Algorithmic Recommendations and Human Discretion[*]

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

October 25, 2022

PRELIMINARY DRAFT – PLEASE DO NOT CITE OR DISTRIBUTE

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 the human biases and mistakes that motivated the adoption of the algorithms in the first place. We develop new quasi-experimental tools to measure the impact of human discretion over an algorithm, 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 generally underperform the algorithm when making a discretionary override, with most judges 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 making 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 less likely to overreact to highly-salient events compared to the low-performing judges.

[*] We thank Alex Albright, David Arnold, Raj Chetty, Dhruv Gaur, Benjamin Goldman, Nathan Hendren, Mitch Hoffman, Peter Hull, Lawrence Katz, Bentley MacLeod, Sendhil Mullainathan, Issac Opper, Emma Rackstraw, Manish Raghavan, and numerous seminar participants for helpful comments and discussions. We are indebted to Kenneth Gu, Dan Ma, and Nada Shalash 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 (e.g., Tversky and Kahneman 1974; Mullainathan 2002; Bordalo, Gennaioli, and Shleifer 2012; Kahneman, Sibony, and Sunstein 2021). These seemingly intractable problems have contributed to the rapid adoption of predictive algorithms in a range of high-stakes settings, from job screening to medical diagnoses to pretrial release decisions. Yet these same settings still require that a human 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. But allowing for such human discretion can also reintroduce the same human biases and mistakes that helped drive the development and introduction of the algorithms to begin with. Distinguishing between these possibilities and measuring the impact of human discretion remains difficult, complicating efforts to develop optimal oversight policies.

This paper develops new quasi-experimental tools to measure the impact of human discretion over an algorithm on the accuracy of decisions. We develop these tools in the context of bail decisions, where the sole legal objective of judges is to allow most defendants to be released before trial while minimizing the risk of pretrial misconduct (such as failing to appear in court or being arrested for a new crime). To help guide these decisions, judges are often given an algorithmic risk assessment that includes the predicted likelihood of misconduct and a recommendation of whether to release or detain the defendant. But judges frequently override these algorithmic recommendations, despite influential work showing that such algorithms can substantially outperform a bail judge working alone (e.g., Kleinberg et al. 2018). The key open question is whether allowing for such human oversight and discretion can yield even more accurate decisions, such that a human and an algorithm working together can be better than an algorithm alone.

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

Estimating the impact of human discretion over an algorithm in this way is complicated by an important selection challenge. We only observe pretrial misconduct among the selected subset of defendants that the judges choose to release before trial. We are therefore unable to directly measure the counterfactual misconduct rate under the algorithm at the judges' existing release rates, as we are missing the required outcomes among the defendants that the algorithm would have released but the judges chose to detain. This selection challenge can be understood as a kind of missing data problem that economists and statisticians have developed methods to overcome in a variety of contexts.

The first part of the paper shows that we can overcome this selection challenge and measure the impact of human discretion 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 show how the selection problem at a given release rate can be solved by estimating the average misconduct potential of defendants with risk scores at or below a relevant cutoff. Second, we show how to estimate the required average misconduct parameter by extrapolating observed misconduct rates across quasi-randomly assigned judges, building on the methods developed in Arnold, Dobbie, and Hull (2022). Finally, we show how to compare each judge to the algorithmic counterfactual at her existing release rate by repeating these extrapolations for a wide range of risk score cutoffs that span the judges' existing release rates. Our approach to measuring the impact of human discretion only requires that average misconduct risk among defendants at different risk scores can be accurately extrapolated from the data and that the judges' legal objectives are well-specified — assumptions that we provide extensive support for throughout the paper.

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

We find that the judges in our setting generally underperform the algorithm when they make discretionary overrides, increasing pretrial misconduct by an average of 2.4 percentage points at the judges' existing release rates (a 15% increase from the mean). This finding indicates that the typical judge in our setting is less skilled at predicting misconduct than the algorithm and that we could substantially decrease misconduct by automating release decisions. But this average impact masks substantial variation in the judges' performance compared to the algorithm, as we might have expected given the variation in judges' predictive skill. The negative average impact of human oversight and discretion is explained by the 90% of judges who generally underperform the algorithm when they make discretionary overrides. 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. Yet we also find that 10% of the judges generally outperform the algorithm when they make discretionary overrides, suggesting that a human and algorithm working together can potentially outperform automated release decisions. These high-skill judges are evenly distributed across the range of release rates and have similar demographic characteristics, political affiliations, caseloads, and years of experience as the low-skill judges, despite the large differences in performance.

We consider several extensions to our main results. One particularly important concern in our setting is that the judges may care about both racial fairness and accuracy, and the judges we identify as low-skill are simply putting more emphasis on racial fairness than the other judges. Yet the opposite pattern emerges in our data. The low-skill judges underperform the algorithm with respect to both accuracy and racial fairness, increasing the release disparity between white and non-white defendants with the same misconduct potential. Conversely, the high-skill judges generally outperform the algorithm in terms of both accuracy and racial fairness, decreasing (but not eliminating) the disparity between white and non-white defendants with the same misconduct potential. We also show that our results are robust to different sample restrictions, extrapolations of misconduct risk, definitions of pretrial misconduct, classifications of pretrial release, and algorithmic comparisons.

The final part of the paper provides more suggestive evidence on the behavior underlying the differences in judge performance and predictive skill. We start by showing that the high- and low-skill judges use the observable information that is available to both the judges and the algorithm in a remarkably similar way, overriding the algorithm at similar rates and at similar parts of the risk score distribution for observably similar defendants. These findings suggest that the high- and low-skill judges instead likely differ in their use of private information that is not available to the algorithm. We provide two sets of results to more directly support this explanation. We first build a new algorithm that predicts the judges' release decisions using information that is observable to both the judges and the original algorithm. The results confirm that a key difference between the high- and low-skill judges is likely how they use private information. The high-skill judges meaningfully outperform their predicted decision rule, suggesting that they are using relevant private information to improve the accuracy of their decisions. In contrast, the low-skill judges underperform their predicted decision rule, suggesting they are instead adding noise and inconsistency to their decisions when they attempt to use such private information. We then examine the judges' reactions to a highly-salient but largely uninformative event to better understand one particular way that private information can lead to differences in judge performance. We find that the judges, particularly the low-skill judges, are much more likely to detain observably low-risk defendants after hearing a case where a different and completely unrelated defendant is arrested for a serious violent offense while on pretrial release. We argue that this is likely an overreaction on the part of the judge, as there are also no detectable changes in conditional misconduct rates and the effects are concentrated among defendants that are either particularly representative of those arrested for serious violent crimes (Kahneman and Tversky, 1972; Bordalo et al., 2016) or with observable characteristics that are particularly overweighted by judges (Bordalo, Gennaioli, and Shleifer, 2015; Sunstein, 2022).[1]

Our findings are an important proof of concept that the most skilled human decision-makers can still add value to the decision-making process. One insight from our work is that there will not necessarily be a single correct human oversight policy since the impact of such policies depends on the predictive abilities of the human decision-makers. The most skilled human decision-makers can significantly improve

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

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. These findings are consistent with recent work showing that strict guidelines can reduce welfare when there is variation in human ability and that more nuanced policies are needed to improve decision-making in such settings (e.g., Currie and MacLeod 2020; Rambachan 2021; Chan, Gentzkow, and Yu 2022). An important question for future work is how to improve the predictive abilities of human decision-makers and, when that is not possible, how to constrain the least skilled decision-makers.

Our paper complements an important literature showing that predictive algorithms generally 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 without an algorithm. Yet recent work shows that judges frequently override the recommendations generated by such predictive algorithms in practice (e.g., Stevenson 2018; Albright 2021; Stevenson and Doleac 2021; Anwar, Bushway, and Engberg 2022). We connect these two streams of work by developing new tools to measure the impact of human oversight and discretion, showing that it is possible to identify counterfactual outcomes under the algorithm using the quasi-random assignment of decision-makers to individuals. These tools are broadly applicable in settings where there is quasi-random variation in human decision-makers and the objective of these decision-makers is both known and well-measured among the subset of individuals that the decision-maker endogenously selects.

Our paper also adds to a small but important literature studying the impact of human oversight and discretion. Hoffman, Kahn, and Li (2018) find that hiring managers with high override rates end up with worse overall hires, suggesting that discretion may decrease the accuracy of decisions in this setting. Conversely, De-Arteaga, Fogliato, and Chouldechova (2020) show that call workers on a child maltreatment hotline are more likely to override incorrectly-calculated algorithmic recommendations. Neither paper is able to quantify the impact of oversight and discretion on the accuracy of decisions, however, nor the variation in performance compared to an algorithm. Our results suggest that most humans decrease the accuracy of decisions, as in Hoffman, Kahn, and Li (2018), but that a human and an algorithm working together can be better than the algorithm alone in some situations, as in De-Arteaga, Fogliato, and Chouldechova (2020).

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 oversight and discretion. Section V explores potential mechanisms, and Section VI concludes. The Online Appendix provides additional results.

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 used by the algorithm and observable to the human decision-maker $\mathbf{X}_i \in \mathscr{X}$ ("observable information"), and another vector of characteristics that is not observable to the algorithm but again observable 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 partial subset of the private information $\mathbf{V}_{i,j}$ are observable to the econometrician.

Each decision-maker's goal is to align $D_{i,j}$ with Y_i^*, which captures the legitimate justification for setting $D_{i,j} = 1$. In the context of bail decisions, which we focus on in the remainder of this section, $D_{i,j} = 1$ indicates that judge j would release defendant i if assigned to her case (with $D_{i,j} = 0$ otherwise) while $Y_i^* = 1$ indicates that the defendant would subsequently fail to appear in court or be rearrested for a new crime if released (with $Y_i^* = 0$ otherwise). Each judge's legal objective is to release individuals without misconduct potential and detain individuals with misconduct potential, but judges may differ in their predictions of which individuals fall into which category. We note that $D_{i,j}$ is defined as the potential decision of judge j for defendant i, setting aside, for now, the judge decision rule which 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 characteristics in $\mathbf{X}_i$. We similarly define an algorithmic recommendation as a suggested decision based on such a mapping, such as $D_{i,s} = \mathbf{1}[a(\mathbf{X}_i) \leq s]$, where s is a threshold set by the algorithmic designer and represents the designer's preference for release. In our setting, $a(\mathbf{X}_i)$ is an algorithmic risk score that is meant to predict an individual's misconduct potential Y_i^* given observable case and defendant characteristics $\mathbf{X}_i$. Higher algorithmic risk scores are associated with a higher predicted misconduct potential, such that the algorithm recommends releasing individuals with low risk scores and detaining individuals with high risk scores.

Each judge j observes the algorithmic risk score $a(\mathbf{X}_i)$, the algorithmic recommendation $D_{i,s}$, the observable information $\mathbf{X}_i$, and the private information $\mathbf{V}_{i,j}$. The judge uses all of this information to form a subjective prediction of misconduct potential, which we define as a mapping $h_j(\cdot) : \bar{a} \times \bar{D} \times \mathscr{X} \times \mathscr{V} \to [0,1]$, where $\bar{a} = [0,1]$ and $\bar{D} \in \{0,1\}$. This mapping allows the judge to form a subjective misconduct prediction for each individual, $h_{i,j}(a(\mathbf{X}_i), D_{i,s}, \mathbf{X}_i, \mathbf{V}_{i,j})$. We include $a(\mathbf{X}_i)$, $D_{i,s}$, and $\mathbf{X}_i$ as separate inputs to the subjective misconduct prediction since the judge may not know the precise mapping from $\mathbf{X}_i$ to $a(\mathbf{X}_i)$ or may be differentially attentive to $a(\mathbf{X}_i)$ and $D_{i,s}$. We assume that each judge releases individuals in order of this subjective prediction, implying that the judge's decision rule can be represented by a threshold τ_j with $D_{i,j} = \mathbf{1}[h_{i,j}(a(\mathbf{X}_i), D_{i,s}, \mathbf{X}_i, \mathbf{V}_{i,j}) \leq \tau_j]$, where she releases defendants with low perceived misconduct potential and detains defendants with high perceived misconduct potential. The release threshold τ_j can be interpreted as judge j's preference for release under a simple model where the judge weighs the expected perceived cost of pretrial misconduct relative to the perceived social benefit to release (e.g., Kleinberg et al. 2018). This decision rule results in a judge-specific release rate $R_j = E[D_{i,j}]$ and judge-specific misconduct rate among released defendants $M_j = E[Y_i^* | D_{i,j} = 1]$.

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

We instead measure the impact of human oversight 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 altogether and isolating predictive skill at the judge's observed release rate. To build up to this measure, let the algorithmic release rule at judge j's existing release rate be:

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

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

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 the judge j's conditional misconduct rate M_j if she disagrees with the algorithm's misconduct predictions and rankings for some or all individuals.[2]

We can therefore measure the impact of human discretion over the algorithm by comparing a judge's observed misconduct rate M_j to the counterfactual misconduct rate of the algorithm at the judge's existing release $M_{s(j)}$:

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

where we say that judge j's discretion lead 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 pretrial decisions is given by the case-weighted average of $\Delta M_{j,s(j)}$ across all judges.

There are two reasons why the judge's subjective ranking may differ from the algorithm's ranking such that $\Delta M_{j,s(j)} \neq 0$, which correspond to two drivers of human decision-making that have historically been the focus of the economics and psychology literatures. The first is that the judge may systematically over- or underweight observable characteristics $\mathbf{X}_i$ relative to the risk score $a(\mathbf{X}_i)$. For example, whether the defendant is currently on parole or probation is a highly salient characteristic and a judge may believe that this trait is more positively correlated with misconduct potential than the algorithm. The effect of

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

such over- or underweighting on accuracy is theoretically ambiguous, as many existing pretrial algorithms (including the one we study) are deliberately simple and may only roughly approximate $E[Y_i^* \mid \mathbf{X}_i]$. The second reason that the judge may disagree with the algorithm's ranking of individuals is that she has access to private information that is not observable to the algorithm, $\mathbf{V}_{i,j}$. This private information may be a predictive signal of misconduct potential or noise that is not predictive of misconduct potential. Judge overrides based on predictive signals, such as relevant mitigating information presented by defense counsel at the bail hearing, will generally increase the accuracy of decisions. Conversely, judge overrides based on noise, such as extraneous factors like whether a local football team won or lost that week (Eren and Mocan, 2018), specific features of the defendant's appearance (Ludwig and Mullainathan, 2022), or sympathetic (but non-predictive) characteristics of the defendant will generally decrease the accuracy of decisions. In Section V, we will provide suggestive evidence on the most likely reasons that the judge's subjective rankings differ from the algorithm's ranking. We will also explore the possibility that judges may incorrectly predict defendants' ability to pay money bail and thus mistakenly release some high-risk defendants and mistakenly detain some low-risk defendants.

We emphasize that our analysis measuring the impact of human discretion over an algorithm relies on the assumption that the judges' legal objectives are well-specified and, as a result, that the judges release individuals in order of their subjective prediction of misconduct risk. We view this assumption as reasonable in our setting, as the sole legal objective of bail judges is to allow most defendants to be released before trial while minimizing the risk of pretrial misconduct. The narrow legal objective of bail judges is in sharp contrast to later stages of the criminal justice system such as sentencing, where judges are permitted to consider multiple objectives (e.g., Stevenson and Doleac 2021). We will also consider the possibility that the judges in our setting may have other objectives besides pretrial misconduct in robustness checks following recent work in this area (e.g., Kleinberg et al. 2018; Arnold, Dobbie, and Hull 2022).

II.B Graphical Intuition

Figure 1 illustrates the intuition for our approach using hypothetical variation in release and misconduct rates. The solid curved line represents the counterfactual misconduct rate of the algorithm at each possible release rate, with the vertical line at R_s denoting the release preference of the algorithmic designer at risk score threshold s. We also plot different hypothetical release rates and misconduct rates for a judge j to illustrate different scenarios that highlight the logic underlying our approach.

Panel A depicts a scenario where judge j has a lower release threshold than the algorithmic designers ($R_j < R_s$) and thus overrides the algorithmic recommendations. However, judge j overrides the algorithm solely because of differences in release preferences and thus follows the algorithm's ranking of individuals. As a result, the judge's misconduct rate is equal to the algorithm's misconduct rate holding fixed the judge's release rate, such that $\Delta M_{j,s(j)} = 0$. We can therefore say that there is no impact of human discretion on the accuracy of decisions in this first example. Panel B of Figure 1 illustrates an alternative scenario where judge j has a higher release threshold than the algorithmic designers ($R_j > R_s$) and thus overrides some of the algorithmic recommendations. But now judge j overrides the algorithm because she has a different ranking of individuals compared to the algorithm. In this particular case, the judge is able to make more accurate predictions than the algorithm and thus achieve a lower misconduct rate among released defendants, such that $\Delta M_{j,s(j)} < 0$. We can therefore say that human discretion increases the accuracy of decisions in this second example. The same underlying logic applies to a scenario where the judge makes less accurate predictions than the algorithm and thus achieves a higher misconduct rate among released defendants.

To summarize, we measure the impact of human discretion over the algorithm by comparing a judge's observed misconduct rate with the counterfactual misconduct rate under the algorithm at the judge's existing release rate. A judge who only overrides the algorithmic recommendation because she prefers a different release threshold will have the same misconduct rate as the algorithm at the same release rate. In contrast, a judge who overrides the algorithm because she ranks individuals differently than the algorithm will have a lower (higher) misconduct rate if she is more (less) skilled at predicting misconduct than the algorithm, again at the same release rate.

II.C Empirical Challenges

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

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

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

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

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

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

We show below that we can recover unbiased estimates of $M_{s(j)}$ using the as-good-as-random assignment of judges to defendants. The unbiased estimates of $M_{s(j)}$ then allow us to calculate unbiased estimates of $\Delta M_{j,s(j)}$ and other statistics of interest, such as the share of high-skill judges with $\Delta M_{j,s(j)} \leq 0$. Our quasi-experimental approach will provide an alternative to other approaches such as imposing strong selection-on-observables assumptions (e.g., Jung et al. 2017) or imposing worst- and best-case bounds on the algorithmic counterfactual (e.g., Kleinberg et al. 2018; Rambachan 2021).[3]

III Setting and Data

III.A Our Setting

We study the impact of human oversight and discretion in the context of a large, mid-Atlantic city that was one of the first jurisdictions in the country to introduce a pretrial risk assessment tool. The pretrial system is meant to allow the vast majority of criminal defendants to be released pending case disposition while minimizing pretrial misconduct. Bail judges are not meant to assess guilt or punishment when determining which individuals should be released from custody. In our setting, bail judges are directed to consider case and defendant characteristics only as that information is relevant to minimizing pretrial misconduct, defined as either failure to appear for a required court appearance (FTA) or new criminal activity after being released from jail and before the case is disposed (NCA), as measured by a new arrest. Bail judges in both our setting and across the country are granted substantial discretion in determining what pretrial conditions to impose in achieving this goal, although they are often provided with a release recommendation from the local pretrial services agency and/or an algorithmic risk score.

The pretrial services agency in our setting first started providing algorithmic risk scores and release recommendations in the mid-2000s. The introduction of algorithmic risk scores and release recommendations were a significant departure from the jurisdiction's previous approach of providing judges with a hand written questionnaire voluntarily filled out by the defendant (with many defendants leaving the form blank), the police affidavit, and a subjective judgment of risk. The algorithmic risk scores and recommendations have been credited with increasing both the share of released defendants and the share of released defendants without money bail in internal reports, and enjoy broad support within the jurisdiction's pretrial system. The risk assessment tool underlying the algorithmic risk scores and release recommendations has been updated several times since the initial roll-out and is locally validated, meaning that it is tested on a recent popula-

[3] Our quasi-experimental approach to measuring judge skill is much more informative than bounding approaches in our setting. Appendix Figure A.1 reports our main estimates under best-case bounds (where we impute $Y_i^* = 0$ for all detained defendants), and worst-case bounds (where we impute $Y_i^* = 1$ for all detained defendants). These bounds yield extremely wide estimates for the share of judges underperforming the algorithmic counterfactual, with anywhere from 0% to 98% of the judges being categorized as low-skill.

tion of defendants arraigned and released in the jurisdiction. We study one of the most recent iterations of this locally-validated risk assessment tool, first implemented in 2016. This iteration of the risk assessment tool was used until the onset of pandemic, when it became infeasible to conduct in-person interviews with defendants.

The risk assessment tool we study creates separate FTA and NCA risk scores based on defendant demographics (e.g., age at the current arrest), defendant criminal history (e.g., age at first arrest, number of prior felony and prior misdemeanor convictions), case characteristics (e.g., number of pending charges and charge types), and current criminal justice status (e.g., parole/probation status and pretrial release status). The specific defendant and case characteristics included in each risk score are based on whether there was a statistically significant association with the relevant misconduct outcome in the data used to build the algorithm, resulting in similar but slightly different set of inputs for the two risk scores. The included defendant and case characteristics are each associated with a certain number of points, which are aggregated to yield a detailed risk score that ranges from 0 to 73 for the FTA score and 0 to 30 for the NCA score. The detailed risk scores are then binned into aggregate scores that range from 1 to 6, with lower scores indicating a lower probability of misconduct and a higher score indicating a higher probability of misconduct.

The risk assessment tool also generates an automatic release recommendation based on the combination of these FTA and NCA binned scores (see Appendix Table A.1). Low FTA and NCA scores generate an automatic recommendation of release with no conditions, generally known as Release on Recognizance (ROR). These ROR recommendations are automatically generated for approximately 25% of cases in our data. More moderate FTA and NCA scores generate an automatic recommendation of release with regular phone or in-person check-ins, with these cases making up approximately 11% and 48% of cases in our data, respectively. Finally, the highest NCA score generates an automatic recommendation of no release. These highest-risk cases make up approximately 16% of cases in our data. The risk assessment never recommends money bail, despite local judges imposing money bail in approximately 47% of cases in our data. We focus on the NCA scores throughout our analysis, as this is the only risk score that generates variation in release recommendations.

The information required for the risk assessment tool is collected via several sources. The case begins with the arresting police officer inputting information on current arrest charges, which is electronically imported into the pretrial services case management system. The pretrial services officer then collects criminal history information from state and federal databases following identification and fingerprinting of the individual at the local jail. Additional factors for the risk assessment tool are collected by a pretrial services officer during a face-to-face interview conducted shortly after the defendant's booking. The pretrial services officer verifies this information and then enters it into the risk assessment tool. The defendant's gender and race are also collected and presented to the judge, but are not used as algorithmic inputs.[4]

Figure 2 shows a redacted example of the computer-generated risk assessment report provided to the

[4] The pretrial services officer assigned to a case is also given the discretion to override the algorithmic recommendation with supervisor approval in certain situations, typically to issue a harsher recommendation or to apply additional conditions of release. In practice, however, the pretrial services officers generally follow the algorithmic recommendation and have little impact on the overall accuracy of the release decisions. We therefore focus on the effect of judicial discretion over the algorithm throughout our analysis, abstracting away from the role of the pretrial services officer.

bail judge. The risk assessment report details the algorithmic release recommendation ($D_{i,s}$) and the binned FTA and NCA scores ($a(\mathbf{X}_i)$), as well as the defendant's arrest date, date of birth, description of the charges, and a description of the relevant risk factors entering the algorithmic risk scores and recommendation ($\mathbf{X}_i$). The report also includes certain demographic characteristics that are not used by the algorithm, such as the defendant's race and gender ($\mathbf{V}_{i,j}$). The bail judge uses this information to decide whether to release on recognizance (ROR), release with non-monetary conditions, impose monetary bail, or detain the defendant. Release on non-monetary conditions includes unsecured bail bond, which requires no money or deposit to secure release, and release on nominal bail, which requires the defendant to typically post $1.00 as deposit and have a designated person or organization act as a surety. Figure 3 presents a timeline of the process in our setting.

A key feature of our setting is that the bail judges are free to deviate from the algorithmic recommendation. Such deviations are common, with the judges in our sample overriding the default algorithmic recommendation in approximately 12% of cases where the algorithm recommends release ("low-risk cases") and over half of cases where the algorithm recommends detention ("high-risk cases"), with an overall override rate of 18%. But at the same time, the judges in our setting are responsive to the algorithmic recommendations, with pretrial release rates sharply falling by nearly 14 percentage points (a 19% decrease from the mean release rate) when the algorithmic recommendation discontinuously changes from release to detain (see Appendix Figure A.2). These patterns indicate that judges do consider and respond to the algorithmic recommendation, but nevertheless choose to override the recommendation in many cases. The combination of a longstanding risk assessment algorithm and frequent overrides for both observably low- and high-risk defendants makes this jurisdiction an ideal setting to study the impact of human discretion on the accuracy of decisions.

We exploit three additional features of the pretrial system in our analysis. First, the legal objective of bail judges is both narrow and measurable among the set of released defendants for whom pretrial misconduct outcomes are observed (although not among detained defendants, for whom such outcomes are not observed). This narrow legal objective yields a natural approach to measuring the accuracy of decisions at a given release rate, with lower pretrial misconduct rates indicating more accurate decisions and higher pretrial misconduct rates indicating less accurate decisions. We also explore the importance of potential objectives such as racial fairness and specific forms of misconduct in robustness checks.

Second, we follow the prior literature in viewing bail judges as effectively making binary decisions, releasing low-risk defendants (generally by ROR, non-monetary bail, or setting a low cash bail amount) and detaining high-risk defendants (generally by setting a high cash bail amount, or outright detaining them). The view that bail judges effectively make binary decisions is controversial in many settings, as many jurisdictions only allow the outright detention of defendants when they are charged with very serious offenses like capital crimes. But this simplification is particularly well-suited to our setting, where judges are constitutionally allowed to detain defendants when no other set of conditions could reasonably assure the safety of the public. As a result, the predictive algorithm in our setting explicitly recommends detaining observably high-risk defendants even when individuals are charged with less serious offenses such as misdemeanors. We also explore alternative characterizations of judge decisions in robustness checks, such as release without monetary conditions.

Third, 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 judges whose full-time assignment is to the court, while nighttime shifts and weekend/holiday shifts are covered by a group of nearby judges and/or senior judges. Appendix Table A.2 confirms that judge assignment to cases is balanced on all observable characteristics conditional on shift-by-time fixed effects, while Appendix Table A.3 shows that judge assignment has a strong first-stage effect on the probability that a defendant is released pretrial.

III.B Data and Summary Statistics

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

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 that are used to calculate the FTA and NCA risk scores, the automatic algorithmic recommendation, the pretrial officer's recommendation, the bail judge assigned to the case, whether the defendant was ultimately released before trial, and whether this release was due to ROR, release with non-monetary conditions, or release conditional on paying money bail. We categorize defendants as either released (including release on recognizance, release with non-monetary conditions, and payment of money bail) or detained (including non-payment of money bail or outright detention). Finally, we observe whether a defendant subsequently failed to appear for a required court appearance or was arrested for new criminal activity before case disposition among the subset of defendants released by the judge. We take either form of pretrial misconduct as the primary outcome of our analysis.[5]

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

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

restrictions leave us with 37,855 cases among 27,503 unique defendants assigned to 62 unique bail judges.

Table 1 summarizes our estimation sample, both overall and by the algorithm's automatic recommendation. Panel A column 1 shows that 83% of all defendants are released at some point before trial. Relatively few of these releases are without conditions, with 52% and 37% of released defendants having been assigned non-monetary and monetary conditions at the first bail hearing, respectively. There are also a small handful of defendants who are initially remanded without bail but later released. Columns 2 and 3 report summary statistics for observably high-risk defendants that the algorithm recommends detaining, separately by whether the judge overrides the recommendation ("lenient override") or follows the recommendation. Columns 4 and 5 present summary statistics for observably low-risk defendants that the algorithm recommends releasing, separately by whether the judge overrides the recommendation ("harsh override") or follows the recommendation. Importantly, these release statistics indicate that judges override the algorithm's recommendation in 54% of observably high-risk cases and 12% of observably low-risk cases. The total override rate is thus 18% after accounting for the distribution of observably low- and high-risk cases (and ignoring overrides that do not impact release decisions, such as moving from ROR to non-monetary conditions). On net, these overrides mean that observably high-risk defendants are less likely to be released than observably low-risk defendants, with a 54% release rate relative to 88% release rate.

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

Panels B and C also indicate that lenient overrides among observably high-risk cases and harsh overrides among observably low-risk cases are not random. Among observably high-risk defendants that the algorithm recommends detaining, defendants for whom judges issue a lenient override have fewer prior arrests and convictions, are less likely to be on parole or probation, more likely to be charged with drug or traffic offenses, and less likely to be male. The pattern is reversed for harsh overrides of observably low-risk defendants that the algorithm recommends releasing, with these defendants having more prior arrests and convictions, more likely to be on parole or probation, more likely to be charged with property and public order charges, more likely to be male and non-white.

Finally, Panel D shows that observably high-risk defendants that are released despite the algorithm's detention recommendation are 14.7 percentage points, or 51%, more likely to be rearrested or have an FTA than low-risk defendants that are released in compliance with the algorithm's release recommendation. Among released defendants who commit pretrial misconduct, the vast majority are rearrested for a new offense only. Importantly, and in contrast to the other statistics in Table 1, the risk statistics in Panel D are only measured among released defendants, which are a selected sample of all defendants. Pretrial misconduct potential is, by definition, not observed among detained individuals.

IV Effects of Human Discretion on Pretrial Release Decisions

IV.A Methods

We measure the impact of human oversight and discretion on the accuracy by comparing each judge's observed misconduct rate M_j to quasi-experimental estimates of the algorithm's performance at the same release rate $M_{s(j)}$. Our approach only requires that average misconduct risk among defendants at different release rates can be accurately extrapolated from the quasi-experimental data and that judges' legal objectives are well-specified.

The first key insight underlying our approach is that when judges are as-good-as-randomly assigned, the problem of measuring the algorithmic counterfactual holding fixed the judge's release rate, $M_{s(j)} = E[Y_i^*|a(\mathbf{X}_i) \leq s(j)]$, reduces to the problem of estimating the average misconduct risk at the appropriate algorithmic release threshold. However, due to the selective labels challenge, we do not observe Y_i^* for all defendants that the algorithm would release but the judge does not release in practice. Under as-good-as-random judge assignment, however, the average misconduct risk at a given algorithmic release threshold is common to all judges and captured by threshold-specific population misconduct risk.

The second key insight underlying our approach is that these average threshold-specific misconduct risk parameters can be estimated from quasi-experimental variation in pretrial release and misconduct rates, building on the methods developed in Arnold, Dobbie, and Hull (2022). To build intuition for this approach, consider a setting with as-good-as-random judge assignment and a supremely lenient bail judge j^* who releases nearly all defendants below a given algorithmic release threshold $s(j)$ for $j \neq j^*$, regardless of their potential for pretrial misconduct. This supremely lenient judge's release rate among these defendants is close to one hundred percent:

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

making the threshold-specific misconduct rate among defendants she releases close to the threshold-specific average misconduct risk in the full population:

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

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

In the absence of such a supremely lenient judge, the required average misconduct parameters can be estimated using model-based or statistical extrapolations of release and misconduct rate variation across as-good-as-randomly assigned judges. As discussed in Arnold, Dobbie, and Hull (2022), this approach is conceptually similar to how average potential outcomes at a treatment cutoff can be extrapolated from nearby observations in a regression discontinuity design, particularly so-called donut designs designs where the data in some window of the treatment cutoff is excluded. Here, released misconduct rates are extrapolated from as-good-as-randomly assigned judges with high leniency to the release rate cutoff of 100% given by a hypothetical supremely lenient judge. Mean risk estimates may, for example, come from the vertical intercept at 100% of linear or local linear regressions of estimated misconduct rates among released individuals $E[Y_i^* \mid D_{i,j}=1, a(\mathbf{X}_i) \leq s(j)]$ on estimated release rates $E[D_{i,j} \mid a(\mathbf{X}_i) \leq s(j)]$ across J judges at a given algorithmic release threshold $s(j)$.

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

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

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

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

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

IV.B Counterfactual Misconduct Under the Algorithm

Figure 4 illustrates our extrapolation-based estimation of the average misconduct risk for two algorithmic release thresholds. Panel A reports results for the full sample of cases, corresponding to an algorithmic release threshold of 100%. Panel B restricts the sample to cases where the algorithm recommends release, corresponding to an algorithmic release threshold of just over 80%. The horizontal axis in each panel plots release rates ($\hat{R}_j$) for each of the 62 judges in our data after regression adjustment for shift-by-time fixed effects. We find sizable variation across judges at each risk score threshold, with many judges releasing a high fraction of defendants. The vertical axis plots conditional misconduct rates for each judge ($\hat{M}_j$), regression-adjusted for shift-by-time fixed effects.

The vertical intercepts of the lines of best fit, at 100%, provide estimates of the threshold-specific average misconduct rates. The lines of best fit are obtained by OLS regressions of judge-specific conditional misconduct rate estimates on judge-specific release rate estimates, with the judge-level regressions weighted inversely by the variance of misconduct rate estimation error. We obtain standard errors using a bootstrap procedure, where we first take independent random draws from the distributions of the estimated judge-specific release rates ($\hat{R}_j$) and conditional misconduct rates ($\hat{M}_j$) and then recalculate the threshold-specific extrapolations. These estimates and associated standard errors are reported at the bottom of each panel. The simple linear extrapolation yield a precise mean misconduct estimate of 14.7% (SE: 1.0) for the full population of cases, which corresponds to a release rate of 100%. The extrapolation yields a mean misconduct estimate of 13.8% (SE: 0.8) for cases where the algorithm recommends release, which again corresponds to a release rate of just over 80%. 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 risks score cutoffs that correspond to release rates ranging from just under 70% to 100% – spanning both the unadjusted and regression-adjusted judge release rates observed in our sample. The results from these extrapolations are plotted in Figure 5 and reported with standard errors in Appendix Table A.4. The extrapolations show that, in practice, we observe risk scores that correspond to closely-spaced release rates (column 1, Appendix Table A.4) and that there is a (weakly) monotonically increasing relationship between risk scores and conditional misconduct rates (columns 2-3, Appendix Table A.4). These threshold-specific estimates allow us to measure the counterfactual misconduct rate under the algorithm and construct $\Delta M_{j,s(j)}$ for each judge in our sample using the linear spline approach discussed above. We take the linear extrapolation as our baseline specification for estimating the impact of judge discretion and explore the robustness of our results to alternative mean risk estimates below.

IV.C Effect of Judicial Discretion on Conditional Misconduct Rates

Figure 5 presents our main findings on the effect of judicial oversight and discretion 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 conditional misconduct rate $\hat{M}_j$ against the judge's release rate for all cases $\hat{R}_j$, both regression adjusted for shift-by-time fixed effects. The dashed orange line shows the estimated conditional misconduct rate for the algorithm at different release rates, estimated using linear extrapolations of average misconduct at each discrete risk score threshold and connected using a linear spline. The dashed gray line shows the conditional misconduct rate under a random release rule that a judge could achieve by ignoring the algorithm completely and releasing defendants by flipping a coin, estimated using a linear extrapolation of average misconduct in the full sample. We also report the share of judges with conditional misconduct rates that are higher than the algorithmic counterfactual and the random release rule, estimated using the posterior average effect approach of Bonhomme and Weidner (2022) that accounts for sampling error in our judge-level estimates. We obtain standard errors for these estimates using a bootstrap procedure, where we first take independent random draws from the distributions of the estimated judge-specific release rate $\hat{R}_j$ and conditional misconduct rates $\hat{M}_j$ and then recalculate the threshold-specific extrapolations and statistics of interest.

Three striking patterns emerge from the distribution of judge conditional misconduct rates and release rates, even before we compare the judges to the algorithmic counterfactual and random release rule. First, there is substantial variation in judges' preferences for release, with the regression-adjusted release rates ranging from just above 70% to over 95%. These patterns are consistent with prior work on the pretrial system (Arnold, Dobbie, and Yang, 2018; Arnold, Dobbie, and Hull, 2022) and highlight the importance of comparing each judge's outcomes to the algorithmic counterfactual at the same release rate. Second, there is also substantial variation in the judges' conditional misconduct rates at the same release rate, with a judge at the case-weighted 90th percentile of conditional misconduct rates having a misconduct rate that is 6.6 percentage points higher than a judge at the case-weighted 10th percentile. These findings suggest important variation in predictive skill across the judges and provide evidence against the standard monotonicity assumption, which implies that judges with the same release rate will have the same conditional misconduct rate (Frandsen, Lefgren, and Leslie, 2019; Chan, Gentzkow, and Yu, 2022). Third, the judges' conditional misconduct rates do not increase with the release rate, with an OLS regression of the judge-specific misconduct rates on the judge-specific release rates yielding a coefficient of -0.05 (SE: 0.08). This pattern again suggests important variation in predictive skill and is inconsistent with a standard model where the monotonicity assumption holds, which implies that the judges' conditional misconduct rate will increase with the release rate (Chan, Gentzkow, and Yu, 2022).

The comparison of the judges and algorithmic counterfactual reveals an even more striking pattern – the vast majority of judges significantly underperform the algorithm, as indicated by a conditional misconduct rate that is above the algorithmic counterfactual at the same release rate. We estimate that 90% (SE: 6.1) of judges generally underperform the algorithm when they make discretionary overrides, with a remarkable 69% (SE: 14.1) of judges underperforming the random release rule. These findings mean that, incredibly, most judges could achieve a lower misconduct rate by flipping a coin or using a random number generator. The system-wide impact of human discretion on the accuracy of release decisions is correspondingly negative, with the judges increasing pretrial misconduct by an average of 2.4 percentage points (SE: 0.5) at their existing release rates (column 1, Table 4). These findings therefore indicate that the typical judge in our setting is less skilled at predicting misconduct than the algorithm and that we could substantially decrease misconduct by automating release decisions.

But this negative system-wide impact of discretion masks substantial variation in the judges' performance compared to the algorithm, as we might have expected given the variation in the judges' performance compared to each other. Importantly, we find that 10% of the judges generally outperform the algorithm when they make discretionary overrides, as indicated by a conditional misconduct rate that is lower than the algorithmic counterfactual at the same release rate. This more positive finding suggests that a human and an algorithm working together can outperform an automated release decision in at least some situations and that a human can still add value to the decision-making process.

The remainder of the paper explores the variation in the judges' performance relative to the algorithm to better understand these results. While there is noise in our individual-level estimates of judge performance, for simplicity, we divide the judges into two exhaustive and mutually-exclusive groups based on their performance. The first group consists of "low-skill" judges that likely underperform the algorithmic counterfactual at the same release rate, while the second group consists of "high-skill" judges who are more likely to outperform the algorithmic counterfactual, and almost certainly more likely to outperform the random decisions counterfactual. We formally define a judge as low-skill if the posterior probability of $\Delta M_{j,s(j)} > 0$ is 0.9 or above, and a judge as high-skill otherwise, again using the approach of Bonhomme and Weidner (2022). There is a large mass of judges with a posterior probability of $\Delta M_{j,s(j)} > 0$ at 0.9 or above and, reassuringly, the set of low- and high-skill judges is nearly identical if we instead use observed estimates of $\Delta M_{j,s(j)}$ to categorize the judges.

Table 2 shows estimates from OLS regressions of an indicator for being a high-skill judge on different judge characteristics. There is no statistically significant relationship between judge performance and the judge's gender, race, political affiliation, experience, or caseload (columns 1-5). We also find that high- and low-skill judges are equally likely to override the algorithm (column 6) and have statistically identical release rates (column 7), ruling out the explanation that high-skill judges are simply more likely to follow the algorithmic recommendations compared to low-skill judges. None of the estimates from Table 2 suggest a clear explanation for the variation in judge performance, although we note that standard errors are large for many of the characteristics we consider.

IV.D Robustness and Extensions

Our approach to measuring the impact of human discretion over algorithms requires that average misconduct risk among defendants at different release rates can be accurately extrapolated from the quasi-experimental data and that judges' legal objectives are well-specified. We now consider several extensions to our main results that test or relax these assumptions.

Racial Fairness. One particularly important concern is that the judges in our setting may care about both racial fairness and overall accuracy, and the judges we identify as low-skill are simply putting more emphasis on racial fairness and less emphasis on accuracy. We explore this concern by measuring release disparities between white and non-white defendants with identical misconduct potential. This measure of racial fairness is consistent with the legal theory of disparate impact, as well as notions of algorithmic discrimination that compare equally "qualified" white and non-white individuals (Berk et al., 2018). We can measure these release disparities using estimates of race-specific misconduct risk to rescale observational release rate comparisons in such a way that makes released white and non-white defendants comparable in terms of misconduct potential within each judge's defendant pool (Arnold, Dobbie, and Hull, 2021, 2022).

Figure 6 plots these release disparities for the 62 judges in our sample, along with counterfactual disparities for the algorithm. The algorithm generates substantial release disparities between white and non-white defendants with identical misconduct potential, despite not including race or ethnicity as an input. This finding indicates that the vector of characteristics used by the algorithm ($\mathbf{X}_i$) are correlated with race and ethnicity after adjusting for misconduct potential (Arnold, Dobbie, and Hull, 2021). We find that nearly half of the judges in our setting generate lower release disparities than the algorithm on average when they make discretionary overrides, leading to generally fairer release decisions. Importantly, we find that there is a positive correlation between outperforming the algorithm in terms of accuracy and in terms of racial fairness in Table 2, columns 7 and 8. These findings indicate that high-skill judges generally outperform the algorithm in terms of both accuracy and racial fairness, decreasing (but not eliminating) release disparities between white and non-white defendants with identical misconduct potential.

Sample of Judges. Our baseline sample includes all bail judges who heard at least 100 cases during our sample period. We explore the sensitivity of our main results to a group of 54 bail judges who heard at least 200 cases in order to conduct our extrapolations using only those judges with substantial caseloads in Appendix Figure A.3. We continue to find that the vast majority of judges underperform the algorithm at their release rate, with 90% of judges underperforming the algorithm and 76% underperforming a random release rule.

Extrapolations of Misconduct Risk. Our baseline specification estimates counterfactual misconduct under the algorithmic using a series of linear extrapolations that control for shift-by-time fixed effects. We then connect the extrapolation-based estimates at each discrete risk score with a linear spline. We find qualitatively similar results when using local linear extrapolations in Appendix Figure A.4 and Appendix Table A.4, or using linear extrapolations that omit the shift-by-time fixed effects in Appendix Figure A.5, with a substantial share of judges underperforming both the algorithm and random release rule in all of the specifications we consider. We also find similar results when we do not connect discrete risk scores with a linear spline, instead constructing best- and worst-case step functions to connect the estimates at each discrete risk score threshold in Appendix Figure A.6.

We also find similar results using a modified approach where we extrapolate to the most lenient judge at a given risk score cutoff and then calculate bounds for the remaining share of defendants in Appendix Figure A.7. Our modified approach proceeds in four steps. First, we identify the release rate of the most lenient judge for each risk score cutoff in the data. Second, we estimate the conditional misconduct rate at the most lenient judge's release rate using our extrapolation-based procedure. Third, we construct worst- and best-case bounds for the remaining share of defendants detained by the most lenient judge. Finally, we calculate the counterfactual misconduct rate using a weighted average of the misconduct rate from the extrapolation at the most lenient judge's release rate and the bounding procedure for the remaining share of defendants. Our results are very similar using this modified approach, as we generally observe at least one judge that releases 100% of defendants at the lower risk scores.

Pretrial Misconduct. Our baseline measure of pretrial misconduct is defined as either FTA or NCA. But judges may care more about certain types of pretrial misconduct more than others. However, we find qualitatively similar results when measuring pretrial misconduct using only FTA in Appendix Figure A.8 or only NCA in Appendix Figure A.9, with approximately 75% to 90% of judges underperforming the algorithm. We also observe qualitative similar patterns when using only violent NCA in Appendix Figure A.10, but we are unable to calculate posterior average effects for this outcome since violent NCAs only occur in about 2% of cases in our sample.

Pretrial Release. Our baseline measure of pretrial release categorizes defendants as either released at any point before the case is disposed or always detained until the case is disposed. Our release variable therefore includes the release decision of the first bail judge assigned to the case, as well as the release decisions of any subsequent bail judge who reviews and potentially amends the first judge's decision. We find very similar results if we instead measure pretrial release in only the first three days of the original bail hearing in an attempt to isolate just the first judge's release decision, with 85% of judges underperforming the algorithm and 50% underperforming the random release rule using this alternative definition of pretrial release in Appendix Figure A.11.

Our baseline measure of release also includes defendants released on money bail, which raises the possibility that judges may not only make errors in predicting risk of pretrial misconduct, but also in predicting ability to pay. For instance, one explanation for our results is that low-skill judges may incorrectly predict defendants' ability to pay money bail and thus mistakenly release some high-risk defendants and mistakenly detain some low-risk defendants. We explore this possibility by replacing our baseline measure of release with an indicator for release on recognizance or release on non-monetary conditions (versus monetary bail or remand without bail) to focus on the set of defendants where there is no possibility of mispredicting the ability to pay. We find that nearly all judges underperform the algorithm when we examine the defendants who are released on recognizance or released on non-monetary conditions in Appendix Figure A.12, similar to our main findings. The share of judges that make override decisions that are no better than random, however, is sharply lower than in our main findings, at only 8.4%. These results suggest that a large portion of the overall underperformance of judges relative to the random release rule may be due to incorrectly predicting the defendants' ability to pay money bail. We caution, however, that it is impossible to for us to know if judges are setting money bail with the intention to release or detain an individual. These estimates should therefore be viewed as only suggestive.

Algorithmic Design. Our baseline analysis compares judges' outcomes to the recommendations of the proprietary algorithm that the judges see at the time of the bail decision. The algorithm itself is based on a simple scoring system, which assigns points to various demographic, criminal history, and current charge characteristics as described in Section III. This comparison allows us to measure the impact of human discretion over an algorithm, but a natural concern is that the best judges in our setting may still underperform a more sophisticated machine learning algorithm. We explore this concern by comparing the judges to the gradient-boosted decision tree algorithm developed by Kleinberg et al. (2018) in Appendix Figure A.13.[6] We

[6] We follow Kleinberg et al. (2018) closely in the construction of the gradient-boosted decision tree algorithm. The model is

find that this more sophisticated machine learning algorithm is more accurate than the proprietary algorithm, with misconduct decreasing by an average of 0.4 percentage points across the observed release thresholds. The share of judges underperforming the more sophisticated algorithm is correspondingly higher, at 95%, with the share of judges outperforming the more sophisticated algorithm falling to 5%.

V Potential Mechanisms

This section provides more suggestive evidence on the behavior underlying the differences in judge performance. We start by showing that the high- and low-skill judges use the observable information that is available to both the judges and the algorithm in a remarkably similar way. We then formally decompose the judges' performance differences into components that are and are not explained by the predictable use of the observable information, before examining the judges' reactions to a highly-salient but largely uninformative event to better understand one particular way that private information can result in inconsistency and noise.

V.A Use of Observable Information

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

We therefore start by exploring the possibility that the high- and low-skill judges override the algorithm at different parts of the risk score distribution in Appendix Figure A.2. Overrides at different parts of the risk score distribution could explain our results if, for example, the high-skill judges only override the algorithm for particularly "close calls" where the recommendation discontinuously changes from release to detain but the low-skill judges override the algorithm throughout the risk score distribution. But the data do not support this idea. Appendix Figure A.2 shows nearly identical release rates for high- and low-skill judges throughout the risk score distribution. Both types of judges release most observably low-risk defendants, with release rates declining monotonically with the risk score. Both types of judges are also equally likely to make overrides for the "close call" cases near the risk score threshold where the algorithmic recommendation changes from release to detain. Nothing in these results suggests that the differences in judge performance are driven by overrides at different parts of the risk score distribution.

We next explore the possibility that the high- and low-skill judges override the algorithm for observably different types of defendants *within* risk scores. For example, our findings could be explained by the high-skill judges differentiating defendants within risk scores by using additional information contained in the observable characteristics, resulting in the two groups of judges releasing a different mix of defendants

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

within each risk score. We explore this possibility in Table 3 by regressing an indicator for pretrial release on the full set of algorithmic inputs ($\mathbf{X}_i$) and several demographic characteristics (a subset of $\mathbf{V}_{i,j}$) separately for high- and low-skill judges. We also control for detailed risk score fixed effects, so that the coefficients reflect the additional weight that judges place on each characteristic relative to the risk score. Column 1 reports these results for high-skill judges, column 2 reports these results for low-skill judges, and column 3 reports the p-value from a test of equality of the coefficients between high-skill and low-skill judges.

We find that the high- and low-skill judges use the observable information that is available to both the judges and the algorithm in a remarkably similar way, even within these narrow risk score bins. While both types of judges do place additional weight on certain characteristics, they generally do so in the same way. For example, high- and low-skill judges are both about 2 percentage points more likely to release defendants charged with a drug offense and about 4 percentage points less likely to release male defendants. Most strikingly, high- and low-skill judges are both about 14 percentage points more likely to release defendants who are currently on probation or parole. There are only two observable characteristics where there is a statistically significant difference between high- and low-skill judges at the 10% level – about what we would expect by chance since we examine fifteen characteristics in total. The first is for traffic charges, where high-skill judges are 5.8 percentage points more likely to release individuals charged with a traffic offense versus other offenses compared to 8.1 percentage points more likely for low-skill judges. The second is for white defendants, where high-skill judges are 0.2 percentage points less likely to release white individuals versus non-white individuals compared to 1.7 percentage points more likely for low-skill judges. We also see a similar pattern in Appendix Table A.5 for lenient and harsh overrides, with very few statistically significant differences between the high- and low-skill judges.

The results from this section suggest that high- and low-skill judges use the observable information available to both them and the algorithm in a similar way, both across and within narrow risk score bins. Rather, these findings indicate that the high- and low-skill judges may instead differ in their use the private information that is not available to the algorithm. We next decompose the judges' performance differences into components that are and are not explained by the predictable use of the observable information that is available to both the judges and the algorithm to better understand and interpret these findings.

V.B Use of Private Information

The second explanation for our results is that the high- and low-skill judges differ in their use the private information that is not available to the algorithm ($\mathbf{V}_{i,j}$). We provide two sets of results to support this explanation and to better understand the potential importance (or lack thereof) of the judges' use of observable information ($\mathbf{X}_i$) versus private information.

We start by building a new algorithm that predicts the judges' release decisions using all of the information that is observable to both the judges and the original algorithm, following Kleinberg et al. (2018). We predict the high- and low-skill judges' release decisions separately using all of the observable information in $\mathbf{X}_i$, with these new predictions denoted by $\hat{p}_i^J(a(\mathbf{X}_i), D_{i,s}, \mathbf{X}_i)$. We then use the predicted release decisions of the high- and low-skill judges to construct counterfactual decision rules for each judge, where we rank order defendants in terms of their predicted probability of release and define judge-specific thresholds $\bar{p}(j)$

that yield each judge's original release rate. The new decision rule based on the release predictions is given by $\hat{p}_i^J \geq \bar{p}(j)$. Next, we estimate the counterfactual misconduct under the predicted release decisions of each judge, $M_{\bar{p}(j)} = E[Y_i^* | \hat{p}_i^J \geq \bar{p}]$, using the extrapolation approach described above, again separately for high- and low-skill judges. These extrapolations allow us to construct the misconduct rate under the high and low-skill predicted judge release rule, $M_{\bar{p}(j)}$, holding fixed each judge j's release rate. The results from this exercise tell us how the high- and low-skill judges would have performed if they had simply followed their own release tendencies using observable information (and/or private information correlated with this observable information set). We can then use the difference between the judges' actual performance and their predicted release decisions to shed light on the importance of privation information, which presumably drives most of the deviations from the predicted release tendencies.

Figure 7 presents the conditional misconduct rate under the high and low-skill predicted judge release rules. Each of the 62 judges in our sample is again represented by a green dot that shows the judge's conditional misconduct rate $\hat{M}_j$ against the judge's release rate for all cases $\hat{R}_j$, regression adjusted for shift-by-time fixed effects. The solid red line shows the estimated conditional misconduct rate for high-skill predicted release decisions, while the solid blue line shows the estimated conditional misconduct rate for low-skill predicted release decisions. We also include the dashed orange line showing the estimated conditional misconduct rate for the algorithm at different release rates for comparison. The misconduct rate under high-skill and low-skill predicted release rules is modestly higher than the misconduct rate under the original algorithm, suggesting that both types of judges make modest but predictable errors in their use of observable information that leads to some underperformance compared to the algorithm. But most importantly, the predicted decision rules yield nearly identical misconduct rates across the entire observed release distribution. These results again indicate that the high- and low-skill judges use the observable information available to both them and the algorithm in a very similar way, consistent with our findings above.

We can understand the role of private information by comparing the judges' actual decisions to the predicted decision rules as deviations are likely driven by some sort of private information. For example, a judge could choose to deviate from predicted release tendencies based on valuable private information, such as relevant mitigating evidence presented by defense counsel regarding a defendant's ties to the community. A judge could also choose to deviate due to noise, such as extraneous factors not related to the case at hand like what has happened in a previous case or sympathetic characteristics of a defendant. Figure 7 shows that the high-skill judges consistently outperform their predicted judge release rule, suggesting that these judges are able to productively use private information that is not available to the algorithm to improve the accuracy of their decisions. By comparison, the low-skill judges underperform their predicted judge release rule, suggesting that these judges are instead adding noise and inconsistency to their decisions when they attempt to use such additional information.

Our second set of results uses the predicted release decisions described above to decompose judge performance compared to the algorithm into two broad components. The first component captures performance differences that are predicted by the observable characteristics $\mathbf{X}_i$, such as the systematic over- or underweighting of particular observable characteristics relative to the algorithmic risk score. The second component instead captures performance differences that are not predicted by the observable characteristics $\mathbf{X}_i$, such as the use of private information that is not available to the algorithm $\mathbf{V}_{i,j}$.[7]

Table 4 reports the results from the decomposition procedure. Column 1 reports results for all judges, column 2 for high-skill judges, and column 3 for low-skill judges. The first row in each column reports the average difference between the observed conditional misconduct rate for each judge and the counterfactual misconduct rates under the algorithm at the same release rate. The second row reports the share of this difference due to predictable performance differences from, for example, the systematic over- or underweighting of observable information. The third row reports the share of this difference due to non-predictable performance differences from, for example, the use of private information that is observable to the judge but not observable to the algorithm. Standard errors come from a bootstrap procedure where we first take random draws from the distributions of the estimated judge-specific release rates and conditional misconduct rates and then recalculate the threshold-specific extrapolations and statistics of interest. All estimates adjust for shift-by-time fixed effects.

We find that the judges make modest errors in their predictable use of information that leads to some underperformance compared to the algorithm, consistent with the descriptive patterns described above. We see, for example, that the judges increase conditional misconduct rates by 2.4 percentage points relative to the algorithm on average, but only 0.4 percentage points (14%) of this higher misconduct rate is explained by predictable differences in the use of observable information or private information that is correlated with observable information. The remaining 2.1 percentage points (86%) is instead explained by the non-predictable and unproductive use of information. We see similar patterns when examining high- and low-skill judges separately. High-skill judges decrease conditional misconduct rates by 1.5 percentage points relative to the algorithm on average, with 1.9 percentage points (more than 100%) explained by the non-predictable use of information. Low-skill judges instead increase conditional misconduct rates by 3.5 percentage points relative to the algorithm on average, with 3.1 percentage points (89%) explained by the non-predictable use of information.

The results from this section confirm that high- and low-skill judges use the observable information available to both them and the algorithm in a very similar way, suggesting that the key difference between the high- and low-skill judges may be how they use private information that is not available to the algorithm. We will next explore one particular way that private information may add inconsistency and noise to release decisions.

V.C An Example of Unhelpful Private Information

Our final set of results examines the judges' reactions to a highly-salient but largely uninformative event to better understand one particular way that private information can lead to inconsistency and noise. We

[7] We caution that the mapping from our decomposition procedure to our conceptual framework is only approximate since we cannot fully disentangle performance differences due to observable information versus private information. The predictable component from the decomposition procedure includes the systematic use of private information $\mathbf{V}_{i,j}$ that is correlated with the observable information $\mathbf{X}_i$, while the non-predictable component omits such systematic use of private information $\mathbf{V}_{i,j}$ that is correlated with the observable information $\mathbf{X}_i$. We nonetheless view the decomposition procedure as helpful in providing a general understanding of why judge performance may differ from the algorithm.

focus on hearings held just after a case where a different and completely unrelated defendant is arrested for a violent first-degree felony while on pretrial release. First-degree felonies are the highest grade and severity offense, consisting primarily of homicides, rapes and sexual assaults, aggravated assaults, and kidnappings. These are highly-salient adverse events for a bail judge, but arguably uninformative on misconduct risk given both the rarity of such events and the fact that the judge assigned to the case is generally not the judge who initially released the defendant.

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

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

where $D_{i,j,t}$ is an indicator variable for pretrial release in an unrelated case i assigned to judge j in shift t, and $K_{j,t}$ is an indicator denoting the number of shifts since the adverse event, ranging from $k = -4$ to $k = 4$. $\mathbf{U}_i$ is a vector of observable case and defendant characteristics (including $\mathbf{X}_i$ and a subset of $\mathbf{V}_{i,j}$), $\mathbf{W}_i$ is a vector of shift-by-time effects, and α_j are judge fixed effects. The coefficients of interest are β_k, which measure the probability of release for cases heard in the four shifts before and four shifts after the adverse event relative to the omitted shift at $k = -1$. We assign cases in the same shift as the adverse event to the omitted shift and bin cases outside the focal shifts into separate indicators that we do not report. We estimate the event-study specification using a balanced panel of 59 judges, including 9 judges who never experienced an adverse event (who are assigned to the omitted shift) and 50 judges who we observe for at least five shifts before and after the first time they experience an adverse event. We focus our main results on the first observed adverse event. Standard errors are clustered at the judge level.

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. We start with the full sample of cases in Panel A, where we observe a sharp decline in pretrial release rates immediately after the adverse event. The response is largest in the second shift after the event, with release rates returning to baseline levels by the fourth shift after the event. The magnitude of the response is substantial, with a 5.0 percentage point decrease in pretrial release rates over the first four shifts following the adverse event (a 6% decrease from the mean). In Panel B, we show that these effects are driven by the low-skill judges, whose release rates decrease by 4.5 percentage points (a 5% decrease from the mean) following the adverse event compared to a statistically insignificant increase of 0.6 percentage points (a 1% increase from the mean) for high-skill judges. We see these results as consistent with low-skill judges being more likely to use private information in a way that leads to inconsistency and noise.

We also explore heterogeneity in the judges' reactions to better understand this behavioral response. Panel A of Appendix Figure A.14 presents results separately by defendant race, as non-white defendants are particularly representative of those arrested for serious violent crimes while on pretrial release (Kahneman and Tversky, 1972; Bordalo et al., 2016) and prior work documents substantial racial discrimination in bail decisions (Arnold, Dobbie, and Yang, 2018; Arnold, Dobbie, and Hull, 2021). We find that non-white defendants are 9.3 percentage points less likely to be released after an adverse event (an 11% decrease from the mean), compared to only 0.9 percentage points for white defendants (a 1% decrease from the mean). Panel

B of Appendix Figure A.14 similarly shows that defendants currently on probation or parole are 7.9 percentage points less likely to be released after an adverse event (a 9% decrease from the mean), compared to only 3.7 percentage points for defendants not on probation or parole (a 4% decrease from the mean). Recall that the judges substantially overweight this characteristic relative to the risk score (Table 3). Taken together, these results suggest that the reactions are concentrated among defendants that are either particularly representative of those arrested for serious violent crimes (Kahneman and Tversky, 1972; Bordalo et al., 2016) or with observable characteristics that are particularly overweighted by judges (Bordalo, Gennaioli, and Shleifer, 2015; Sunstein, 2022).

We view these patterns as overreactions to a highly-salient but largely uninformative event, rather than a rational updating of beliefs. We show in Appendix Figure A.15 that the judges' reactions are concentrated among observably low-risk defendants that the algorithm recommends releasing, such that judges increase their probability of harsh overrides by 5.1 percentage points (a 45% increase from the mean) after the adverse event. We also show in Appendix Figure A.16 that conditional misconduct rates decrease by only a statistically insignificant 1.0 percentage points following an adverse event (a 4% decrease from the mean), despite the large decrease in release rates documented above. Both results, as well as the fact that judges' release rates return to baseline levels relatively soon after the adverse event, are inconsistent with most models of rational updating but consistent with a behavioral response. For example, the judges in our setting may be particularly prone to the "availability" heuristic described in Tversky and Kahneman (1974), where they overreact to recent and salient events that are "top of mind" (Bordalo, Gennaioli, and Shleifer, 2013; Bordalo et al., 2016; Sunstein, 2022).[8]

VI Conclusion

This paper shows that there is substantial variation in the effects of human discretion over an algorithm 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. But the remaining 10% outperform the algorithm and significantly decrease pretrial misconduct compared to an algorithmic counterfactual. These performance differences are most likely driven by how the judges use private information that is unavailable to the algorithm, with high-skill judges using such information to improve the accuracy of their decisions and low-skill judges only adding inconsistency and noise when they attempt to use such information.

Our findings suggest that there will not necessarily be a single correct policy on human oversight of algorithms, as the impact of such policies depends on the ability of the human decision-makers in a particular context. The quasi-experimental methods developed in this paper may also prove useful in measuring the impact of human oversight and discretion in other high-stakes settings. Our approach to measuring the impact of human discretion is appropriate whenever there is the quasi-random assignment of decision makers

[8] Similar overreactions impact macroeconomic expectations (Bordalo et al., 2020), medical treatment decisions (Choudhry et al., 2006; Singh, 2021), coverage of crime and judicial errors (Philippe and Ouss, 2018), and sentencing reversals on appeal (Bhuller and Sigstad, 2021). As one such example, Singh (2021) finds that experiencing adverse obstetric events in one delivery mode (such as C-section versus vaginal delivery) makes the physician more likely to switch to the other delivery mode on the next patient, regardless of patient indication, resulting in worse patient outcomes and inefficient resource use.

and the objective of these decision makers is both known and well-measured. Our test can therefore be used to explore the impact of human discretion in other settings where algorithms are widely used, such as hiring, lending, and medical testing decisions.

References

Albright, Alex. 2021. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions." *Unpublished Working Paper.*

Anwar, Shamena, Shawn D. Bushway, and John Engberg. 2022. "The Impact of Defense Counsel at Bail Hearings." RAND Working Paper No. WR-A1960-1.

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

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

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

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

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research*, 1–42.

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

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

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

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

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

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

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

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

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

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

De-Arteaga, Maria, Riccardo Fogliato, and Alexandra Chouldechova. 2020. "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

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

Frandsen, Brigham R., Lars J. Lefgren, and Emily C. Leslie. 2019. "Judging Judge Fixed Effects." *NBER Working Paper No. 25528.*

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

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. "Simple Rules for Complex Decisions." SSRN Working Paper No. 2919024.

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

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

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

Ludwig, Jens, and Sendhil Mullainathan. 2022. "Algorithmic Behavioral Science: Machine Learning as a Tool for Scientific Discovery." Chicago Booth Working Paper No. 22-15.

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

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

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

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

Rambachan, Ashesh. 2021. “Identifying Prediction Mistakes in Observational Data.” *Unpublished Working Paper*.

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

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

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

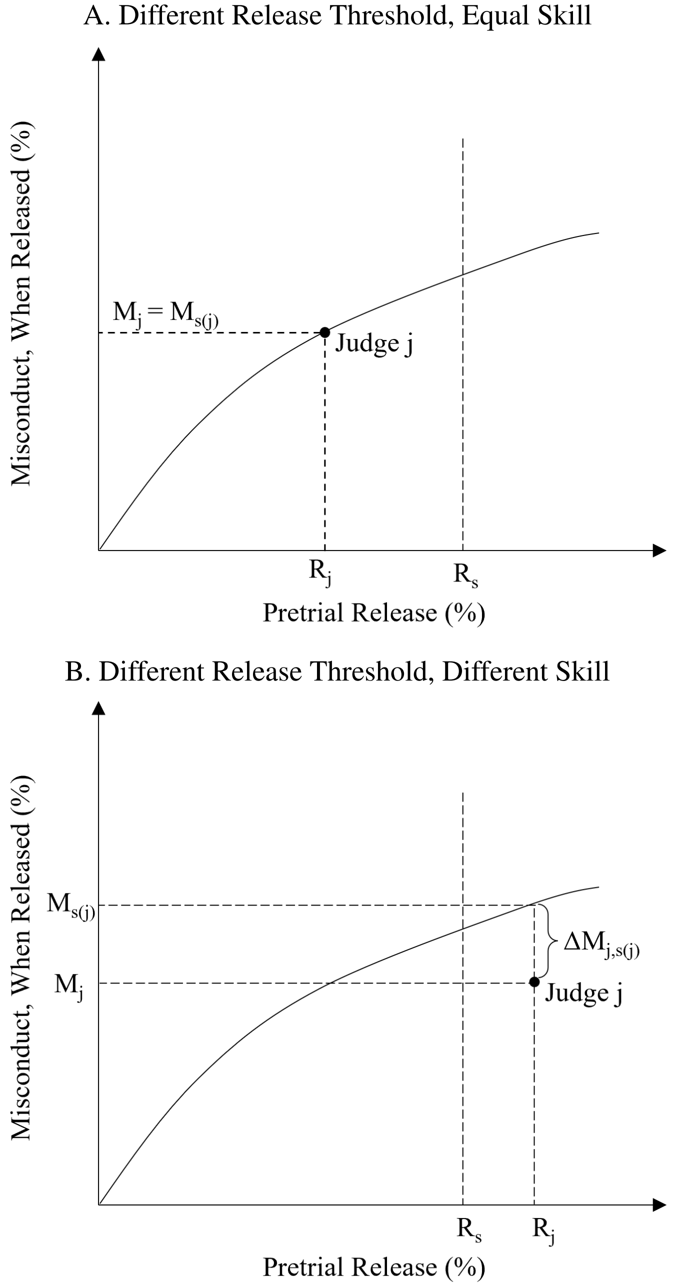
Stevenson, Megan, and Jennifer Doleac. 2021. “Algorithmic Risk Assessment in the Hands of Humans.” *Unpublished Working Paper*.

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

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

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

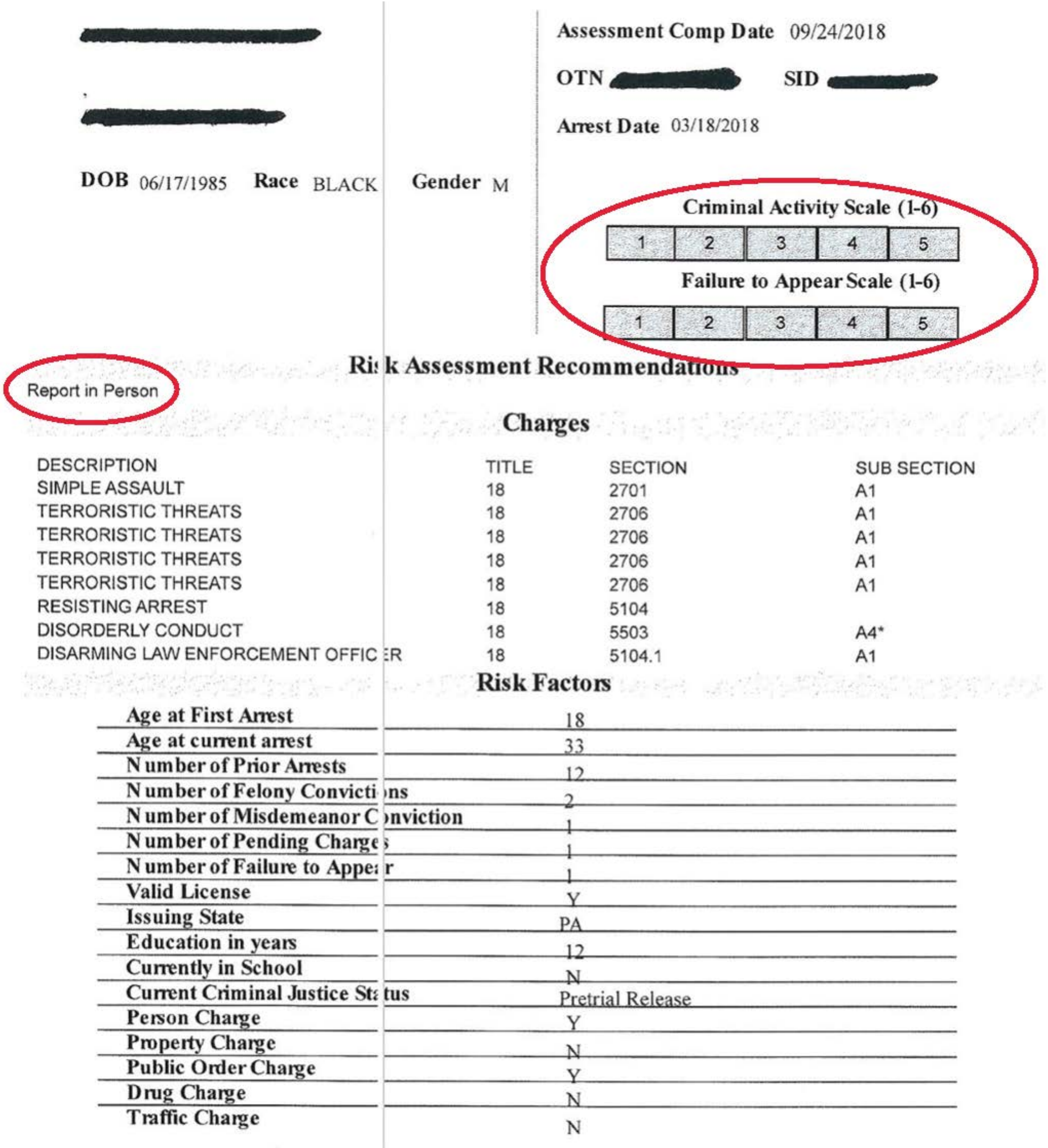
Figure 1: Hypothetical Variation in Release Thresholds and Predictive Skill


A. Different Release Threshold, Equal Skill

B. Different Release Threshold, Different Skill

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

Figure 2: Example Risk Assessment



[redacted]

[redacted]

DOB 06/17/1985 **Race** BLACK **Gender** M

Assessment Comp Date 09/24/2018

OTN [redacted] **SID** [redacted]

Arrest Date 03/18/2018

Criminal Activity Scale (1-6)

1	2	3	4	5

Failure to Appear Scale (1-6)

1	2	3	4	5

Risk Assessment Recommendations

Report in Person

Charges

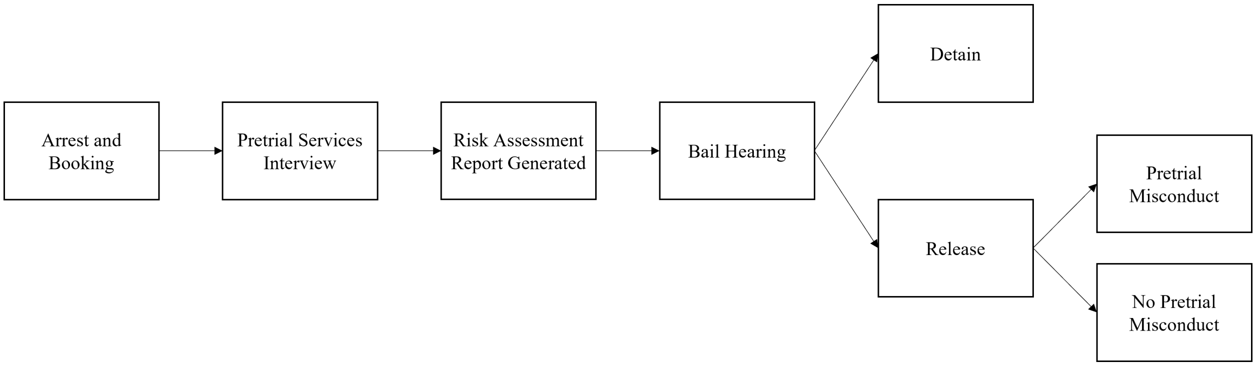
DESCRIPTION	TITLE	SECTION	SUB SECTION
SIMPLE ASSAULT	18	2701	A1
TERRORISTIC THREATS	18	2706	A1
TERRORISTIC THREATS	18	2706	A1
TERRORISTIC THREATS	18	2706	A1
TERRORISTIC THREATS	18	2706	A1
RESISTING ARREST	18	5104	
DISORDERLY CONDUCT	18	5503	A4*
DISARMING LAW ENFORCEMENT OFFICER	18	5104.1	A1

Risk Factors

Age at First Arrest	18
Age at current arrest	33
Number of Prior Arrests	12
Number of Felony Convictions	2
Number of Misdemeanor Conviction	1
Number of Pending Charges	1
Number of Failure to Appear	1
Valid License	Y
Issuing State	PA
Education in years	12
Currently in School	N
Current Criminal Justice Status	Pretrial Release
Person Charge	Y
Property Charge	N
Public Order Charge	Y
Drug Charge	N
Traffic Charge	N

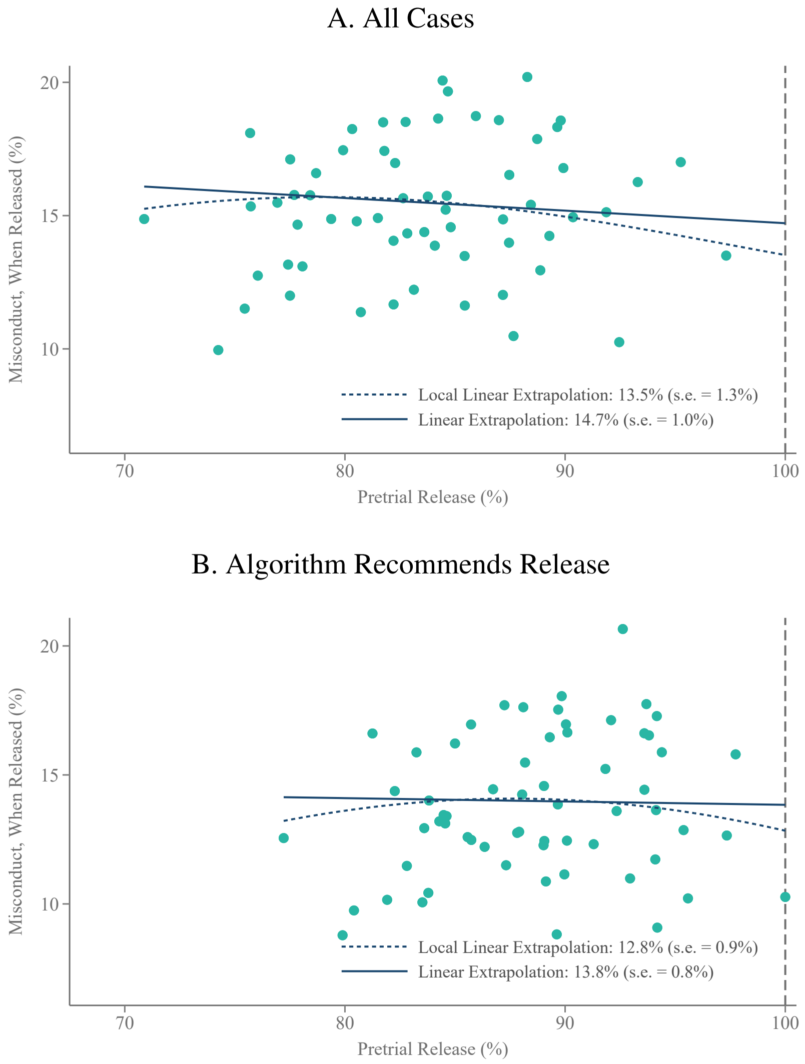
Notes. This figure shows an example risk assessment report in our setting. We add red ovals around the risk assessment scores and algorithmic recommendation. The example risk assessment report also includes the defendant's arrest date, date of birth, race, gender, description of the charges, and a description of the relevant risk factors entering the algorithmic risk scores and recommendation.

Figure 3: Pretrial Process



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

Figure 4: Extrapolations of Release and Conditional Misconduct Rates

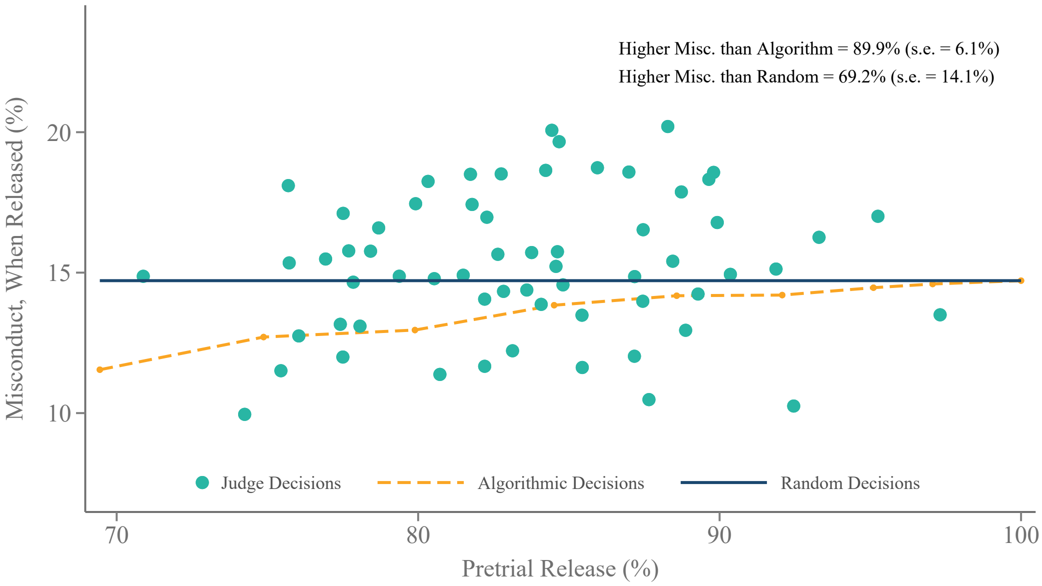


A. All Cases

B. Algorithm Recommends Release

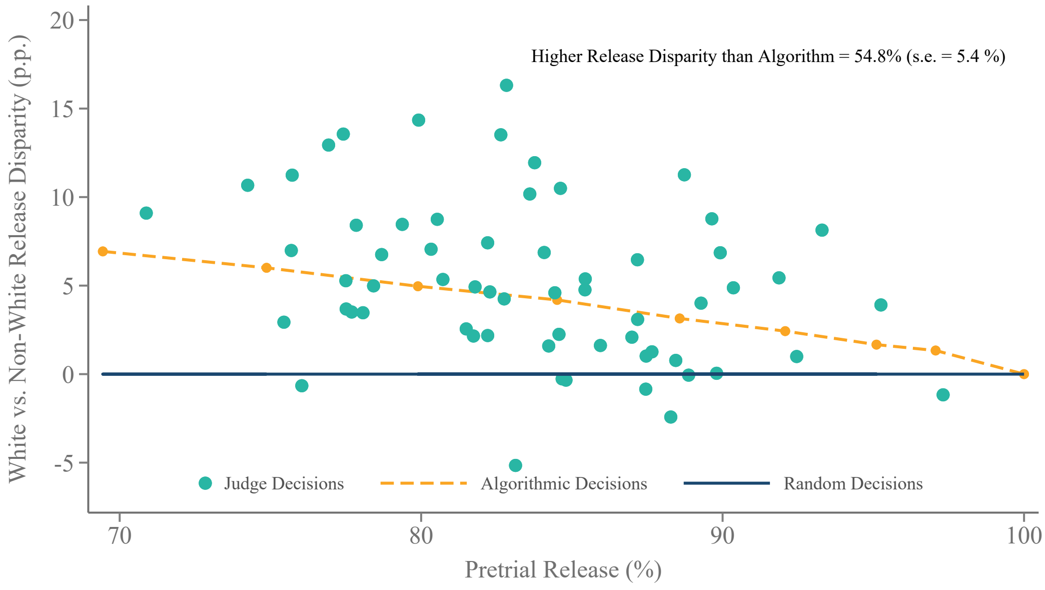
Notes. This figure plots judge release rates against misconduct rates among released defendants at two algorithmic risk score cutoffs. Each point represents the mean release and conditional misconduct rate for each judge, adjusted for shift-by-time fixed effects. Panel A reports results for the full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

Figure 5: Judge Discretion and Conditional Misconduct Rates



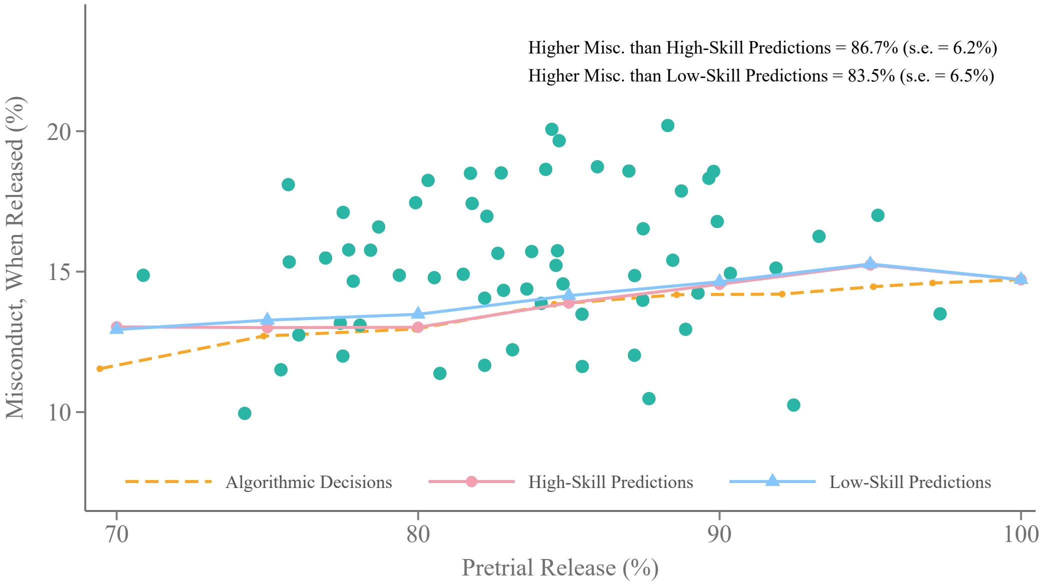
Notes. This figure plots release rates against misconduct rates among released defendants for the 62 judges in our sample, along with counterfactual misconduct rates among released defendants for algorithmic and random decisions. Conditional misconduct rates under the algorithm are estimated using linear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

Figure 6: White vs. Non-White Release Disparities



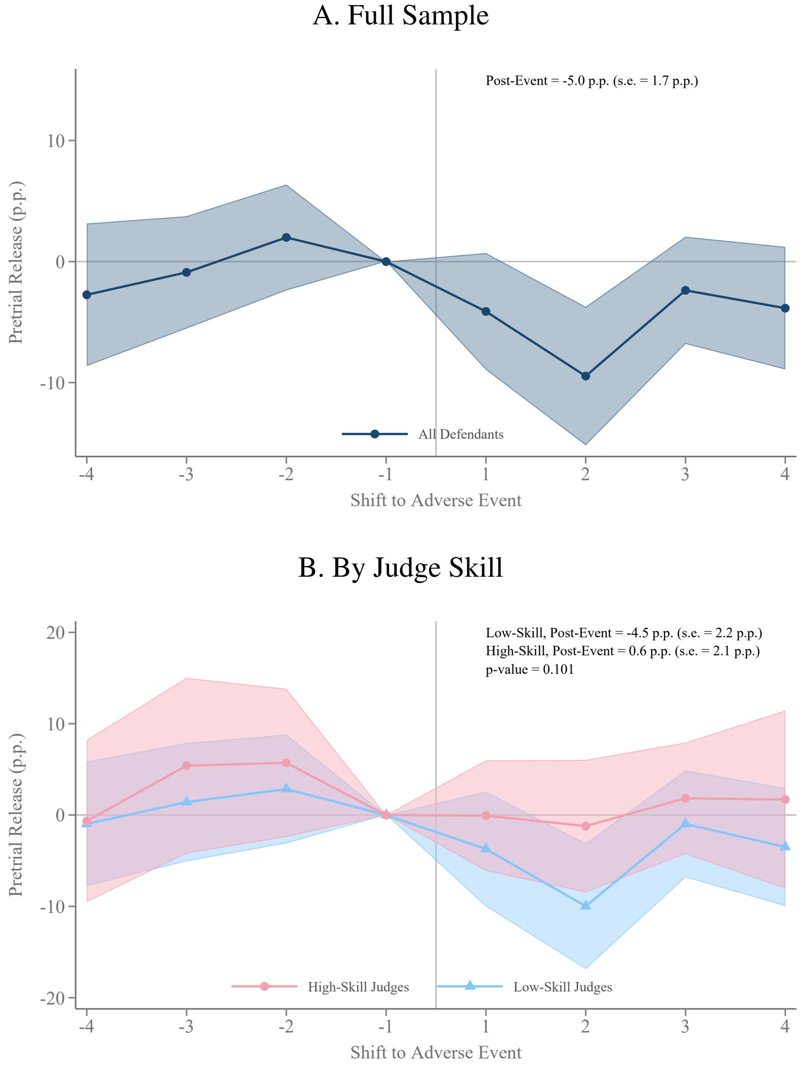
Notes. This figure plots release rates against release disparities between white and non-white defendants with identical misconduct potential, along with counterfactual disparities for algorithmic and random decisions. All estimates are based on a linear extrapolation of the race-specific misconduct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

Figure 7: Predicted Release Decisions



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

Figure 8: Effect of Adverse Events on Pretrial Release



A. Full Sample

B. By Judge Skill

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

Table 1: Descriptive Statistics

		Recommend Detain		Recommend Release	
	All Cases	Lenient Override	Follow Algorithm	Harsh Override	Follow Algorithm
A. Pretrial Release	(1)	(2)	(3)	(4)	(5)
Released Before Trial	0.83	1.00	0.00	0.00	1.00
Share ROR	0.10	0.02	—	—	0.11
Share Non-Monetary	0.52	0.15	—	—	0.56
Share Monetary Bail	0.37	0.82	—	—	0.32
Share Remanded	0.01	0.01	—	—	0.00
B. Algorithmic Inputs					
Age at Current Arrest	34.59	33.02	33.04	35.50	34.80
Age at First Arrest	21.63	16.91	16.59	20.06	22.85
Prior Arrests	9.39	17.86	19.89	12.86	6.96
Prior Felonies	1.39	3.04	3.52	2.10	0.91
Prior Misdemeanors	2.38	4.70	5.26	3.26	1.73
Pending Charges	0.57	1.94	2.07	0.52	0.27
Property Charge	0.20	0.26	0.29	0.26	0.18
Drug Charge	0.28	0.47	0.40	0.27	0.24
Public Order Charge	0.45	0.50	0.57	0.53	0.43
Traffic Charge	0.14	0.28	0.17	0.06	0.14
Parole/Probation	0.27	0.46	0.65	0.52	0.18
Pretrial Release	0.32	0.91	0.89	0.33	0.20
Violent Charge	0.41	0.17	0.20	0.38	0.46
C. Demographics					
Male	0.74	0.84	0.87	0.84	0.71
White	0.44	0.38	0.37	0.39	0.46
D. Pretrial Misconduct, When Released					
Any Misconduct	0.16	0.29	—	—	0.14
Share NCA Only	0.74	0.62	—	—	0.76
Share FTA Only	0.17	0.21	—	—	0.16
Share NCA and FTA	0.09	0.17	—	—	0.08
Cases	37,855	3,142	2,721	3,784	28,208

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

Table 2: Characteristics of High-Skill Judges

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male	-9.89								-1.83
	(15.20)								(17.91)
White		11.58							4.21
		(19.23)							(24.99)
Registered Democrat			-14.51						-13.06
			(13.86)						(14.12)
Above Median Experience				0.73					-3.23
				(11.93)					(13.06)
Above Median Caseload					-3.23				-10.95
					(11.90)				(13.03)
Override Rate (0-100)						-1.44			-2.97
						(1.67)			(2.15)
Release Rate (0-100)							0.05		-1.62
							(1.16)		(1.35)
White vs. Non-White Disparity (0-100)								-16.40	-19.41
								(6.76)	(8.44)
Judges	62	62	62	62	62	62	62	62	62

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

Table 3: Characteristics of Released Defendants

	High-Skill Judges	Low-Skill Judges	p-value
	(1)	(2)	(3)
Age at Current Arrest	0.62	0.25	0.52
	(0.51)	(0.29)	
Age at First Arrest	-0.03	-0.16	0.28
	(0.10)	(0.05)	
Prior Arrests	-0.27	-0.13	0.26
	(0.10)	(0.07)	
Prior Felonies	0.28	-0.15	0.21
	(0.27)	(0.21)	
Prior Misdemeanors	0.10	-0.08	0.57
	(0.28)	(0.15)	
Pending Charges	-0.21	-1.72	0.16
	(0.97)	(0.48)	
Property Charge	-1.38	-0.83	0.78
	(1.85)	(0.67)	
Drug Charge	2.13	2.33	0.86
	(1.04)	(0.47)	
Public Order Charge	-2.81	-3.99	0.28
	(1.01)	(0.41)	
Traffic Charge	5.80	8.09	0.07
	(0.88)	(0.90)	
Parole/Probation	-13.06	-14.70	0.42
	(1.83)	(0.85)	
Pretrial Release	-1.45	1.28	0.16
	(1.68)	(0.94)	
Violent Charge	-1.70	-0.62	0.49
	(1.20)	(1.00)	
Male	-4.36	-4.19	0.86
	(0.84)	(0.42)	
White	-0.19	1.70	0.04
	(0.63)	(0.67)	
Shift x Time FE	Yes	Yes	
Risk Score FE	Yes	Yes	
Cases	7,909	29,946	

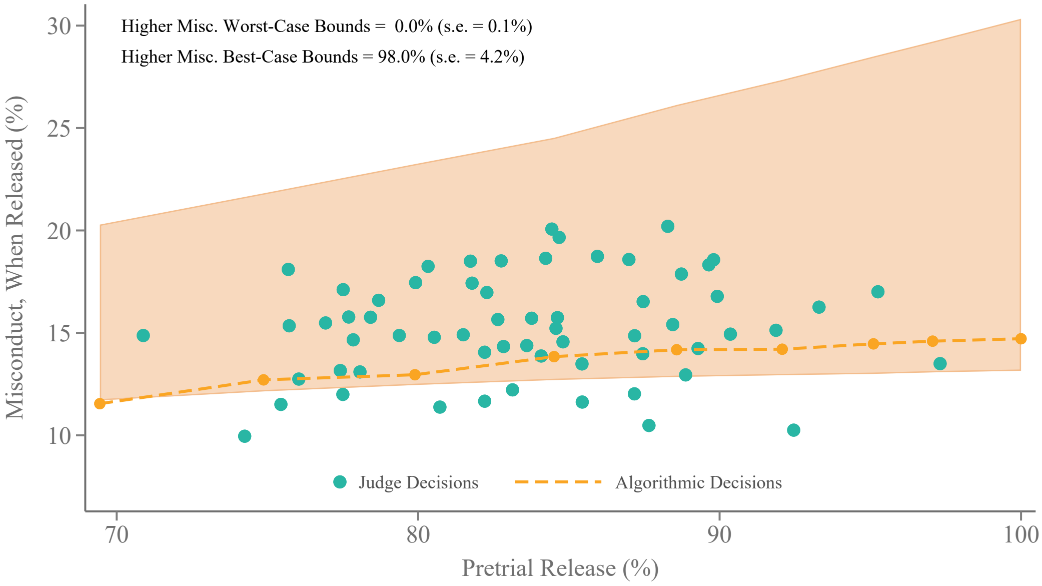
Notes. This table reports OLS estimates of regressions of an indicator for release on case and defendant characteristics with shift-by-time and NCA risk score fixed effects. Column 1 reports results for high-skill judges with lower conditional misconduct rates than the algorithm holding fixed release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

Table 4: Decomposing the Judges' Performance

	All Judges	High-Skill Judges	Low-Skill Judges
	(1)	(2)	(3)
Judge Misconduct vs. Algorithm	2.42	-1.54	3.47
	(0.45)	(0.40)	(0.40)
Predictable Performance Differences	0.35	0.38	0.34
	(0.55)	(0.58)	(0.57)
Non-Predictable Performance Differences	2.07	-1.93	3.13
	(0.48)	(0.65)	(0.54)
Judges	62	19	43

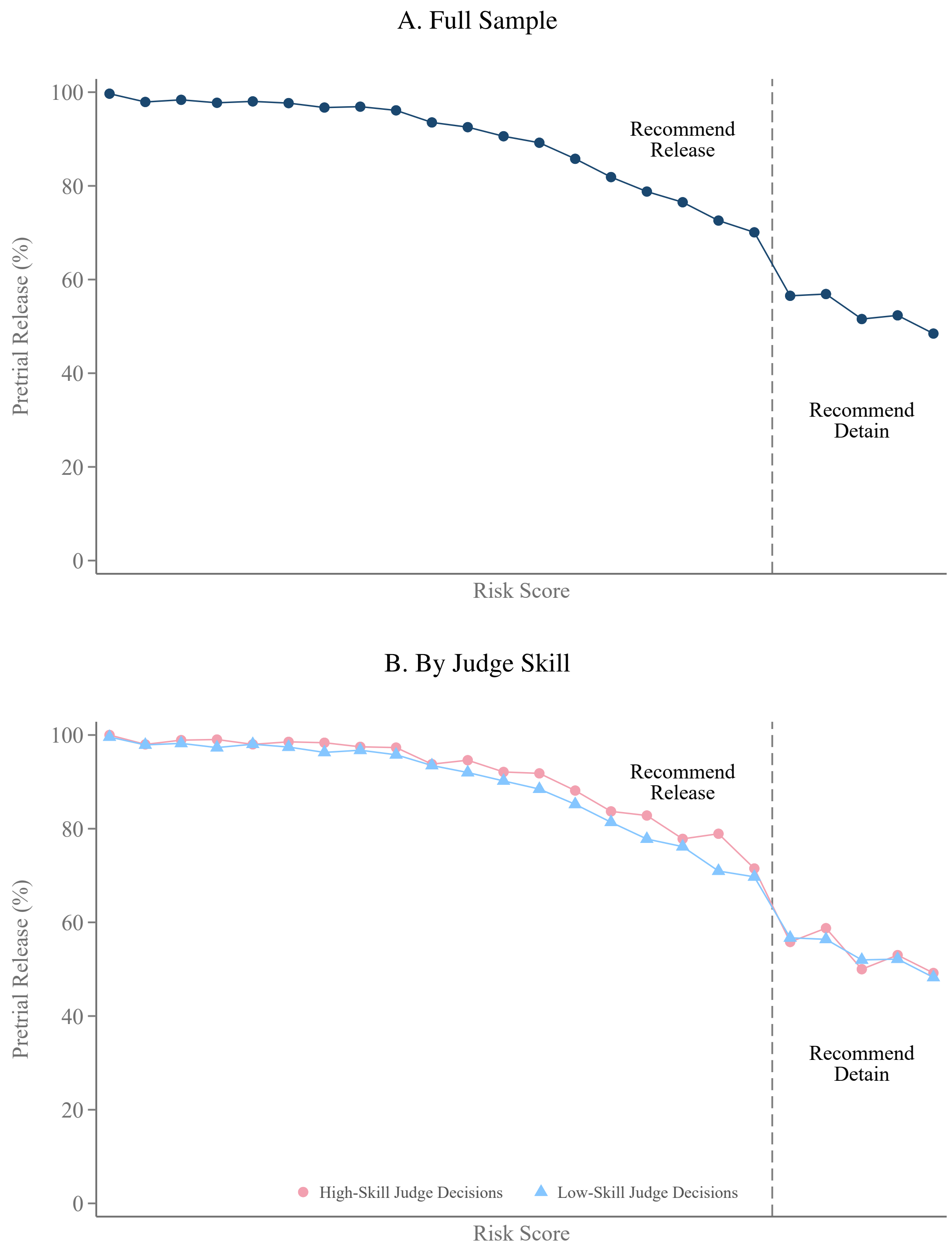
Notes. This table reports estimates decomposing the impact of human discretion on conditional misconduct rates into the share that is predictable based on observable inputs versus non-predictable based on observable inputs. Column 1 reports results for all judges, column 2 for high-skill judges with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

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



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

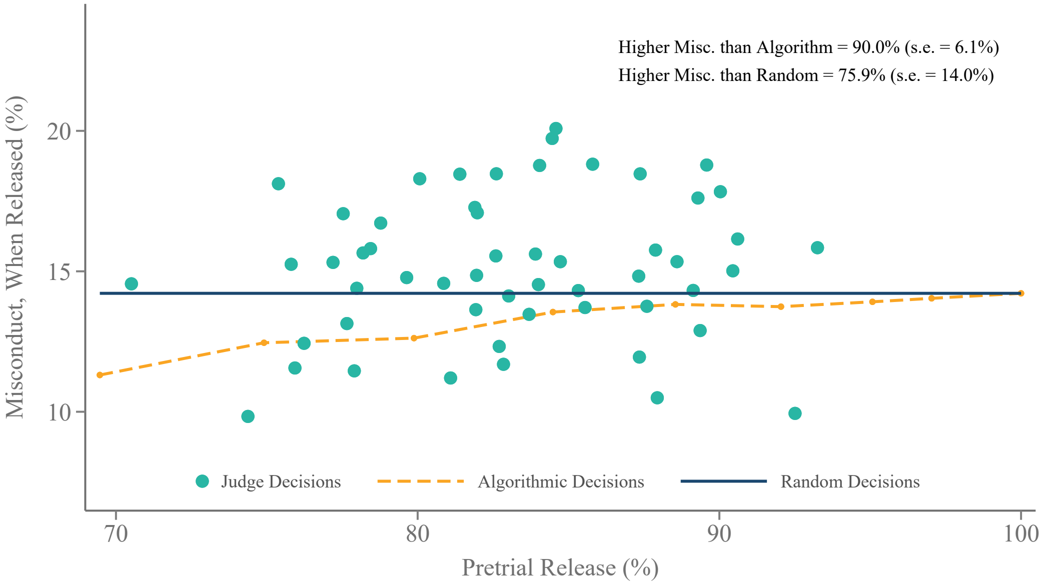
Figure A.2: Pretrial Release Rates by Risk Score


A. Full Sample

B. By Judge Skill

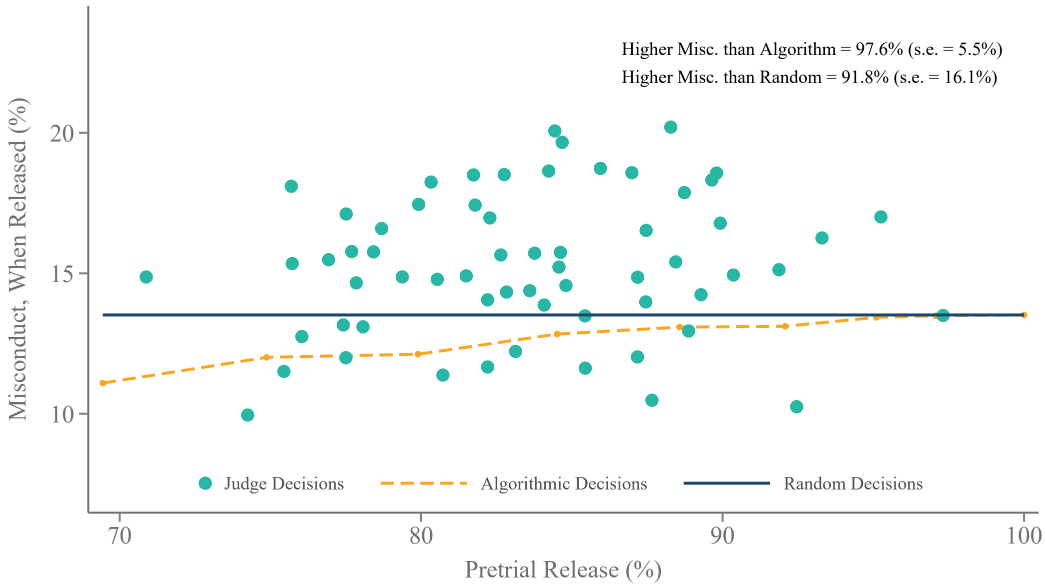
Notes. This figure plots algorithmic risk scores against pretrial release rates, along with a marker for the score at which the algorithmic recommendation changes from release to detain. In Panel A, each point represents one of the discrete risk scores and the corresponding mean pretrial release rate for defendants assigned the risk score. Panel B plots the mean pretrial release rate at each discrete risk score by judge skill.

Figure A.3: Results for Judges with High Caseloads



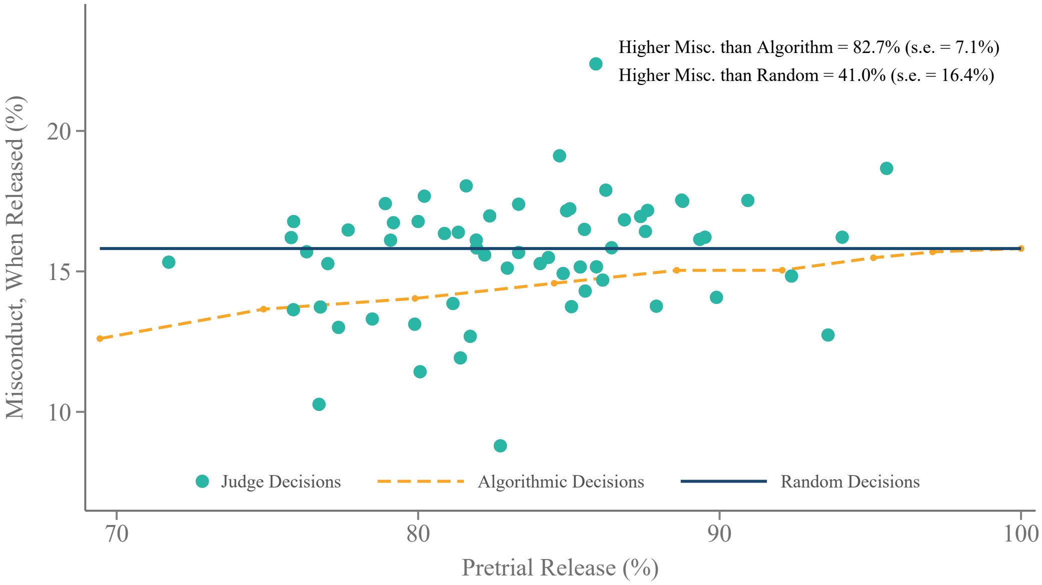
Notes. This figure shows results for the 54 judges in our sample who heard 200 or more cases during the span of our study. See the notes for Figure 5 for additional details.

Figure A.4: Results with Local Linear Extrapolations of Conditional Misconduct Rates



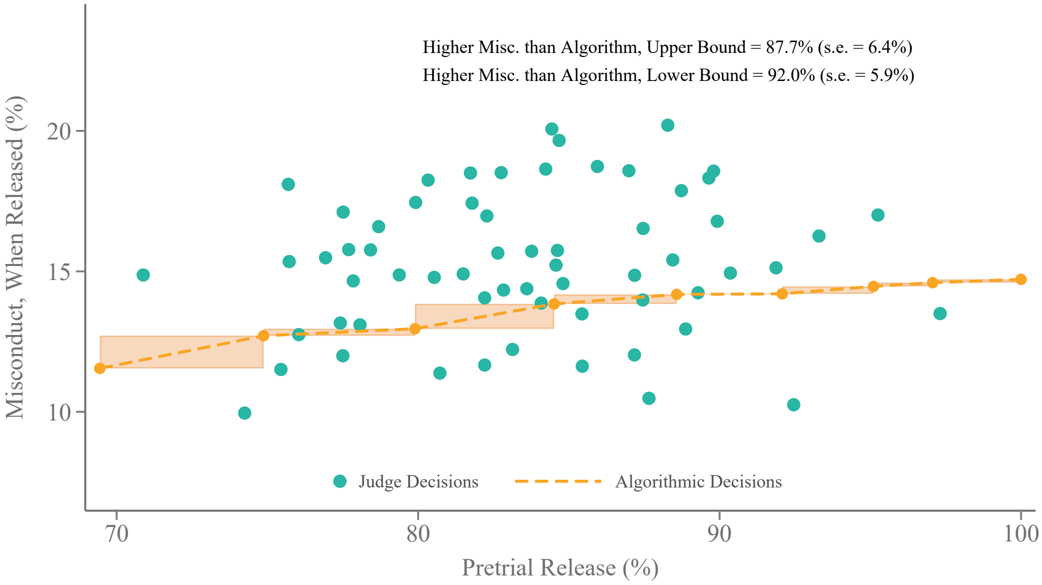
Notes. The figure shows results when we construct counterfactual misconduct rates under the algorithm using local linear extrapolations of mean risk. See the notes for Figure 5 for additional details.

Figure A.5: Results Without Shift-by-Time Effects



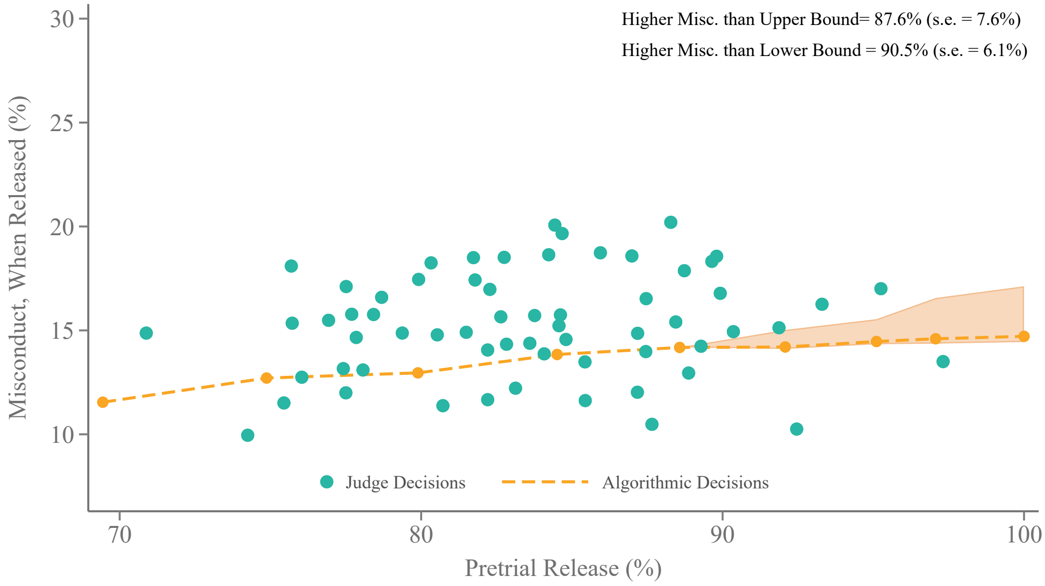
Notes. This figure shows results when we omit shift-by-time effects. See the notes for Figure 5 for additional details.

Figure A.6: Stepwise Connections of Discrete Risk Scores



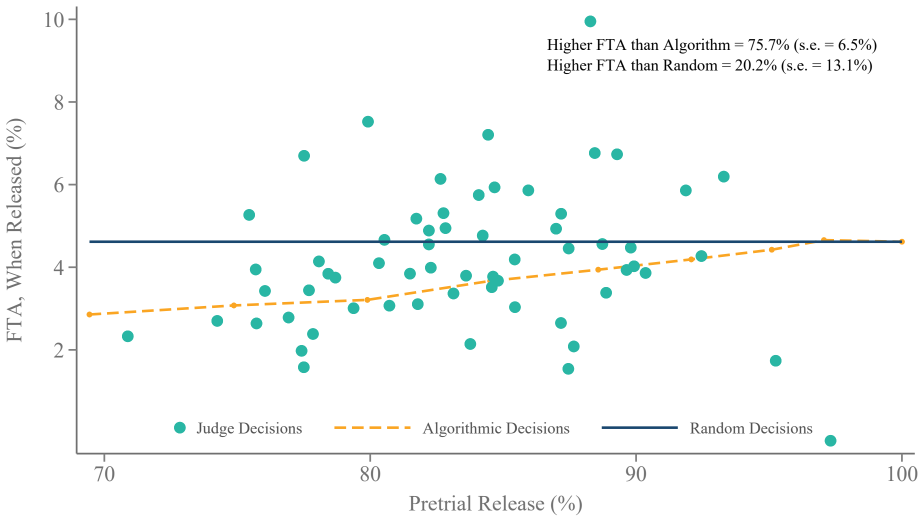
Notes. This figure shows the stepwise connections of the discrete risk scores. See the notes for Figure 5 for additional details.

Figure A.7: Results with Extrapolations to the Most Lenient Judge



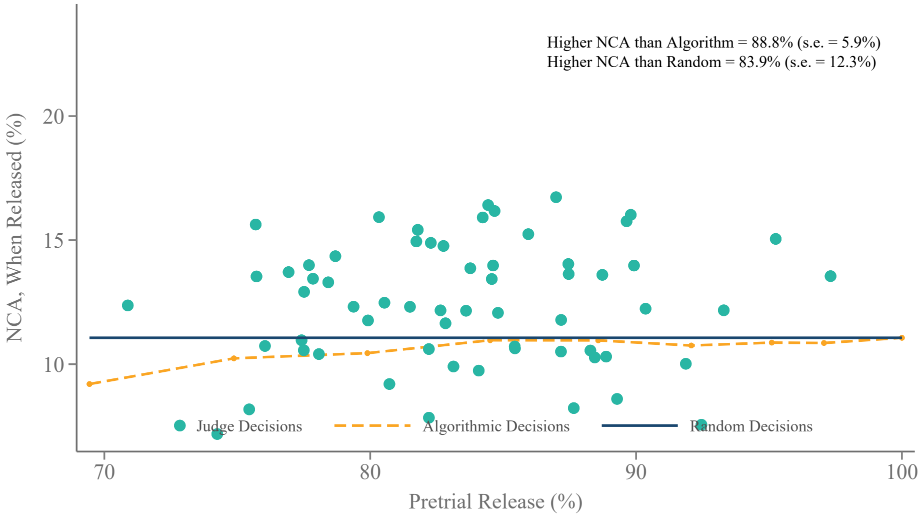
Notes. This figure shows results when we construct counterfactual misconduct rates under the algorithm using extrapolations to only the most lenient judge at each risk score cutoff and then calculate worst- and best-case bounds for the remaining fraction of defendants. The region between the constructed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

Figure A.8: Results for Failure to Appear



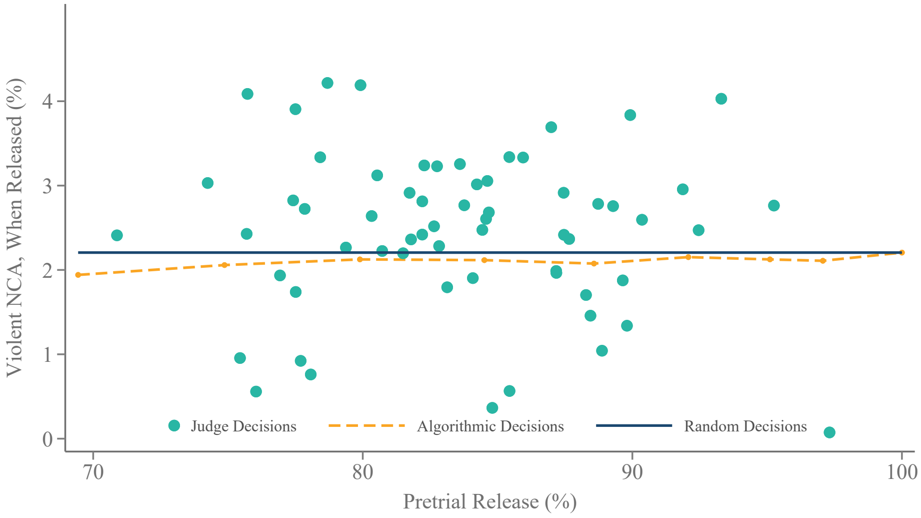
Notes. This figure shows results for failures to appear. See the notes for Figure 5 for additional details.

Figure A.9: Results for New Criminal Activity



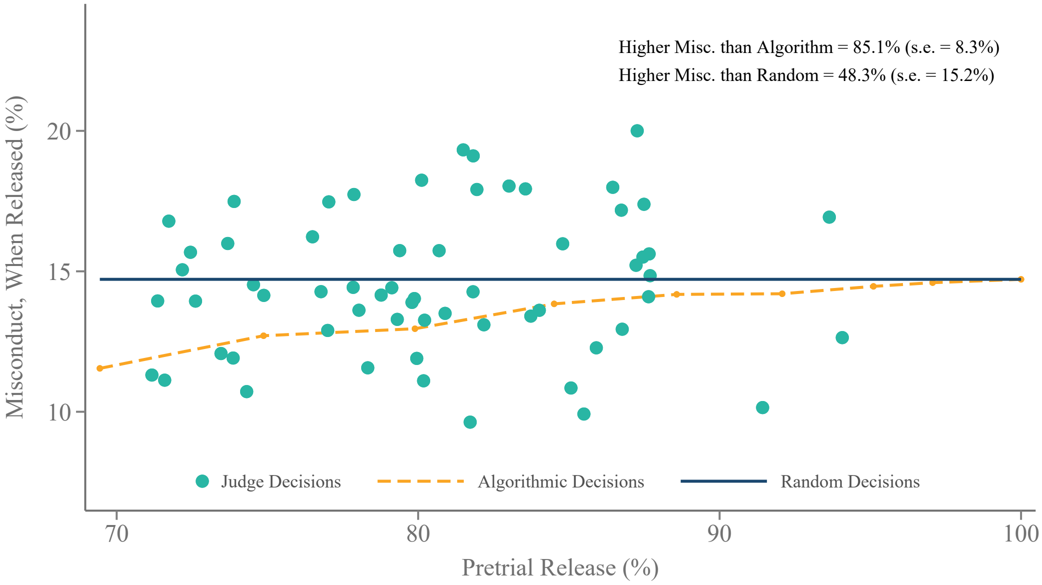
Notes. This figure shows results for new criminal activity. See the notes for Figure 5 for additional details.

Figure A.10: Results for Violent New Criminal Activity



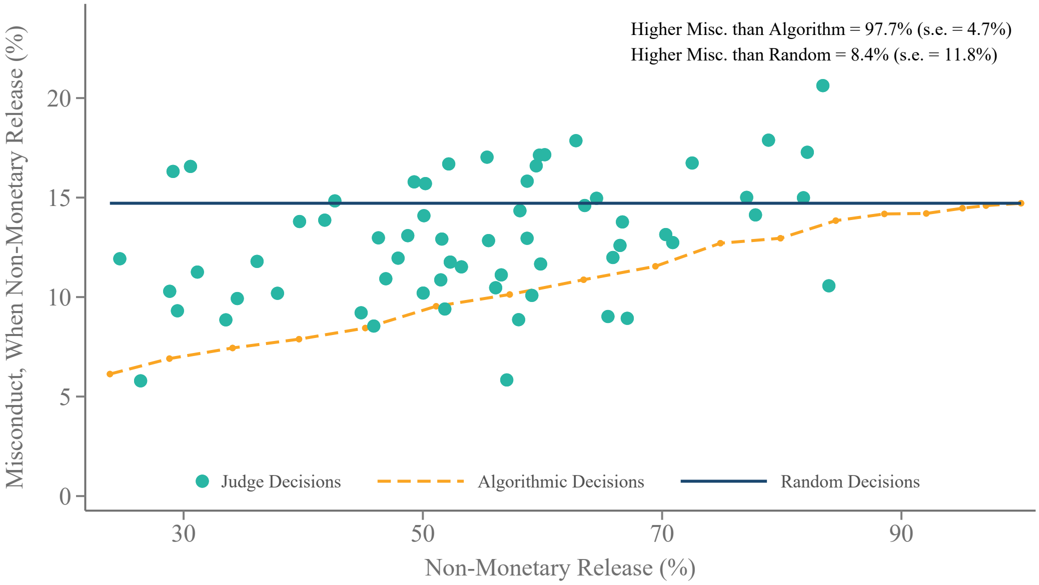
Notes. This figure shows results for violent new criminal activity. See the notes for Figure 5 for additional details.

Figure A.11: Results for Defendants Released Within Three Days



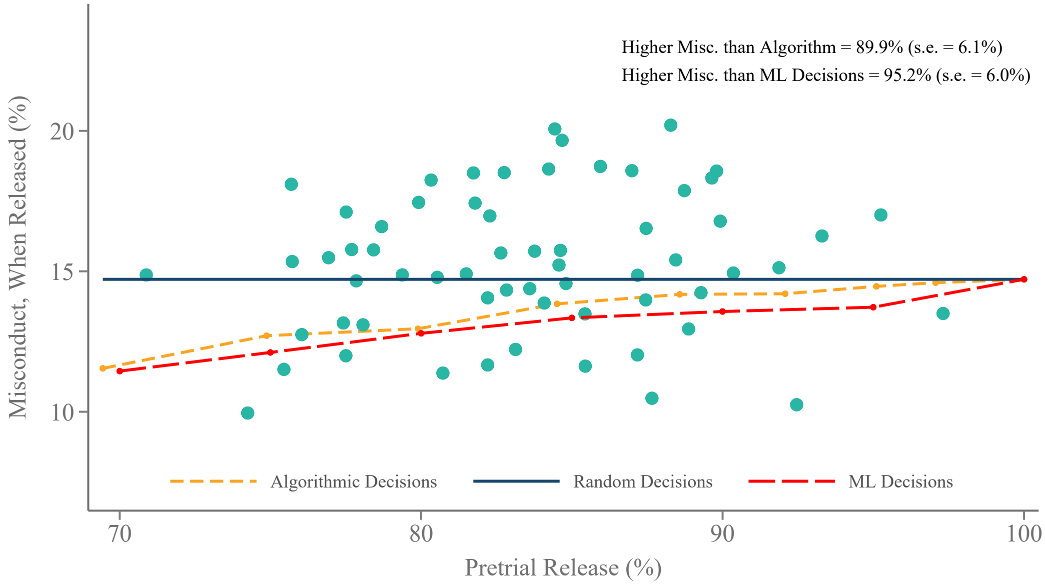
Notes. This figure shows results when we categorize defendants as released if they are released within the first three days of the original bail hearing. See the notes for Figure 5 for additional details.

Figure A.12: Results for Non-Monetary Release



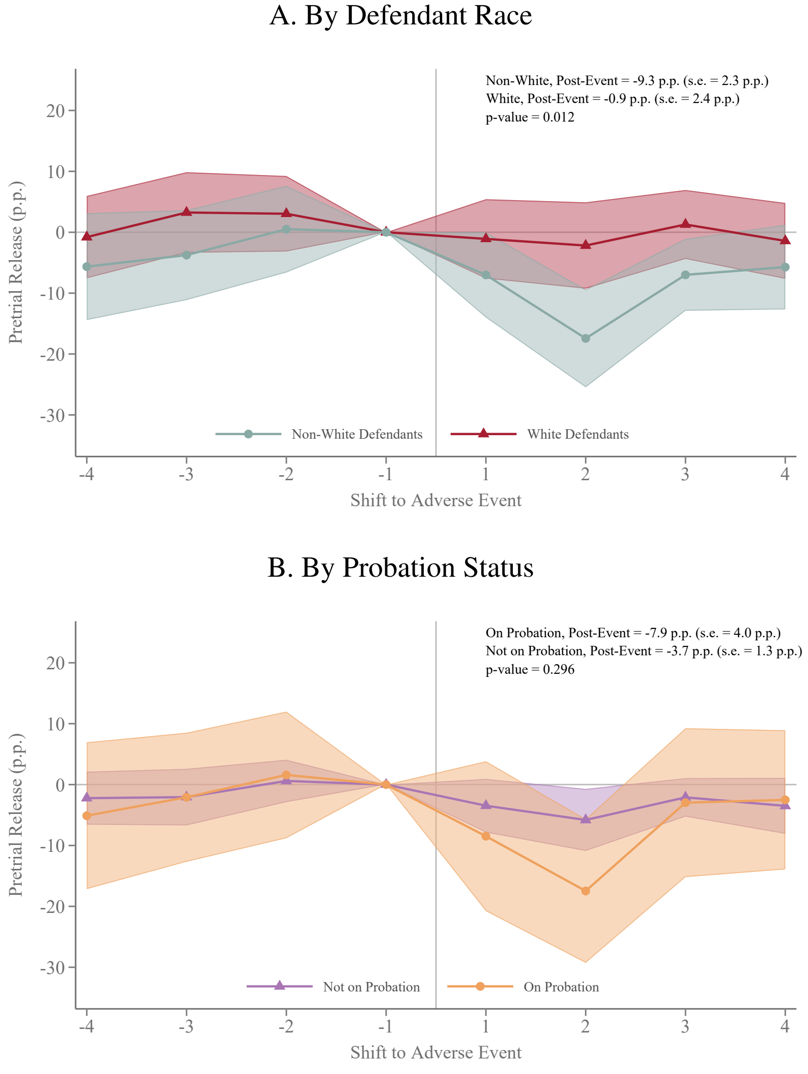
Notes. This figure shows results when we categorize defendants as released with no monetary conditions (including ROR and non-monetary release) versus released with monetary conditions or not released. See the notes for Figure 5 for additional details.

Figure A.13: Comparison to Machine Learning Algorithm



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

Figure A.14: Effect of Adverse Events on Pretrial Release by Defendant Race and Probation Status

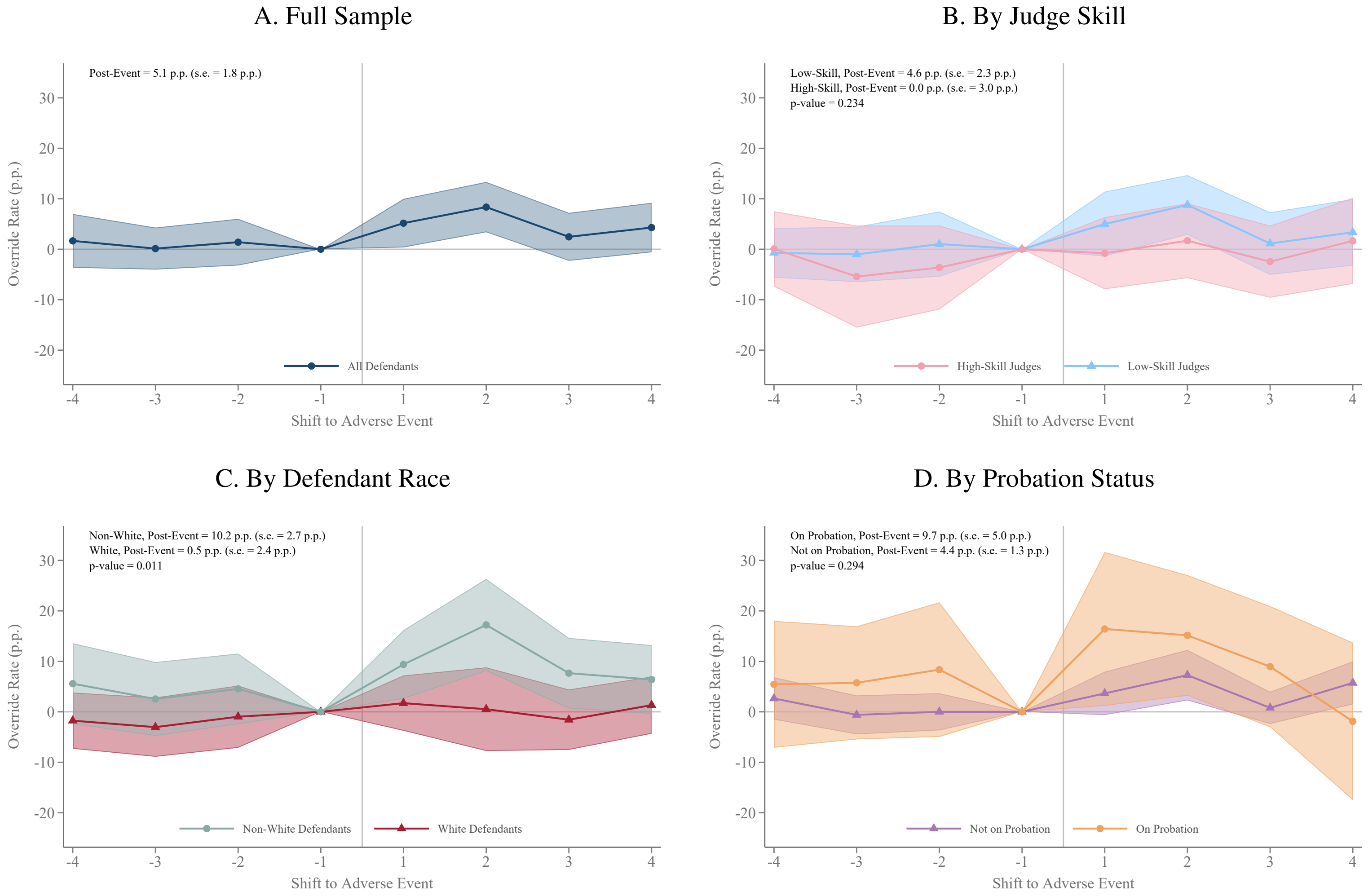


A. By Defendant Race

B. By Probation Status

Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release by defendant race and probation status. See the notes of Figure 8 for additional details.

Figure A.15: Effect of Adverse Events on Harsh Overrides


A. Full Sample

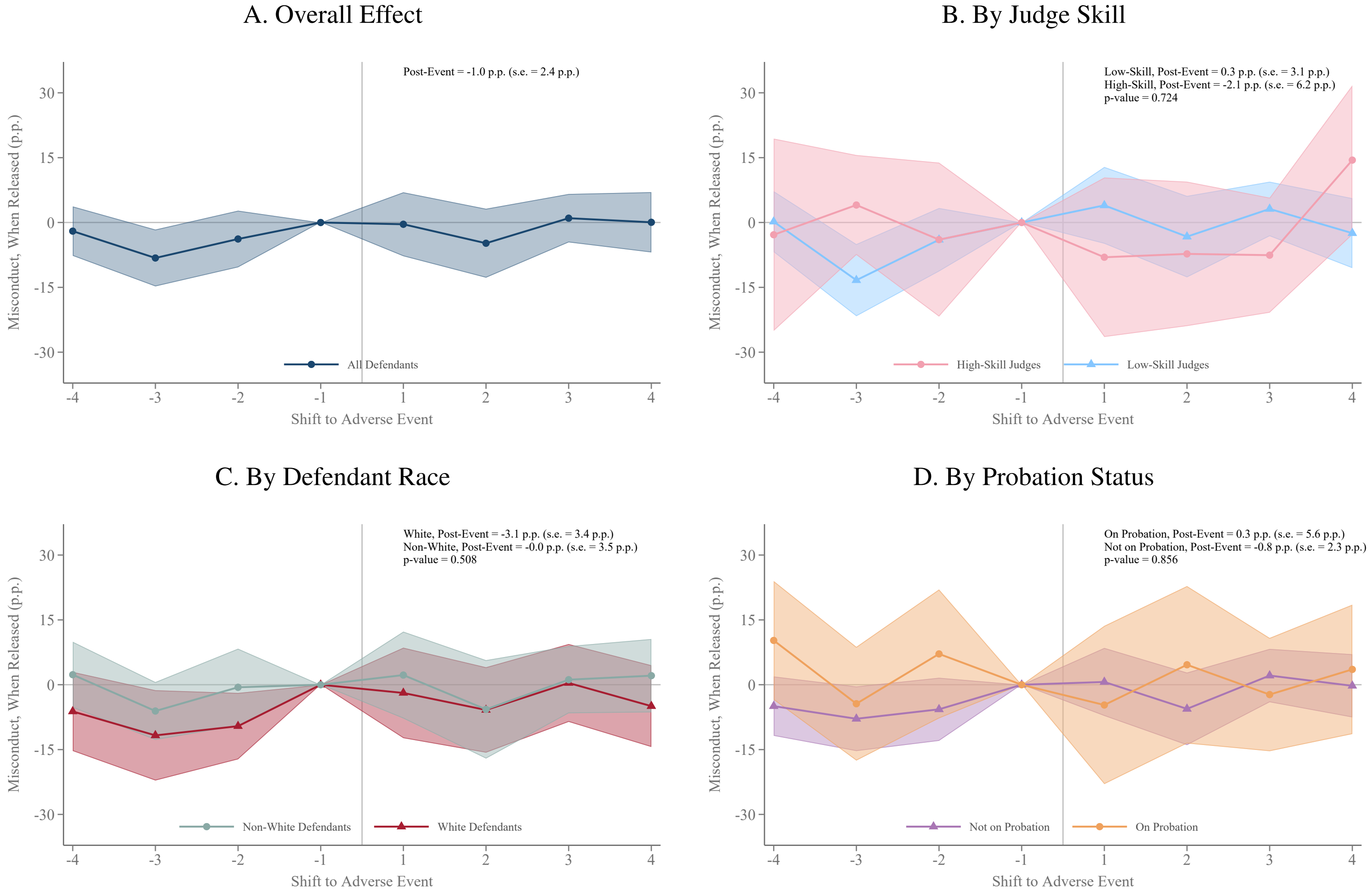
B. By Judge Skill

C. By Defendant Race

D. By Probation Status

Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release on harsh overrides for observably low-risk defendants. See the notes of Figure 8 for additional details.

Figure A.16: Effect of Adverse Events on Conditional Misconduct



Notes. This figure plots event-study estimates of hearing a case involving a defendant arrested for a serious violent felony while on pretrial release on conditional misconduct rates. See the notes of Figure 8 for additional details.

Table A.1: Algorithmic Release Recommendations

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

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

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

	All Defendants	Recommend Release	Recommend Detain
	(1)	(2)	(3)
Age at Current Arrest	0.00000	0.00001	-0.00006
	(0.00002)	(0.00002)	(0.00005)
Age at First Arrest	-0.00001	-0.00002	-0.00003
	(0.00002)	(0.00002)	(0.00012)
Prior Arrests	0.00001	0.00002	-0.00001
	(0.00003)	(0.00004)	(0.00006)
Prior Felonies	-0.00003	-0.00006	0.00017
	(0.00008)	(0.00009)	(0.00021)
Prior Misdemeanors	0.00000	0.00000	0.00012
	(0.00007)	(0.00009)	(0.00015)
Pending Charges	-0.00007	0.00008	0.00009
	(0.00021)	(0.00034)	(0.00031)
Property Charge	0.00049	0.00098	-0.00167
	(0.00057)	(0.00062)	(0.00141)
Drug Charge	0.00024	0.00002	0.00133
	(0.00034)	(0.00031)	(0.00104)
Public Order Charge	-0.00068	-0.00048	-0.00092
	(0.00042)	(0.00042)	(0.00088)
Traffic Charge	0.00041	0.00066	-0.00065
	(0.00049)	(0.00054)	(0.00090)
Parole/Probation	0.00008	0.00024	-0.00026
	(0.00035)	(0.00031)	(0.00110)
Pretrial Release	0.00055	0.00069	0.00039
	(0.00092)	(0.00106)	(0.00134)
Violent Charge	-0.00082	-0.00058	-0.00163
	(0.00062)	(0.00060)	(0.00108)
Male	-0.00027	-0.00002	-0.00136
	(0.00047)	(0.00046)	(0.00123)
White	-0.00005	-0.00012	0.00057
	(0.00037)	(0.00038)	(0.00073)
Joint p-value	[0.356]	[0.328]	[0.514]
Shift x Time FE	Yes	Yes	Yes
Cases	37,855	31,992	5,795

Notes. This table reports OLS estimates of regressions of judge leniency on case and defendant characteristics. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a given bail judge. All regressions control for shift-by-time fixed effects. The p-values reported at the bottom of each column are from F-tests of the joint significance of the variables listed in the rows. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses. After accounting for the shift-by-time fixed effects, there are 68 singleton observations in the recommend detain sample, which are automatically dropped in the regression in column 3.

Table A.3: First Stage Effects of Judge Leniency

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

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

Table A.4: Extrapolations of Conditional Misconduct Rates

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

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

Table A.5: Defendant Characteristics for Lenient and Harsh Overrides

	Lenient Override			Harsh Override		
	High-Skill Judges	Low-Skill Judges	p-value	High-Skill Judges	Low-Skill Judges	p-value
	(1)	(2)	(3)	(4)	(5)	(6)
Age at Current Arrest	-1.49	-1.47	0.99	-0.32	-0.29	0.95
	(2.72)	(1.74)		(0.49)	(0.29)	
Age at First Arrest	-0.49	0.37	0.12	-0.01	0.18	0.07
	(0.48)	(0.29)		(0.09)	(0.05)	
Prior Arrests	-0.13	-0.08	0.88	0.32	0.15	0.21
	(0.29)	(0.15)		(0.11)	(0.08)	
Prior Felonies	0.40	0.07	0.64	-0.03	0.16	0.62
	(0.56)	(0.44)		(0.31)	(0.20)	
Prior Misdemeanors	0.48	-0.43	0.22	-0.35	0.02	0.33
	(0.67)	(0.31)		(0.33)	(0.17)	
Pending Charges	0.42	-0.79	0.39	0.30	3.39	0.03
	(1.18)	(0.76)		(1.26)	(0.74)	
Property Charge	1.02	0.80	0.97	1.44	0.77	0.67
	(5.73)	(2.46)		(1.48)	(0.57)	
Drug Charge	9.90	3.13	0.22	-0.74	-2.04	0.21
	(5.16)	(1.93)		(0.83)	(0.63)	
Public Order Charge	-1.03	-5.69	0.47	2.91	3.62	0.43
	(6.13)	(1.77)		(0.81)	(0.41)	
Traffic Charge	9.17	13.12	0.41	-2.88	-6.99	0.00
	(4.00)	(2.63)		(1.16)	(0.86)	
Parole/Probation	-19.50	-18.78	0.89	11.08	13.42	0.30
	(5.07)	(1.73)		(2.06)	(0.88)	
Pretrial Release	-9.25	1.46	0.13	0.86	-3.48	0.03
	(5.93)	(3.97)		(1.76)	(0.96)	
Violent Charge	-3.33	-1.28	0.74	2.23	0.46	0.22
	(5.63)	(2.59)		(1.11)	(0.94)	
Male	-7.47	-7.50	1.00	4.31	4.07	0.80
	(5.44)	(2.90)		(0.85)	(0.44)	
White	-3.52	2.35	0.14	-0.22	-1.40	0.17
	(2.85)	(2.79)		(0.61)	(0.61)	
Shift x Time FE	Yes	Yes		Yes	Yes	
Risk Score FE	Yes	Yes		Yes	Yes	
Cases	1,244	4,619		6,665	25,327	

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