the judges and the algorithm.

V.A Objective Function

One potential explanation for our results is that the high- and low-performing judges consider different objectives, with the high-performing judges focusing on pretrial misconduct and the low-performing judges prioritizing other objectives. For example, the high-performing judges may only consider the legal objective of minimizing pretrial misconduct while the low-performing judges may consider extra-legal objectives such as exhibiting mercy to younger defendants, or even explicitly discriminating against certain types of defendants. Another possibility is that the high-performing judges consider all types of pretrial misconduct while the low-performing judges consider only certain types of misconduct such as being rearrested for a new crime.

We explore the possibility that the high- and low-performing judges consider different objectives in two ways. First, we measure judge performance separately for different defendant and case subgroups. Figure 5 shows the fraction of judges who underperform the algorithm both overall and for different defendant subgroups defined by race, age, education, and felony charge. We find that the fraction of low-performing judges is similar across different defendant and case subgroups, with at least 74% (SE: 6.4) of judges underperforming the algorithm in each subgroup. We also find a strong positive relationship between posterior estimates of our baseline measure of judge-specific performance and performance in each defendant and case subgroup for each of the 62 judges in our main sample (Appendix Figure A.7). These findings indicate that our main results are unlikely to be explained by different objective functions for high- and low-performing judges or different preferences for release across different defendant subgroups.

Second, we measure judge performance separately for different pretrial misconduct outcomes. Figure 5 shows the fraction of judges who underperform the algorithm both overall and for only FTA, only NCA, only violent NCA, and for a modified misconduct measure that captures the social cost of different types of misconduct.[14] We find that the fraction of low-performing judges is also similar across these different misconduct outcomes, with at least 76% (SE: 6.5) of judges under-

[14]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 outcomes, we use the lowest social cost estimate available (e.g., we use the cost estimate for assault instead of murder for persons offenses). Specifically, we use $1,278 for FTA, $83,743 for DUI, $25,351 for person offenses, $10,590 for motor vehicle offenses, $10,147 for drug offenses, $3,725 for weapons offenses, $3,091 for property offenses, $1,819 for public order offenses, and $501 for all other non-traffic criminal offenses. 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 weight the incidence of each misconduct type by these social costs to calculate the average social cost of misconduct. We then perform our analysis using this new social cost measure as the outcome variable.

performing the algorithm in each specification. Perhaps most importantly, we find that 89% (SE: 6.7) of judges underperform the algorithm when we use a misconduct measure that captures the social cost of different types of misconduct outcomes, remarkably similar to the 90% (SE: 6.1) of judges underperforming the algorithm in our baseline specification. We also find a strong positive correlation between posteriors estimates of our baseline measure of judge-specific performance and performance for each type of misconduct outcome (Appendix Figure A.7). These findings indicate that our main results are also unlikely to be explained by different preferences for different misconduct outcomes for high- and low-performing judges, suggesting that the performance differences are instead driven by how judges predict pretrial misconduct risk.

V.B Observable and Private Information

The second potential explanation for our results is that the high- and low-performing judges share the same objective as the algorithm – to minimize pretrial misconduct – but have different disagreements with the algorithm's risk predictions and rankings for some or all individuals. Table 2 shows that, on average, the high- and low-performing judges are equally likely to override the algorithm, meaning that any differences in judge performance must be driven by the types of defendants for which the high- and low-performing judges choose to override the algorithm. The high- and low-performing judges may, for example, override different types of defendants based on differential use of observable information $\mathbf{X}_i$ available to both judges and the algorithm or the use of private information $\mathbf{V}_{i,j}$ only available to the judges.

We explore the relative importance of observable and private information by building an alternative algorithm using a gradient-boosted decision tree that predicts the judges' release decisions (not pretrial misconduct) using all the observable information, following Kleinberg et al. (2018). We use these predicted release decisions to construct counterfactual decision rules for the high- and low-performing judges, where we order defendants by 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 judge release decisions using the extrapolation approach described above, again separately for high- and low-performing judges. These extrapolations allow us to construct the misconduct rate associated with the predicted release rules for high- and low-performing judges, holding fixed each judge's release rate. The results from this exercise tell us how the high- and low-performing judges would have performed if they had simply followed their own release tendencies using only 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 performance based on predicted release decisions to shed light on the importance of private information, which presumably drives most of the deviations

from the predicted release rules.

Figure 6a presents the conditional misconduct rate under the predicted judge release rules for high- and low-performing judges. The solid red line shows the estimated conditional misconduct rate under the predicted release decisions for high-performing judges, while the solid blue line shows the estimated conditional misconduct rate under the predicted release decisions for low-performing judges. The dashed orange line shows the estimated conditional misconduct rate under the original algorithm for comparison. The estimated conditional misconduct rate under both the high- and low-performing judges' predicted release rules is modestly higher than under the original algorithm, suggesting that both types of judges make modest but predictable errors in their use of observable information that lead 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 indicate that the high- and low-performing judges use the observable information available to both them and the algorithm in a very similar way.[15] Figure 6a also shows that the high-performing 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-performing 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 private information.

The above results suggest that high- and low-performing judges primarily differ in how they use private information that is not available to the algorithm $\mathbf{V}_{i,j}$. Three sets of results more directly support this explanation and show how private information introduces noise and inconsistency for low-performing judges.

The first way private information adds noise and inconsistency for low-performing judges is the over- or underweighting of information contained in the detailed risk assessment reports, which largely capture the information available to judges when they make their release decisions. We use these reports to collect several pieces of private information not used by the algorithm, such as information on gender and race, whether the defendant is homeless, whether the defendant is missing a telephone number, and whether the charge involves violence against adults or children. We also collect information on whether the pretrial services officer recommended an override of the algorithmic recommendation and noted any supporting aggravating factors, such as whether the defendant poses a threat to others or suffers from mental illness. We explore whether the high- and low-performing judges use this private information differently by constructing a second alternative algorithm that predicts the judges' release decisions (not pretrial misconduct) using both the full

[15] We also find that the high- and low-performing judges override the algorithm at very similar parts of the risk score distribution (Appendix Figure A.8) and for observably similar types of defendants within narrow risk score bins (Appendix Table A.5).

set of observable characteristics $\mathbf{X}_i$ and this private information $\mathbf{V}_{i,j}$, separately for high- and low-performing judges. Figure 6b presents the conditional misconduct rates under these new release rules. We find a modest deterioration in performance for the low-performing judges and a modest improvement in performance for the high-performing judges when we use both observable and private information to predict release decisions relative to when we use only observable information (Figure 6a), although we caution that the estimates are statistically imprecise.

We also find several notable differences in how the judges use the private information available from the pretrial risk assessment reports, after controlling for detailed risk score fixed effects and the full set of observable characteristics $\mathbf{X}_i$ (Panel B of Appendix Table A.5). For example, low-performing judges are much less likely to release defendants with an out-of-state address compared to high-performing judges, despite these factors not being particularly predictive of misconduct risk in unreported results. The low-performing 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-performing judges, despite these factors being predictive of misconduct risk.

The second way private information adds noise and inconsistency is how the judges use financial release conditions that require private information not included in the algorithm. For example, the judges in our setting assign monetary bail in 47% of all cases, with 37% of released defendants assigned an average bail amount of $11,325 and 93% of detained defendants assigned an average bail amount of $18,954.[16] In theory, the judges assign monetary bail in these cases to encourage the 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. However, the algorithm and the corresponding risk assessment report do not provide information on a defendant's financial resources, so 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, requiring counseling for mental health issues, issuing no contact orders with victims, and requiring supervision by pretrial services. These non-financial release conditions again require that the judges use private information to accurately address a defendant's needs.

Figure 7a explores whether high- and low-performing judges differ in their use of financial and non-financial release conditions by regressing each release condition on an indicator for being a high-performing judge, controlling for risk score fixed effects. We find that the low-performing

[16]The judges assign monetary bail in 93% of cases with harsh overrides where the algorithm recommends release, and 82% of cases with lenient overrides where the algorithm recommends detention. The average bail amounts for harsh and lenient overrides involving monetary bail are $18,662 and $13,667, respectively.

judges are 9.9 percentage points (SE: 5.0) more likely to assign monetary bail than the high-performing judges conditional on the risk score. The low-performing judges also perform particularly poorly when they assign monetary bail (Appendix Figure A.9), suggesting that these judges may mistakenly release some high-risk defendants and mistakenly detain some low-risk defendants when setting monetary bail. Meanwhile, the high-performing judges are much more likely to use non-financial conditions of release that are meant to directly address the defendants' underlying needs. The high-performing judges are 9.3 percentage points (SE: 3.2) more likely to impose drug and alcohol treatment compared to the low-performing judges, 1.3 percentage points (SE: 1.1) more likely to impose mental health treatment, 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-performing judges may be unnecessarily adding noise and inconsistency to their release decisions by overusing financial release conditions and underusing non-financial release conditions that address the defendants' needs.

The third way private information adds noise and inconsistency to release decisions is how the judges react to uninformative details or events. To study this channel, we focus on the judges' reactions to a highly salient but largely uninformative event. Specifically, we study hearings held just after a case where a different defendant was arrested for a homicide or violent first-degree felony while on pretrial release. 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 bail judge assigned to the case is generally not the judge who initially released the defendant.

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

$$D_{i,j,t} = \sum_{k \neq -1} \beta_k \mathbf{1}\{K_{j,t} = k\} + \mathbf{U}_i'\omega + \mathbf{W}_i'\gamma + \alpha_j + \varepsilon_{i,j,t}, \tag{7}$$

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 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 whom 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 so that we can observe the

entire sequence of each judge's caseload. Standard errors are clustered at the judge level.[17]

Figure 8 plots our event-study estimates and corresponding 95% confidence intervals. We also report average treatment effects, pooling the first four post-treatment shifts. In the full sample of judges (Figure 8a), 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. Figure 8b shows that these effects are driven by the low-performing judges, whose pretrial 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-performing judges. The effects are similar among observably low- and high-risk defendants (Appendix Figure A.11a) but are substantially larger among non-white defendants (Appendix Figure A.11b). The effects by race are notable given prior work documenting racial discrimination in bail decisions (Arnold, Dobbie, and Yang, 2018; Arnold, Dobbie, and Hull, 2022).

We view the above patterns as overreactions to a highly salient but largely uninformative event among low-performing judges, rather than a rational updating of beliefs. Consistent with this view, 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 (Appendix Figure A.12). These results, as well as the judges' release rates returning to baseline levels soon after the adverse event, are inconsistent with most models of rational updating.

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-performing 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 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

[17]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 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.10a) and that our effects are not driven by changes in the types of cases assigned to judges (Appendix Figure A.10b).

diagnosis and history, and financial resources). We surveyed 28 of the judges in our sample, with similar response rates for the high- and low-performing judges. The Online Appendix provides additional details on the survey administration, response rates, and a full list of questions.

Figure 7b presents the bivariate relationship between being a high-performing judge and three mutually exclusive indices, as well as for the individual factors within each index. We use the following three indices: (1) observable information available to both the judges and the algorithm, (2) private demographic information available to only the judges, and (3) private non-demographic information available to only the judges. The survey data reveal several striking patterns. The high- and low-performing 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 (Appendix Table A.5). High-performing judges report putting 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-performing judges do not report placing different weights on factors such as whether the current offense is a violent offense and whether the defendant has a past history of pretrial misconduct.

However, there are important differences in how the high- and low-performing judges report using the private information that is not available to the algorithm, again consistent with what we observe in the administrative court data (Appendix Table A.5 and Figure 6). In the survey data, high-performing judges are 19.0 percentage points (SE: 9.8) less likely to report putting weight on private demographic information such as gender and race, matching what we found in the administrative data. In contrast, the high-performing judges are 16.4 percentage points (SE: 8.1) more likely to report putting 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-performing judges place on these factors is particularly illuminating in light of our findings that low-performing judges' underperformance could be driven by their ineffective use of monetary bail while high-performing judges' outperformance could be driven by their greater use of non-financial conditions such as treatment for substance abuse disorders.

Our findings here and in the administrative court data suggest that the use of relevant private information can help explain why the high-performing 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 private information 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 help them learn what private information is relevant.

VI Conclusion

This paper shows that there is substantial variation in the impact of human discretion over an algorithm 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 the algorithmic counterfactual. These performance differences are most likely driven by how the judges use the private information that is unavailable to the algorithm, with high-performing judges using such information to improve the accuracy of their decisions and low-performing judges only adding noise and inconsistency 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 performance 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-performing 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 outcome of interest is both known and well-measured among the subset of treated individuals. 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, medical diagnoses, lending, and foster care decisions.

Data Availability Statement

The data and code underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.15148966.

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*.

American Bar Association. 1981. "American Bar Association Standards Relating to the Administration of Criminal Justice - Pretrial Release." *Operations of the Pretrial Services Agencies - Hearings*, 497–552.

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. 2022. "Measuring Racial Discrimination in Bail Decisions." *American Economic Review*, 112(9): 2992–3038.

Baron, E. Jason, Joseph J. Doyle Jr., Natalia Emanuel, Peter Hull, and Joseph Ryan. 2024. "Discrimination in Multiphase Systems: Evidence from Child Protection." *Quarterly Journal of Economics*, 139(3): 1611–1664.

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.

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

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

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.

Dobbie, Will, Andres Liberman, Daniel Paravisini, and Vikram Pathania. 2021. "Measuring Bias in Consumer Lending." *Review of Economic Studies*, 88(6): 2799–2832.

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.

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

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

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*.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. "Simple Rules for Complex Decisions." *Working Paper*.

Kesavan, Saravanan, and Tarun Kushwaha. 2020. "Field Experiment on the Profit Implications of Merchants' Discretionary Power to Override Data-Driven Decision-Making Tools." *Management Science*, 66(11): 5182–5190.

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.

Ludwig, Jens, and Sendhil Mullainathan. 2021. "Fragile Algorithms and Fallible Decision-Makers." *Journal of Economic Perspectives*, 35(4): 71–96.

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

Ouss, Aurélie, and Megan Stevenson. 2023. "Does Cash Bail Deter Misconduct?" *American Economic Journal: Applied Economics*, 15(3): 150–182.

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.

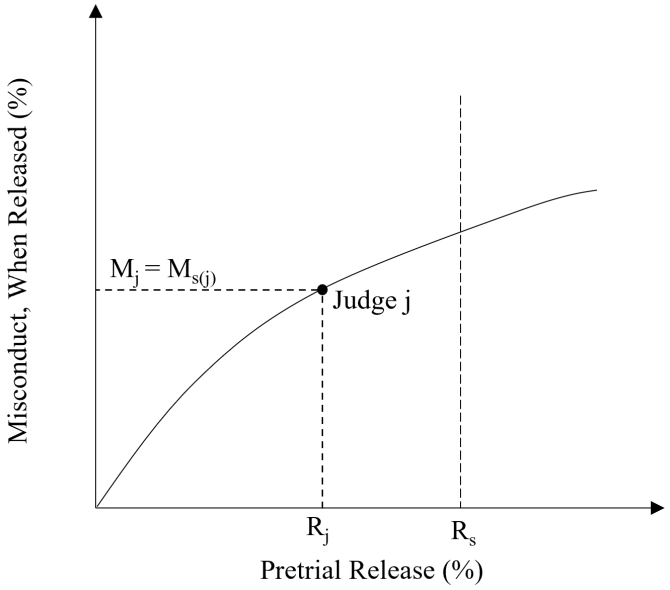
Stevenson, Megan. 2018. "Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes." *Journal of Law, Economics, and Organization*, 34(4): 511–542.

Stevenson, Megan T., and Jennifer L. Doleac. Forthcoming. "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy*.

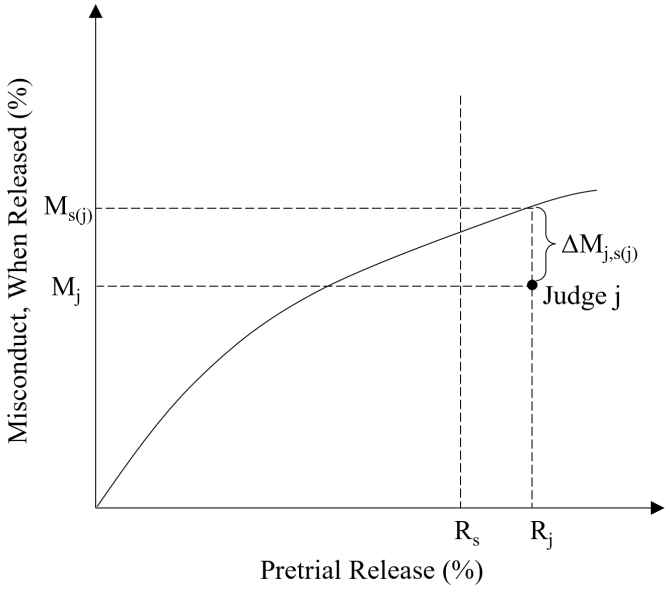
Sun, Jiankun, Dennis J. Zhang, Haoyuan Hu, and Jan A. Van Mieghem. 2022. "Predicting Human Discretion to Adjust Algorithmic Prescription: A Large-Scale Field Experiment in Warehouse Operations." *Management Science*, 68(2): 846–865.

Figure 1: Hypothetical Variation in Release Thresholds and Predictive Performance

A. Different Release Threshold, Equal Performance

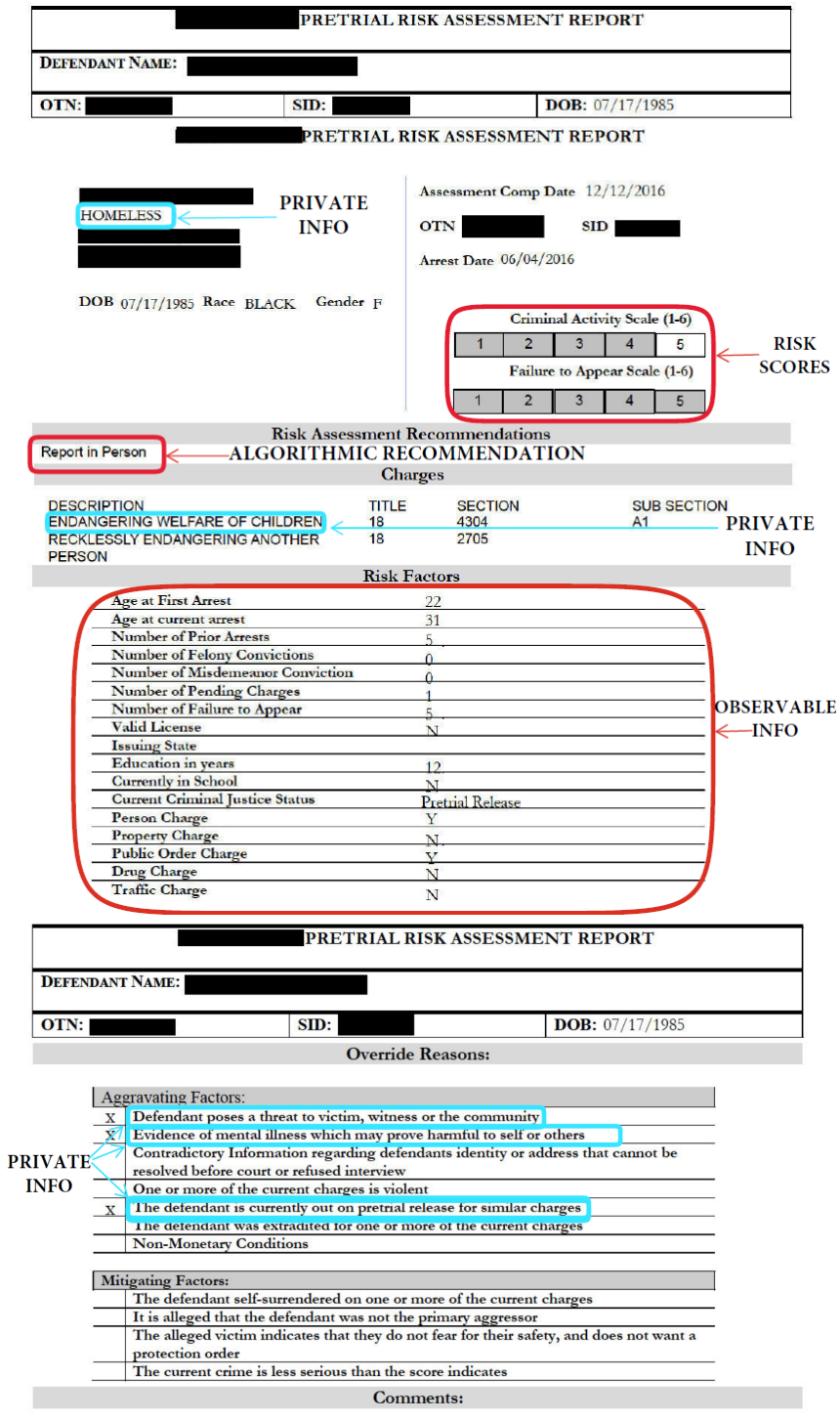


B. Different Release Threshold, Different Performance



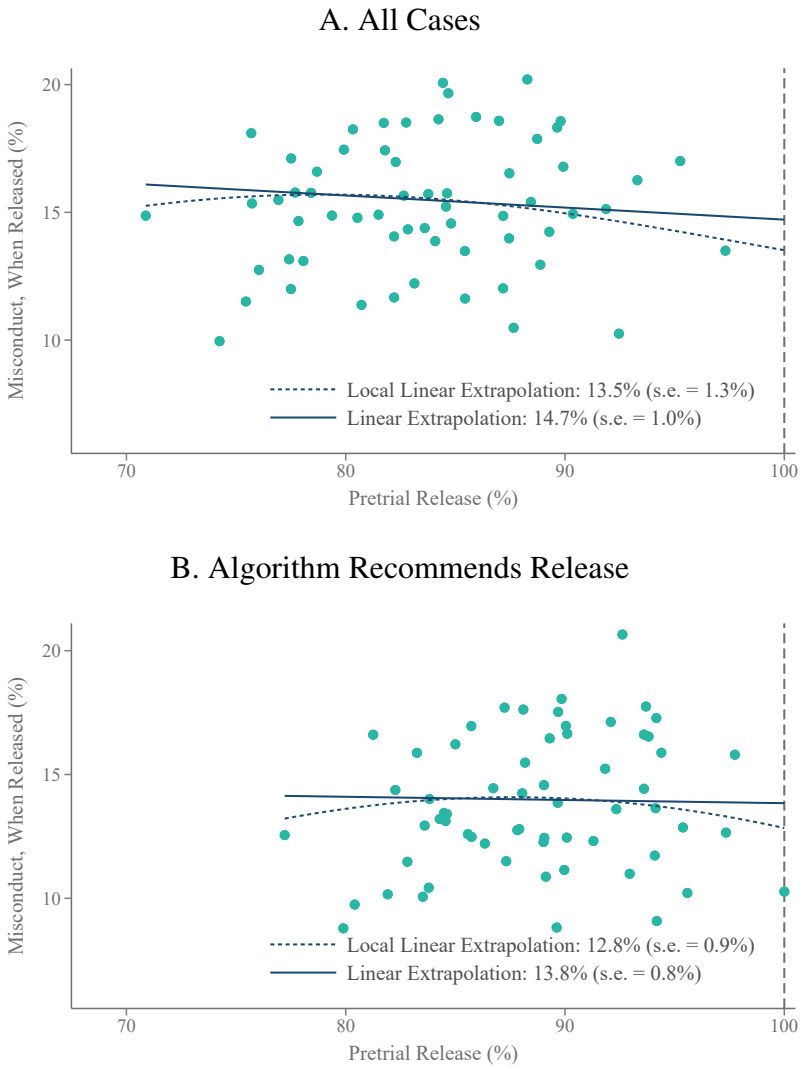
Notes. This figure plots observed misconduct rates among released defendants at each release rate for a hypothetical algorithm, as represented by the solid curve. This figure also plots the observed conditional misconduct rate and release rate for a hypothetical judge j, as represented by the black dot. Panel A presents a scenario where the hypothetical judge has a lower release threshold than the algorithm's recommended release threshold ($R_j < R_s$) but performs equally well relative to the algorithm at her release rate ($\Delta M_{j,s(j)} = 0$). Panel B presents a scenario where the hypothetical judge has a higher release threshold than the algorithm's recommended release threshold ($R_j > R_s$) but performs better relative to the algorithm at her release rate ($\Delta M_{j,s(j)} < 0$).

Figure 2: Example Pretrial Risk Assessment Report



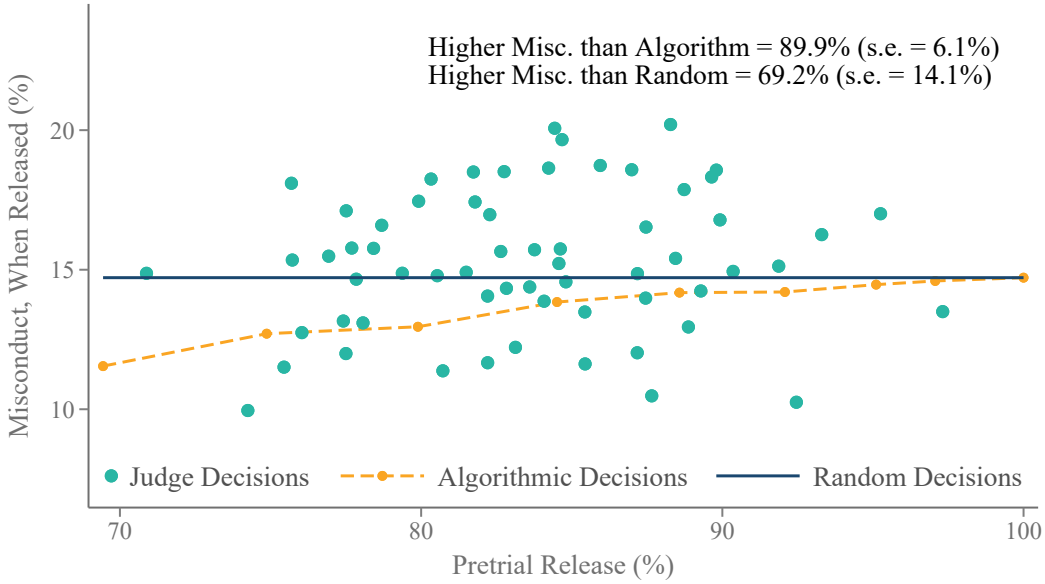
Notes. This figure shows an example pretrial risk assessment report in our setting. Red boxes indicate risk assessment scores, the algorithmic recommendation, and observable information. Blue boxes indicate examples of private information not included in the algorithm. See the main text for additional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

Figure 3: Extrapolations of Release and Conditional Misconduct Rates



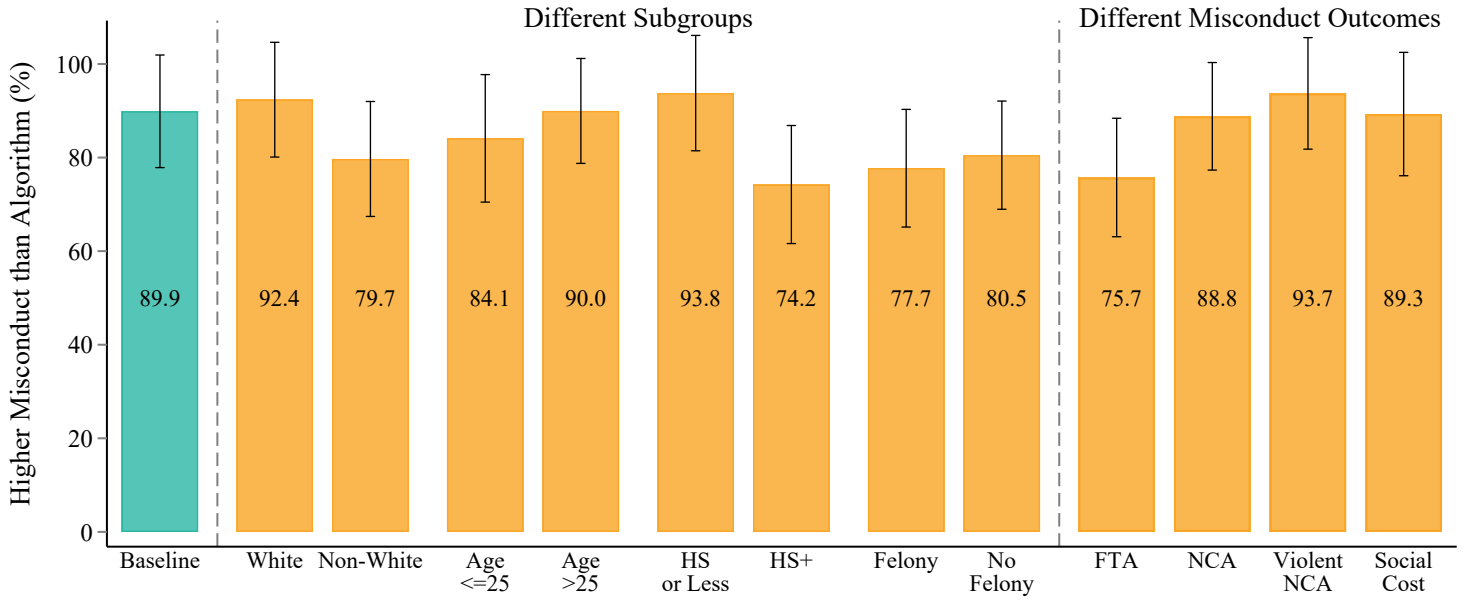
Notes. This figure plots observed misconduct rates among released defendants against judge release rates at two algorithmic risk score cutoffs. Each green dot represents the mean release rate and conditional misconduct rate for each judge, adjusted for shift-by-time fixed effects. Panel A reports results 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 plots local linear and linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the predicted misconduct rate among released defendants. The local linear regression uses a Gaussian kernel with a fixed bandwidth. We also report the estimated intercept and standard error at a cutoff-specific release rate 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 4: Conditional Misconduct Rates Relative to the Algorithm



Notes. This figure plots observed misconduct rates among released defendants against release rates 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 algorithmic release rule are estimated using linear extrapolations of mean risk at different risk score cutoffs as illustrated in Figure 3. Conditional misconduct rates under the random release rule are estimated using linear extrapolations of mean risk for the full sample as described in the main text. All estimates are adjusted for shift-by-time fixed effects. This figure also reports the fraction of judges with higher misconduct rates compared to the algorithmic and the random release rules using the posterior average effects approach described in the main text. 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.

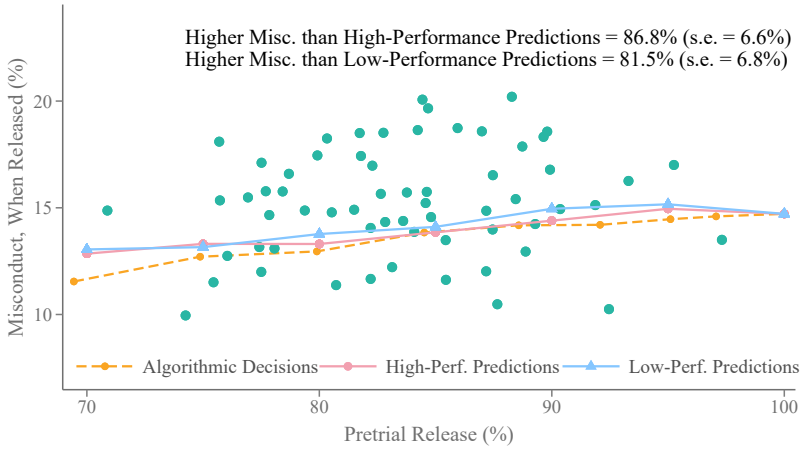
Figure 5: Fraction of Judges Underperforming the Algorithm for Different Subgroups and Misconduct Outcomes



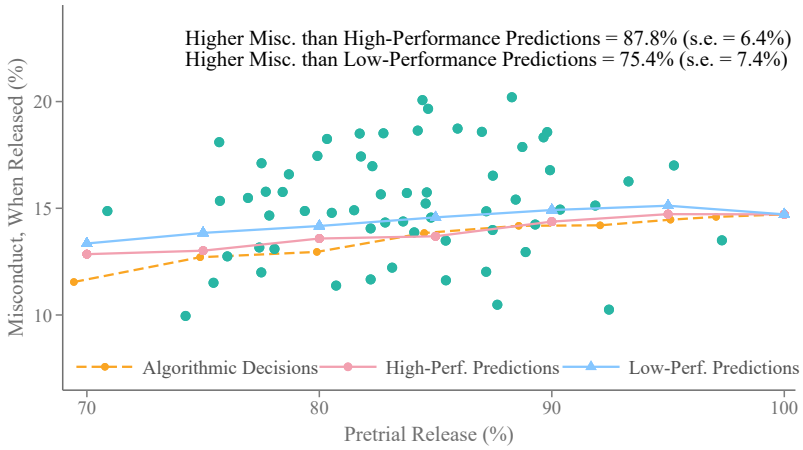
Notes. This figure shows the fraction of judges underperforming the algorithm using the posterior average effects approach described in the main text. The first bar shows the fraction of judges underperforming the algorithm in the full sample, as reported in Figure 4. The second set of bars shows the fraction of judges underperforming the algorithm in different defendant subgroups. The third set of bars shows the fraction of judges underperforming the algorithm for different misconduct outcomes. We also plot 95% confidence intervals for all estimates. See the Figure 4 notes for additional details.

Figure 6: Predicted Pretrial Release Decisions

A. Predictions Based on Observable Information

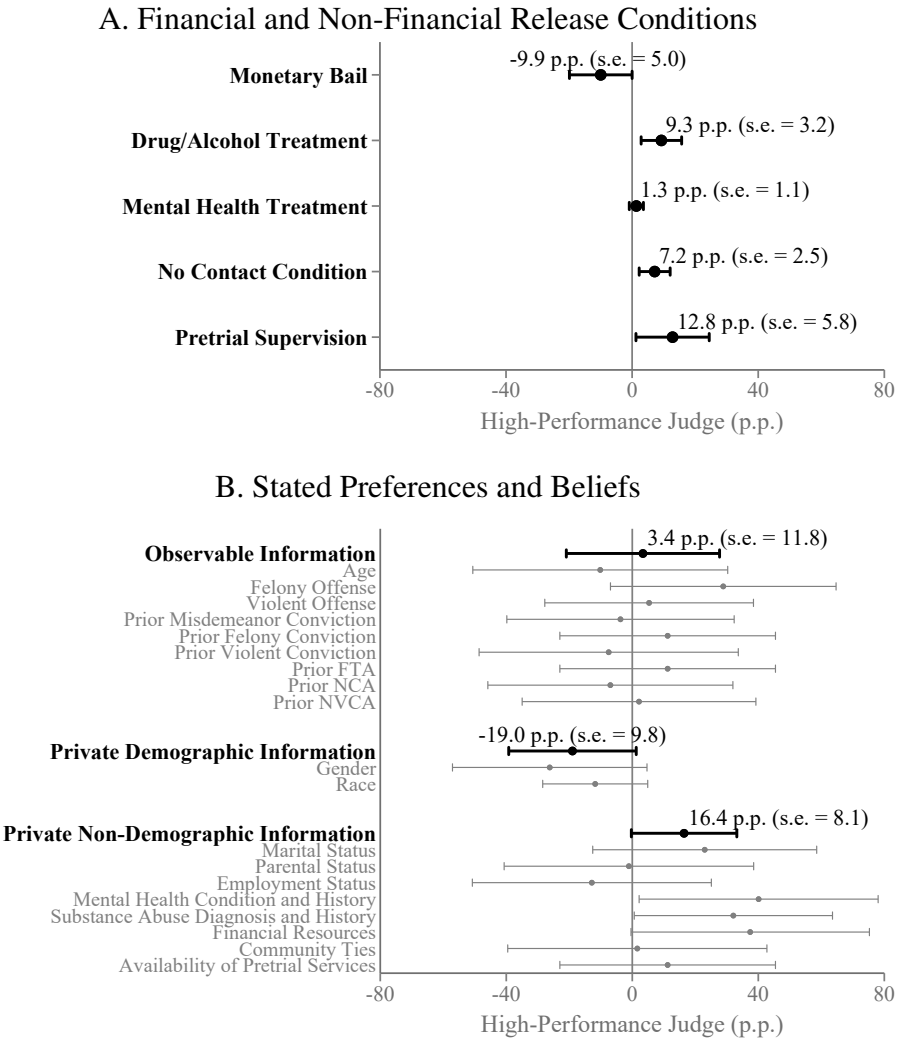


B. Predictions Based on Observable and Private Information



Notes. This figure plots observed misconduct rates among released defendants against release rates for the 62 judges in our sample, along with counterfactual misconduct rates from the high- and low-performing 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 Panel C of 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 are adjusted for shift-by-time fixed effects. This figure also reports the fraction of judges with higher misconduct rates compared to the predicted release decisions using the posterior average effects approach described in the main text. 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.

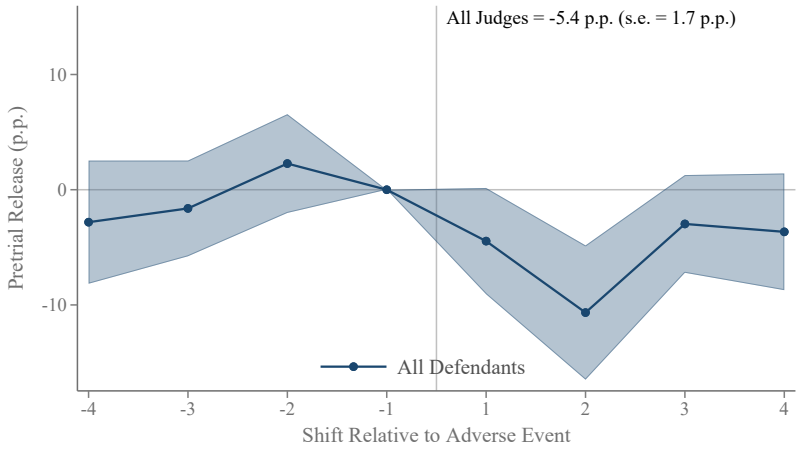
Figure 7: Revealed and Stated Preferences of High-Performance Judges



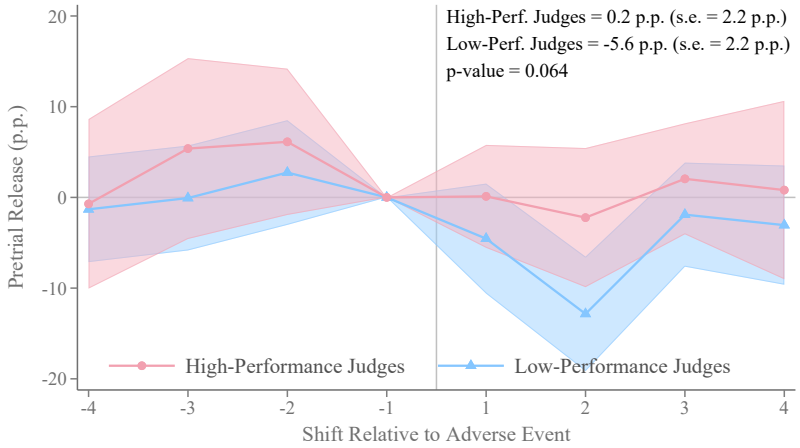
Notes. This figure reports OLS estimates showing the revealed and stated preferences of high-performance judges. Panel A reports OLS estimates from regressions of indicators for financial and non-financial release conditions on an indicator for being a high-performance judge for all 62 judges in our sample, controlling for risk score fixed effects. Panel B reports OLS estimates from regressions of stated judge preferences and beliefs on an indicator for being a high-performance judge for the 28 judges in our sample who took the survey. The survey asked judges to rank the importance of different defendant and case characteristics when making the decision to impose monetary bail. The bolded rows in Panel B 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 in both panels.

Figure 8: Effect of Adverse Events on Pretrial Release

A. All Judges



B. By Judge Performance



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 all 62 judges in our sample, and Panel B reports results separately for high-performing and low-performing 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 at the judge level. 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 main text for additional details on the sample and regression specification.

Table 1: Descriptive Statistics

		Recommend Detain		Recommend Release	
	All Cases	Lenient Override	Follow Algorithm	Harsh Override	Follow Algorithm
A. Pretrial Outcomes	(1)	(2)	(3)	(4)	(5)
Released Before Trial	0.83	1.00	0.00	0.00	1.00
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
Person Charge	0.41	0.17	0.20	0.38	0.46
Parole/Probation	0.27	0.46	0.65	0.52	0.18
Pretrial Release	0.32	0.91	0.89	0.33	0.20
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	0.12	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
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 main text. Information on case and defendant characteristics and pretrial outcomes is derived from court records as described in the main text. 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-Performing 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 from regressions of an indicator for being a high-performing judge 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 main 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 main text for additional details.